Text-based Classification as a Data Mining Technique - The Case of Consumer Complaints Dataset

Yara Salman [*]

Dr. Kinda Abou Kasem [**]

(Received 4 / 5 / 2025. Accepted 15 / 7 / 2025)

□ ABSTRACT □

Text-based classification is a technique that can be used to identify different types of data from an application perspective. Various research is underway to identify methods for identifying data categories from a set of input data. In this paper, we implemented an SVM model to classify text contained within consumer complaints in the Consumer Financial Protection Bureau (CFPB) database.

The data was preprocessed and then split into training and test data, after which features were extracted in preparation for model building. A set of tests were conducted to validate the selected classifier, and experimental results showed that the SVM classifier achieved an accuracy of Training data accuracy is 99.576% and test data accuracy is 82.72%. The proposed model offers an incremental approach to text classification because it dynamically trains the classifier from a new set of user-provided data.

Keywords: Data mining, consumer complaint database, text classification, SVM classifier.

[*] **PhD student, Department of Computer Engineering and Automatic Control, Faculty of Mechanical and Electrical Engineering, Latakia University (Formerly Tishreen), Latakia, Syria. (Email: yarahsalman1991@gmail.com);**

**** Professor, Department of Computer Engineering and Automatic Control, Faculty of Mechanical and Electrical Engineering, Latakia University (Formerly Tishreen), Latakia, Syria.**

التصنيف القائم على النص كتقنية للتنقيب في البيانات – حالة مجموعة بيانات شكاوى المستهلكين

يارا سلمان *

د. كندة أبو قاسم **

(تاريخ الإيداع 4 / 5 / 2025. قُبِل للنشر في 15/ 7 / 2025)

□ ملخّص □

التصنيف القائم على النص Text-based classification هو تقنية يمكن استخدامها لتحديد أنواع مختلفة من البيانات من وجهة نظر التطبيقات. تجري أبحاث مختلفة لتحديد طرق معرفة فئات البيانات من مجموعة من بيانات الإدخال. في هذه الورقة، قمنا بتنفيذ موديل SVM في تصنيف النصوص الواردة ضمن شكاوي المستهلكين في قاعدة بيانات شكاوى المستهلكين المقدمة إلى مكتب الحماية المالية للمستهلك (Consumer Financial Protection Bureau (CFPB).

تم تطبيق المعالجة المسبقة للبيانات ثم تقسيمها إلى بيانات تدريب واختبار، وبعدها تم استخراج الميزات تحضيراً لإنشاء الموديل. تم إجراء مجموعة من الاختبارات للتحقق من المصنف المحدد وقد أظهرت النتائج التجريبية أن مصنف SVM وصل إلى دقة بيانات التدريب 99.576 % ودقة بيانات الاختبار 82.72%. يقدم النموذج المقترح نهجاً تدريجياً لتصنيف النص لأنه يدرب المصنف ديناميكياً من مجموعة جديدة من البيانات التي يوفرها المستخدمون.

الكلمات المفتاحية: التنقيب في البيانات، قاعدة بيانات شكاوي المستهلكين، التصنيف النصي، مصنف SVM.

* **طالبة دكتوراه – قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية، جامعة اللاذقية (تشرين سابقاً)، اللاذقية، سورية. (ايميل: yarahsalman1991@gmail.com).**

** **أستاذ – قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية، جامعة اللاذقية(تشرين سابقاً)، اللاذقية، سورية.**

مقدمة:

تستفيد الصناعات بشكل كبير من تطوير أنظمة آلية للتنقيب في البيانات المنظمة القابلة للاستخدام من مصادر نصية غير منظمة [1]. وفي العصر الرقمي، شهد مجال تصنيف النصوص نمواً تحويلياً من خلال تطبيق خوارزميات التعلم الآلي والتنقيب في البيانات [2].

تصنيف النصوص هو عملية تصنيف المعلومات في صيغة نصية حسب محتواها، أي حسب الرسائل التي قد تنقلها الكلمات الموجودة فيها [3]. تعد أتمتة هذه العملية أمراً بالغ الأهمية حتى نتمكن من تصنيف كمية كبيرة من المعلومات النصية بطريقة حرجة من حيث الوقت. نظراً للكم الهائل من المعلومات النصية التي تحتاج إلى معالجة، فإن التصنيف النصي الآلي يجد تطبيقاً واسع النطاق في مجموعة متنوعة من المجالات مثل استرجاع النص، والتلخيص، واستخراج المعلومات والإجابة على الأسئلة، من بين أمور أخرى [4].

يندرج تصنيف النصوص Text classification ضمن المجال الأوسع للتنقيب في النصوص text mining، وهو المصطلح العام لعملية استخلاص أي معلومات من نص معين باستخدام مجموعة متنوعة من تقنيات معالجة النصوص. وعادةً ما ينطوي ذلك على معالجة النص المدخل لجعله مناسباً للتنقيب فيما يتعلق بالمشكلة المعينة، وإيجاد أنماط في البيانات ثم تقييم المخرج الناتج. بصرف النظر عن تصنيف النصوص، فإن التنقيب في النصوص ينطوي عادةً أيضاً على تلخيص المستندات، وتحليل المشاعر، وتجميع النصوص، والتنقيب في الويب [5].

تؤدي مجموعة البيانات المُسمَّاة للمصنف لأغراض التدريب في عمليات تصنيف النصوص إلى التعلم الخاضع للإشراف. يمكن تعريف التعلم الخاضع للإشراف على أنه مهمة استنتاج المعلومات من البيانات التي تم تصنيفها. في هذا النهج، يتكون كل عنصر بيانات في مجموعة البيانات من زوج من العناصر؛ يتكون العنصر الأول من بيانات الإدخال، والعنصر الثاني هو التسمية المرتبطة أو قيمة الإخراج المطلوبة لهذا الإدخال. ثم يتم تحليل مجموعة البيانات المكونة من أزواج البيانات هذه واستنتاج دالة تسمح للخوارزمية بإنتاج قيم إخراج أو تسميات لبيانات إدخال جديدة غير مُسمَّاة. يتناقض مفهوم التعلم الخاضع للإشراف هذا مع مفهوم التعلم غير الخاضع للإشراف حيث لا يتم تصنيف مجموعة البيانات المستخدمة للتدريب [6].

عرّفت الورقة البحثية في [7] تصنيف النصوص على أنه "مهمة فرز مجموعة من المستندات تلقائياً إلى فئات (أو أصناف أو مواضيع) من مجموعة محددة مسبقاً" وذكرت أنها كانت في مجال استرجاع المعلومات والتعلم الآلي. وعلى هذا النحو، وجدت العديد من التقنيات المستمدة من هذه المجالات تطبيقاً في تنفيذ مصنفات النصوص.

اقترح Joachims [8] استخدام آلات الدعم المتجهة ((Support Vector Machines (SVMs) لتصنيف النصوص وأظهر أيضاً أن SVMs يمكن أن تقدم أداءً أفضل لمصنفات النصوص مقارنة بتقنيات التعلم الآلي الأخرى المعروفة مثل مصنفات Naïve-Bayes ومصنفات kNN.

قارن Cahyani and Patasik في دراستهما [9] بين أداء نموذجي TF-IDF وWord2Vec لتمثيل الميزات في تصنيف النص العاطفي. استخدم الباحثان أساليب آلة الدعم المتجه ((support vector machine (SVM) وطريقة بايز متعدد الحدود ((Multinomial Naïve Bayes (MNB) لتصنيف النص العاطفي لركاب الخطوط الجوية على تويتر. استخدمت هذه الدراسة ثلاثة سيناريوهات، وهي SVM مع TF-IDF وSVM مع Word2Vec وMNB مع TF-IDF. أظهرت النتائج أن SVM مع طريقة TF-IDF تولد أعلى دقة مقارنة بالطرق الأخرى، ثم

يتبعها MNB مع TF-IDF، والأخير هو SVM مع Word2Vec. تُظهر هذه الدراسة أن نمذجة TF-IDF تتمتع بأداء أفضل من نمذجة Word2Vec وتحسن هذه الدراسة نتائج أداء التصنيف مقارنة بالدراسات السابقة [9].

أجرى Hassan وآخرون [10] تحليل مقارن لتصنيف النصوص حيث يتم تحليل ومقارنة كفاءة خوارزميات التعلم الآلي المختلفة على مجموعات بيانات مختلفة. تعتمد خوارزميات: آلة الدعم المتجه (SVM) وأقرب جار (k-Nearest Neighbor (k-NN) والانحدار اللوجستي ((Logistic Regression (LR) وخوارزمية بايز المتعددة الحدود (MNB) والغابات العشوائية ((Random Forest (RF) على التعلم الآلي. تم استخدام مجموعتين مختلفتين من البيانات لإجراء تحليل مقارن لهذه الخوارزميات. تكشف النتائج أن الانحدار اللوجستي وآلة دعم المتجهات يتفوقان على النماذج الأخرى في مجموعة بيانات IMDB، ويتفوق kNN على النماذج الأخرى لمجموعة بيانات SPAM وفقاً للنتائج التي تم الحصول عليها من النظام المقترح.

أهمية البحث وأهدافه:

تصنيف النصوص هو المجال الأكثر أهمية في معالجة اللغة الطبيعية، حيث يتم فرز بيانات النص تلقائياً إلى مجموعة محددة مسبقاً من الفئات، وهو يلغي الحاجة إلى تصنيف البيانات اليدوي، وهي آلية أكثر تكلفة وتستغرق وقتاً طويلاً [10]. يتم تطبيق تصنيف النصوص بشكل واسع في الأعمال التجارية كما في حالات تصفية البريد العشوائي واتخاذ القرار واستخراج المعلومات من البيانات الخام والعديد من التطبيقات الأخرى.

تُعد مجموعات بيانات شكاوى المستهلكين مفيدة لفهم استياء العملاء وتحسين المنتجات والخدمات [11]، وتهدف هذه الورقة إلى تطبيق تقنيات التصنيف النصي على مجموعة بيانات شكاوى المستهلكين لتصنيف وتحليل الشكاوى بشكل فعال، ويتمحور العمل حول توقع المنتج موضع الشكوى بناءً على محتوى موضوع الشكوى.

تم اختيار مصنف النصوص باستخدام **SVM** نظراً لما يتمتع به من قدرة عالية على التعامل مع بيانات نصية عالية الأبعاد، إضافة إلى أدائه المستقر في مهام تصنيف النصوص المثبت من قبل [9]. كما يتميز هذا النموذج بقدرته على التعلم التدريجي، مما يجعله ملائماً لتحديث مستمر على قواعد بيانات مثل قاعدة CFPB المتغيرة. استُخدم مصنف SVM في هذا البحث من أجل **تصنيف حالات البيانات غير المرئية**، ويتمحور العمل حول توقع المنتج موضع الشكوى بناءً على محتوى موضوع الشكوى. تم اتباع **نهج تدريجي للتنقيب في النصوص**، حيث تُضاف البيانات غير المصنفة التي تم التنبؤ بها إلى مجموعة التدريب الأصلية لتعزيزها. هذا يتيح للمستخدم لاحقاً الاستفادة من قاعدة بيانات موسعة تعكس تعلماً تراكمياً من التكرارات السابقة.

طرائق البحث ومواده:

أدوات البحث

مجموعة بيانات شكاوى المستهلكين المقدمة إلى مكتب الحماية المالية للمستهلك (Consumer Financial Protection Bureau (CFPB) عبارة عن قاعدة بيانات newSQL مؤلفة من مجموعة من الشكاوى التي تلقاها مكتب حماية المستهلك المالي بشأن مجموعة من المنتجات والخدمات المالية الاستهلاكية التي تقدمها البنوك والمؤسسات المالية الأخرى في جميع أنحاء الولايات المتحدة الأمريكية.

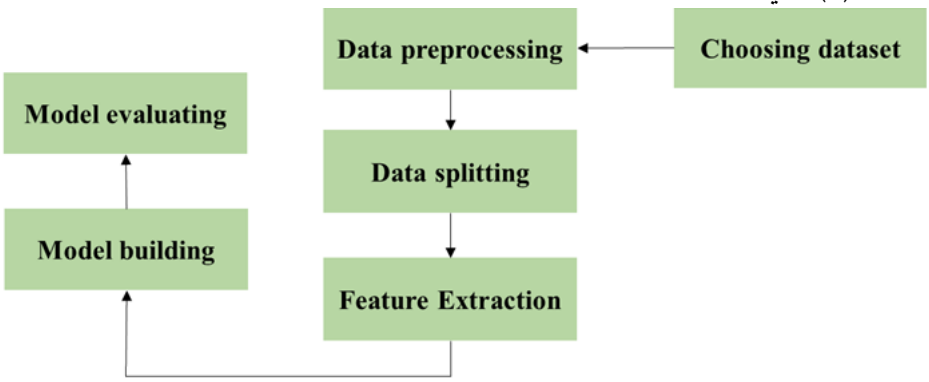
تتكون كل شكوى من سمات يمكنها وصفها وتحديدها بشكل فريد، ويتم استخدام هذه الميزات لتحليل البيانات.

تتوفر البيانات بتنسيقات CSV وJSON ويمكن الوصول إليها من مصادر متعددة بما في ذلك مكتب حماية المستهلك المالي (CFPB) وKaggle التي تُعد مصدراً شائعاً لمجموعات البيانات الجاهزة ومعروفة بدعمها المجتمعي لمشاريع تعلم الآلة وتحديات تحليل البيانات [12].

تم في هذا البحث العمل على منصة google colab لإجراء المحاكاة من حيث تصميم واختبار النظام المقترح، نسخة python3، باستخدام ذاكرة وصول عشوائي 12.7 GB ومعالج رسوميات GPU بذاكرة 15 GB وحجم التخزين 78.2 G.

مخطط العمل

تتألف خطوات العمل من مجموعة من المراحل التي تم اتباعها وصولاً إلى تحقيق الهدف من البحث، وهذه الخطوات موضحة في الشكل (1) وهي:



الشكل 1. مخطط مراحل العمل

A. المعالجة المسبقة للبيانات

1. تنظيف البيانات Data Cleaning: إزالة التكرارات والقيم الفارغة والمعلومات غير ذات الصلة [13]، حيث تم عزل الأعمدة المهمة من مجموعة البيانات dataset الكلية وهما عمودا "الشكاوي" و"المنتجات". ويبين الشكل (2) جزءاً من قاعدة البيانات قبل (a) وبعد (b) إجراء عملية التنظيف مع جزء الكود البرمجي الخاص بالعملية.

```
[ ]  1 # Building a dataset by taking only the important columns we need for the operation.
     2
     3 raw_text_df = train_cat_data[['Product', 'Consumer complaint narrative']]
     4 raw_text_df.head(5)
```

	Product	Consumer complaint narrative
0	Money transfer, virtual currency, or money ser...	i attempted to make a payment to my landlord u...
1	Credit reporting, credit repair services, or o...	the following inquiries are unknown to me and ...
2	Money transfer, virtual currency, or money ser...	this complaint is towards coinbase.com my acco...
3	Credit reporting, credit repair services, or o...	i am on the federal do not call list. however ...
4	Credit reporting, credit repair services, or o...	NaN

a

```
[ ]  1 # Removing missing values
     2 text_df = raw_text_df.dropna(axis=0).reset_index()
     3 text_df.head(5)
```

	index	Product	Consumer complaint narrative
0	0	Money transfer, virtual currency, or money ser...	i attempted to make a payment to my landlord u...
1	1	Credit reporting, credit repair services, or o...	the following inquiries are unknown to me and ...
2	2	Money transfer, virtual currency, or money ser...	this complaint is towards coinbase.com my acco...
3	3	Credit reporting, credit repair services, or o...	i am on the federal do not call list. however ...
4	5	Credit reporting, credit repair services, or o...	xxxx xxxx xxxx xxxx is monitoring my credit in...

b

الشكل 2. جزء من قاعدة البيانات قبل (a) وبعد (b) إجراء عملية تنظيف البيانات.

2. تطبيع النص Text Normalization: تحويل النص إلى أحرف صغيرة وإزالة علامات الترقيم وتطبيق التذييل أو التقسيم [14].

3. التقسيم إلى رموز Tokenization: تقسيم النص إلى كلمات أو رموز فردية، حيث يتم إعطاء كل منتج رقم معرف خاص به [15] (الشكل (3)). تتضمن هذه العملية أيضاً إزالة الكلمات الشائعة التي لا تساهم في التصنيف (على سبيل المثال، "the" و"and")، وهو ما يعرف بـ إزالة الكلمات المتوقفة Stop Words Removal.

```
[ ]  1 # Create a new column 'category_id' with encoded categories
     2 text_df['Product_id'] = text_df['Product'].factorize()[0]
     3 product_id_df = text_df[['Product', 'Product_id']].drop_duplicates()
     4 product_id_df
```

	Product	Product_id
0	Money transfer, virtual currency, or money ser...	0
1	Credit reporting, credit repair services, or o...	1
5	Debt collection	2
6	Student loan	3
14	Mortgage	4
15	Checking or savings account	5
23	Credit card or prepaid card	6
25	Vehicle loan or lease	7
108	Payday loan, title loan, or personal loan	8

الشكل 3. اعطاء كل منتج رقم معرف خاص

B. تقسيم البيانات

تم تقسيم البيانات إلى جزءين: 75% تدريب، 25% اختبار باستخدام الكود البرمجي المبين في الشكل (4).

```
[ ]  1 X_data = text_df['Consumer complaint narrative']
     2 y_data = text_df['Product']
     3 X_train, X_test, y_train, y_test = tts(X_data, y_data, test_size=0.25, random_state=5)
```

الشكل 4. تقسيم قاعدة البينات الى بيانات تدريب واختبار

C. استخراج الميزات Feature Extraction

تمت عملية استخراج الميزات وفق الخطوات التالية [16]:

1. حقيبة الكلمات (Bag of Words (BoW)): تمثيل النص كمجموعة من ترددات الكلمات.

2. تردد المصطلح – تردد المستند العكسي (Term Frequency-Inverse Document Frequency (TF-IDF): وزن أهمية الكلمات بناءً على ترددها في المستند وفي جميع المستندات.

3. تضمين الكلمات (Word Embeddings): استُخدمت نماذج مدربة مسبقاً مثل Word2Vec أو GloVe لالتقاط المعنى الدلالي.

تم استخدام أسلوب تردد المصطلح – تردد المستند العكسي (Term Frequency-Inverse Document Frequency (TF-IDF) لتحديد مدى أهمية كل كلمة داخل الشكوى النصية، حيث تساهم هذه القيم في تمثيل الوثائق بشكل رقمي يعكس وزن الكلمات الأكثر تميّزاً في النص، مما يساعد في تعزيز أداء نموذج التصنيف لاحقاً.

إن أداة تحويل المتجهات باستخدام تردد المصطلح وتردد المستند العكسي (TF-IDF) هي تقنية مستخدمة على نطاق واسع في معالجة النصوص تعكس أهمية الكلمات في المستند بالنسبة لمجموعة من المستندات أو مجموعة النصوص. وهي مفيدة في مهام مثل تصنيف المستندات واسترجاع المعلومات ونمذجة الموضوعات.

ينتج عن تطبيق طريقة TF-IDF تردد المصطلح (TF) وتردد المستند العكسي (IDF) إحصائية رقمية تعكس مدى أهمية الكلمة بالنسبة لمستند في مجموعة [17].

تردد المصطلح (TF)

يقيس تردد المصطلح مدى تكرار ظهور المصطلح في المستند. ونظراً لأن كل مستند يختلف في الطول، فمن الممكن أن يظهر المصطلح مرات أكثر بكثير في المستندات الأطول من المستندات الأقصر.

وبالتالي، غالباً ما يتم تقسيم تردد المصطلح على طول المستند (إجمالي عدد المصطلحات في المستند) كطريقة للتطبيع:

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \quad (1)$$

تردد المستند العكسي (IDF)

يقيس تردد المستند العكسي مدى أهمية المصطلح. أثناء حساب تردد المستند العكسي، يتم اعتبار جميع المصطلحات بنفس الأهمية.

ومع ذلك، قد تظهر مصطلحات معينة، مثل "is" و"of" و"that"، مرات عديدة ولكنها لا تتمتع بأهمية كبيرة. وبالتالي، نحتاج إلى ترجيح المصطلحات المتكررة مع زيادة أهمية المصطلحات النادرة، من خلال حساب ما يلي:

$$IDF(t,D) = \log(\frac{N}{|\{d \in D : t \in d\}|}) \quad (2)$$

حيث:

N هو العدد الإجمالي للمستندات في مجموعة النصوص D.

$|\{d \in D : t \in d\}|$ هو عدد المستندات التي يظهر فيها المصطلح t.

إذا لم يكن المصطلح موجوداً في مجموعة النصوص، فسيؤدي هذا إلى القسمة على الصفر. لذلك من الشائع تعديل المقام إلى $1+|\{d \in D : t \in d\}|$

حساب TF-IDF

يتم حساب درجة TF-IDF عن طريق ضرب الإحصائيتين التاليتين:

$$\text{TF-IDF}\ (t, d, D) = TF(t,d) \times IDF(t,D) \qquad (3)$$

تحويل النص إلى متجهات Vectorization

عند تحويل النص إلى متجهات باستخدام TF-IDF، يتم تمثيل كل مستند كمتجه. يتوافق كل بُعد من أبعاد المتجه مع مصطلح منفصل عن مفردات النص.

القيمة في كل بُعد هي درجة TF-IDF للمصطلح في المستند. للتعامل مع الحجم الهائل للمفردات وندرة متجهات المستند الفردية، تستخدم التنفيذات implementations عادةً تمثيلات مصفوفة متفرقة.

بعد تطبيق خوارزمية TF-IDF، ينتج مصفوفة الأهمية التي تحتوي على أهمية كل كلمة ذكرت بقاعدة البيانات بالنسبة لكل شكوى قدمت في قاعدة البيانات [17]، ويبين الشكل (5) مصفوفة الأهمية.

```
1 tfidf = TfidfVectorizer(sublinear_tf=True, min_df=5, ngram_range=(1, 2), stop_words='english')
2 fitted_vectorizer = tfidf.fit(X_train)
3 tfidf_vectorizer_vectors = fitted_vectorizer.transform(X_train)
4 print(tfidf_vectorizer_vectors)
```

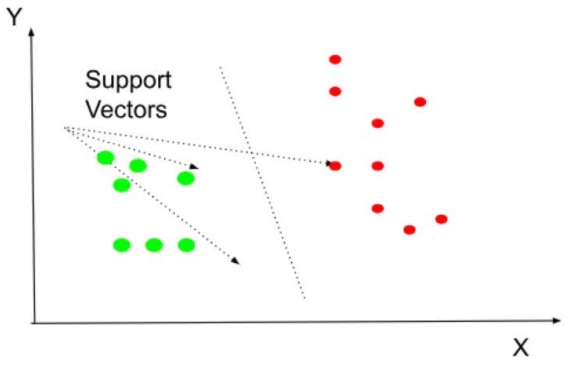
```
  (0, 18343)    0.05628357334510446
  (0, 18251)    0.07962774639892328
  (0, 18245)    0.0728774066260431
  (0, 17716)    0.08184164036896449
  (0, 17316)    0.09667709037676102
  (0, 17274)    0.05163318044956238
  (0, 17223)    0.08534201441750862
  (0, 17184)    0.06663920337176346
  (0, 17019)    0.042825878390626904
  (0, 16906)    0.04175980607555296
  (0, 16561)    0.08409768691229087
  (0, 16522)    0.04893957681717883
  (0, 16356)    0.05674141031818904
  (0, 16349)    0.09835781824530715
  (0, 16348)    0.09835781824530715
  (0, 16336)    0.1056957681981469
  (0, 16311)    0.09667709037676102
  (0, 16310)    0.09686727098386284
  (0, 16261)    0.05407857536212146
  (0, 16107)    0.09386843558773522
  (0, 16102)    0.13975165211115134
  (0, 16047)    0.08046717486221171
  (0, 16031)    0.047343339101403474
  (0, 15811)    0.036953758613683926
  (0, 15697)    0.09266878903118568
  :     :
  (6373, 230)   0.13251817856776402
  (6374, 17484) 0.22385770684339276
  (6374, 17274) 0.04609023855155113
  (6374, 16102) 0.16660238364304233
  (6374, 13143) 0.180613429897238
  (6374, 13105) 0.24077781132463505
  (6374, 11986) 0.3059708934684468
  (6374, 11981) 0.17993638441064l7
```

الشكل 5. شكل مصفوفة الأهمية

D. إنشاء الموديل SVM

نموذج آلة الدعم المتجه ((support vector machine (SVM) هو خوارزمية تعلم آلي قوية ومستخدمة على نطاق واسع ويمكن استخدامها للتصنيف والانحدار واكتشاف القيم المتطرفة، ولكنها الأفضل لمشاكل التصنيف [18]. ينشئ تصنيف SVM خط تقسيم مثالي أو مستوى فائق في مساحة مكونة ذات أبعاد أعلى لرسم خريطة للمعلومات بأقل قدر من المخاطر [18].

تم بناء نموذج تصنيف ثنائي (Binary Classification) باستخدام SVM، حيث يعتمد على تمثيل النصوص باستخدامTF-IDF . يهدف النموذج إلى فصل فئتين من الشكاوى (مثل نوع المنتج أو نوع المشكلة) عبر مستوى فائق في فضاء متعدد الأبعاد. وقد تم اعتماد نهج تدريجي يسمح بتحديث نموذج التصنيف ببيانات جديدة، مما يزيد من دقته مع مرور الوقت ويجعله قابلاً للتكيف مع بيانات المستخدم المستقبلية.



الشكل 6. مبدأ موديل SVM

تم استخدام الوزن الناتج عن TF-IDF ومتجه الكلمات في المرحلة السابقة (مصفوفة الأهمية) كخصائص تصنيف.

بعد إعداد البيانات وتقسيمها، يتم تدريب النموذج، حيث تحاول خوارزمية SVM العثور على المستوى الفائق الذي يفصل نقاط فئة واحدة عن نقاط الفئة الأخرى إلى أقصى حد. وهي تفعل ذلك من خلال العثور على المستوى الفائق الذي يحتوي على أكبر هامش أو مسافة بين نقاط الفئتين.

التنبؤ: بمجرد تدريب النموذج، يمكن استخدامه للتنبؤ بالبيانات الجديدة. لإجراء تنبؤ، يتم تحويل نقطة البيانات الجديدة إلى نفس مساحة الميزة مثل بيانات التدريب ثم يتم تحديد تسمية الفئة بناءً على أي جانب من المستوى الفائق تقع النقطة عليه.

E. تقييم الموديل

تم استخدام الدقة Accuracy لحساب نسبة التعرف في مجموعة البيانات لتقييم أداء النموذج، وهي تحسب من العلاقة :[19]

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Negative} \quad (4)$$

حيث أن:

True Positive الحالات الصحيحة الموجبة،

True Negative الحالات الصحيحة السالبة،

False Positive الحالات الخاطئة الموجبة،

False Negative الحالات الخاطئة السالبة،

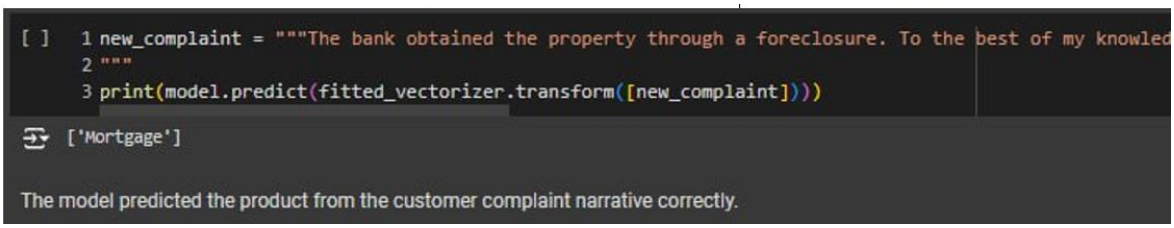
تشتق الحالات السابقة من خلال مصفوفة الارتباك Confusion matrix، وهي أداة قوية تستخدم لتقييم أداء نماذج التصنيف في التعلم الآلي، إذ توفر ملخصاً جدولياً لتوقعات النموذج مقارنة بالنتائج الفعلية، مما يوفر رؤى قيمة حول أنواع الأخطاء التي يرتكبها النموذج [20].

سيناريوهات العمل

هدف البحث إلى استخلاص استنتاجات حول طبيعة شكاوى المستهلكين والمجالات المحتملة للتحسين في المنتجات والخدمات المالية من خلال التنقيب في هذه البيانات باستخدام موديل SVM وخوارزمية التصنيف القائم على النص Text-based classification لمحتوى الشكاوي المسجلة في قاعدة البيانات CFPB. ومن أجل تقييم فعالية نموذج التصنيف المقترح باستخدام خوارزمية SVM، تم اختبار النموذج عبر أربعة سيناريوهات مختلفة، تختلف فيما بينها من

حيث نوع المدخلات المستخدمة والغرض التصنيفي المطلوب من النموذج. يهدف هذا التقييم المتعدد إلى قياس دقة النموذج في سياقات مختلفة تماثل حالات الاستخدام الواقعية:

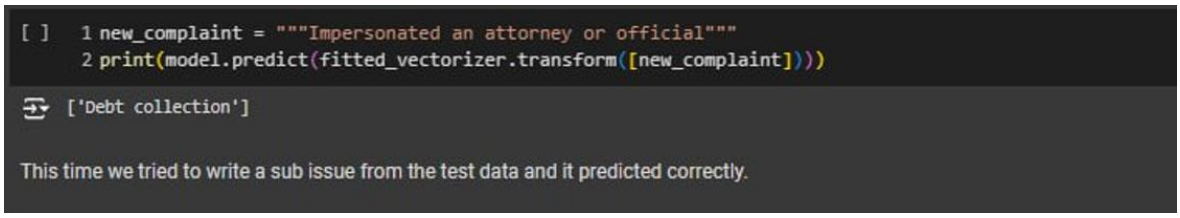
السيناريو الأول: التنبؤ بالمنتج قيد الشكوى كما هو موضح في الشكل (7). يعتمد على تقديم النص الكامل للشكوى كمُدخل للنموذج، ويُطلب منه التنبؤ باسم المنتج المالي المرتبط بالشكوى (مثل: Mortgage أو Credit Card). يُمثّل هذا السيناريو الحالة التقليدية لتصنيف الشكاوى وفق المنتج.



الشكل 7. السيناريو الأول

عند تطبيق السيناريو الأول وطلب التنبؤ بالمنتج قيد الشكوى باستخدام موديل SVM وخوارزمية التصنيف القائم على النص، نجح الموديل في التعرف على المنتج، حيث كان "Mortgage".

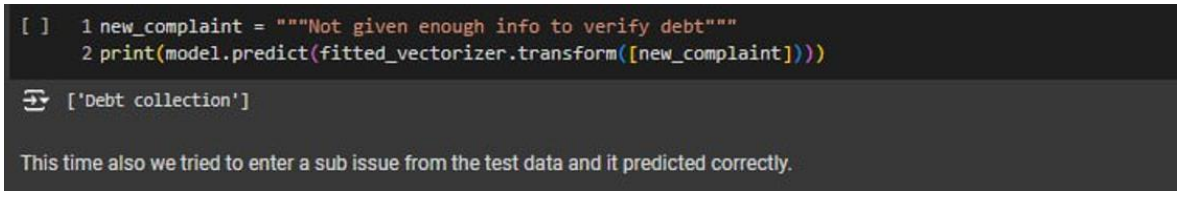
السيناريو الثاني: كتابة الموضوع الفرعي للشكوى كما هو موضح في الشكل (8). يتم فيه إدخال نص مكتوب يعكس محتوى الشكوى أو وصف المستخدم للمشكلة، ويُطلب من النموذج التنبؤ بـ الموضوع الفرعي للشكوى (Sub-issue)، مثل Debt Collection أو Account Management. يُبرز هذا السيناريو قدرة النموذج على فهم الدلالات النصية وتحديد طبيعة المشكلة.



الشكل 8. السيناريو الثاني

عند تطبيق السيناريو الثاني وكتابة الموضوع الفرعي للشكوى باستخدام موديل SVM وخوارزمية التصنيف القائم على النص، نجح الموديل في ذلك حيث كان الموضوع الفرعي للشكوى هو "Debt collection".

السيناريو الثالث: إدخال الموضوع الفرعي للشكوى كما هو موضح في الشكل (9). يتضمن تقديم وصف موجز أو عنوان نصي مختصر يمثل الموضوع الفرعي للشكوى، دون إدخال الشكوى الكاملة. يهدف هذا السيناريو إلى اختبار قدرة النموذج على التصنيف انطلاقاً من مدخلات محدودة ومباشرة، تحاكي الحالات التي يكون فيها المستخدم قدّم فقط عنواناً أو جملة قصيرة.



الشكل 9. السيناريو الثالث

عند تطبيق السيناريو الثالث وإدخال الموضوع الفرعي للشكوى باستخدام موديل SVM وخوارزمية التصنيف القائم على النص، نجح الموديل في ذلك.

السيناريو الرابع: اكتشاف المنتج قيد الشكوى وموضوع المشكلة كما هو موضح في الشكل (10). يُطلب من النموذج تصنيف الشكوى المدخلة لتحديد كل من اسم المنتج والموضوع الفرعي معاً. ويمثل هذا السيناريو الحالة الأكثر شمولاً، إذ يتعامل النموذج مع مهمة تصنيف مركّبة (Multi-label Classification)، مما يتطلب دقة أعلى في تمثيل النص وتحليله.

```
[ ]  1 new_complaint = "PayPal is randomly sending PayPal account holders debit cards without permission
     2 print(model.predict(fitted_vectorizer.transform([new_complaint])))

['Money transfer, virtual currency, or money service']
```

الشكل 10. السيناريو الرابع

عند تطبيق السيناريو الرابع واكتشاف المنتج قيد الشكوى باستخدام موديل SVM وخوارزمية التصنيف القائم على النص، نجح الموديل في ذلك حيث كان "Money transfer, virtual currency, or money service".

النتائج والمناقشة:

نلاحظ من النتائج السابقة أن موديل SVM وخوارزمية التصنيف القائم على النص قد نجح في التنبؤ بسيناريوهات العمل الأربعة. وقد تم تقييم دقة أداء الموديل في تصنيف الشكاوي استناداً للمحتوى النصي وتصنيفه باستخدام الدقة Accuracy، ويبين الشكل (11) مصفوفة الارتباك التي تم حسابها من خلال اختبارات الموديل باستخدام السطر المبين في الشكل (12).

الشكل 11. مصفوفة الارتباك

```
# Compute the confusion matrix
cm = confusion_matrix(y_test, y_pred)
```

الشكل 12. حساب مصفوفة الارتباك

تم تقييم دقة الموديل من خلال كل من الدقة Accuracy، الصحة Precision، Recall، وF1 score من خلال مصفوفة الارتباك كما هو مبين في الشكل (13). بلغت دقة بيانات اختبار الموديل Accuracy 82.72، مما يعني أن الموديل قادر على التنبؤ بالشكوى من خلال التصنيف المعتمد على النصوص بدرجة جيدة جداً.

Accuracy: 0.8272
Precision (weighted): 0.8264
Recall (weighted): 0.8272
F1 Score (weighted): 0.8196

الشكل 13. دقة الموديل

يتفق الأداء الجيد للخوارزمية المستخدمة في البحث مع دراسة Joachims [8] التي أظهرت أداءً أعلى لخوارزمية SVM من Naïve Bayes وk-NN عند استخدامها لتصنيف مستندات نصية ذات أبعاد مرتفعة. كما دعمت دراسة Sarkar[5] هذه النتائج، مؤكدةً أن SVM أكثر استقراراً وفعالية في التصنيف النصي، خاصة عند استخدام تمثيلات مثل TF-IDF.

في دراسة Cahyani and Patasik [9]، تم استخدام ثلاث مجموعات: TF-IDF + SVM، Word2Vec + SVM، وWord2Vec + MNB، وبلغت أعلى دقة لديهم حوالي 94.3% باستخدام TF-IDF + SVM، مما يجعل نتائج البحث الحالي (دقة بيانات التدريب 99.576%) متقدمة نسبياً. قد يعود ذلك إلى طبيعة البيانات الأكثر تنظيماً في CFPB.

من جهة أخرى، أظهرت دراسة Hassan في [10] أن SVM والانحدار اللوجستي يتفوقان على خوارزميات مثل Random Forest وk-NN في تصنيف مجموعات نصوص مثل IMDB وSPAM، وهو ما يدعم اختيار SVM في هذا العمل.

الاستنتاجات والتوصيات:

الجزء الأكثر أهمية في معالجة اللغة الطبيعية هو تصنيف النص، والذي يصنف تلقائياً بيانات النص إلى مجموعة من الفئات المرغوبة. من جهة أخرى، فمن الضروري توظيف التنقيب في النصوص لاستخلاص معلومات مفهومة من بيانات غير معلومة. تعد التقنيات القائمة على التعلم الآلي ضرورية لتصنيف النص والتنقيب في بيانات النصوص.

أظهرت نتائج هذا البحث أن خوارزمية آلة الدعم المتجه (SVM) تمثل خياراً فعالاً في مهمة تصنيف النصوص ضمن سياق شكاوى المستهلكين، حيث حقق النموذج المصمم دقة تصنيف مرتفعة وصلت إلى 99.576%. وقد تبين أن استخدام تمثيل النصوص بطريقة TF-IDF أسهم بشكل مباشر في تحسين أداء النموذج، من خلال توفير تمثيل رقمي يعكس بوضوح أهمية الكلمات ضمن كل شكوى. كما أثبت النهج التدريجي المُعتمد في تدريب النموذج فعاليته، حيث مكّن من تعزيز الأداء التراكمي عبر دمج الشكاوى المصنفة حديثاً في عملية التدريب، ما يتيح للنموذج القدرة على التكيف مع بيانات جديدة باستمرار. بناءً على ذلك، يوصي البحث بتبنّي SVM في تطبيقات تصنيف النصوص ذات

الطبيعة المتغيرة والديناميكية، والاعتماد على TF-IDF كأداة أساسية لاستخلاص الصفات النصية، إلى جانب دراسة جدوى إدماج تقنيات التعلم العميق مستقبلاً كمكمل أو بديل في سيناريوهات أكثر تعقيداً.

References:

[1] N. Sanchez-Pi, L. Martí, and A. C. B. Garcia, *Text classification techniques in oil industry applications*. In International Joint Conference SOCO'13-CISIS'13-ICEUTE'13: Salamanca, Spain, September 11th-13th, 2013 Proceedings (pp. 211-220). Springer International Publishing. (2014).

[2] A. A. Mustafa, and A. M. Abdulazeez. A Review of Text Classification Based on ML and Data Mining Algorithms. *The Indonesian Journal of Computer Science*, Vol. **13**(3), pp. 31, (2024).

[3] X. Luo, Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal*, Vol. **60**(3), pp 3401-3409, (2021).

[4] D. M. Abdullah, and A. M. Abdulazeez, Machine learning applications based on SVM classification a review. *Qubahan Academic Journal*, Vol. **1**(2), pp 81-90, (2021).

[5] A. Sarkar, S. Chatterjee, W. Das, and D. Datta, Text classification using support vector machine. *International Journal of Engineering Science Invention*, Vol. **4**(11), pp 33-37., (2015).

[6] A. Sinha, M. N. B. Naskar, M. Pandey, and S. S. Rautaray, *Text classification using machine learning techniques: comparative analysis*. In 2022 OITS International Conference on Information Technology (OCIT), pp. 102-107, (2022).

[7] F. Sebastiani, *Text categorization*. In Encyclopedia of database technologies and applications (pp. 683-687). IGI Global. (2005).

[8] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. In European conference on machine learning (pp. 137-142). Berlin, Heidelberg: Springer Berlin Heidelberg. (1998).

[9] D. E. Cahyani, and I. Patasik, Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, Vol. **10**(5), pp 2780-2788, (2021).

[10] S. U. Hassan, J. Ahamed, and K. Ahmad, Analytics of machine learning-based algorithms for text classification. Sustainable Operations and Computers, Vol. **3**, pp 238-248, (2022).

[11] https://www.consumerfinance.gov/data-research/consumer-complaints/

[12] C. Chen, and A. Zhang, Consumer Financial Protection Bureau dataset." Consumer Financial Protection Bureau, www.consumerfinance.gov/data-research/. (2019).

[13] A. Makarov, and D. Namiot, Overview of data cleaning methods for machine learning. International Journal of Open Information Technologies, Vol. **11**(10), pp 70-78, (2023).

[14] A. T. H. Nguyen, D. H. Nguyen, N. T. Nguyen, K. T. D. Ho, and K. V. Nguyen, *Automatic Textual Normalization for Hate Speech Detection*. In International Conference on Intelligent Systems Design and Applications (pp. 1-12). Cham: Springer Nature Switzerland. (2023).

[15] M. Sockin and W. Xiong, Decentralization through tokenization. The Journal of Finance, Vol. **78**(1), pp 247-299, (2023).

[16] A. I. Kadhim, An evaluation of preprocessing techniques for text classification. International Journal of Computer Science and Information Security (IJCSIS), Vol. **16**(6), pp 22-32, (2018).

[17] S. Qaiser, and R. Ali, Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, Vol. **181**(1), pp 25-29, (2018).
[18] R. Kurnia, Y. Tangkuman, and A. Girsang, Classification of user comment using word2vec and SVM classifier. Int. J. Adv. Trends Comput. Sci. Eng, Vol. **9**(1), pp 643-648, (2020).
[19] M. Congedo, L. Korczowski, and A. Delorme, Spatio-temporal common pattern: A companion method for ERP analysis in the time domain. Journal of neuroscience methods, Vol. **267**, pp 74-88, (2016).
[20] J. Liang, Confusion matrix: Machine learning. POGIL Activity Clearinghouse, Vol. **3**(4), pp 74-81, (2022).