مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (١٠) العدد (٣) ٢٠٢٦
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (10) No. (3) 2026

تحسين توقع رضا العملاء بدمج تحليل المشاعر (DistilBERT) مع خوارزميات تعلم الآلة

د. كندة أبو قاسم*

يارا سلمان **

(تاريخ الإيداع ٢٠٢٥/١٢/٢ . قُبِل للنشر في ٢٠٢٦/١/٢٧)

□ ملخّص □

يهدف هذا البحث إلى تطوير نموذج تنبؤي دقيق للتنبؤ باعتراض المستهلك استنادًا إلى بيانات شكاوى المستهلكين المقدمة إلى مكتب الحماية المالية للمستهلك (CFPB) باستخدام خوارزميات التعلم الآلي. اعتمدت منهجية العمل على معالجة البيانات الجدولية، ومعالجة اختلال توازن الفئات باستخدام تقنية SMOTE، بالإضافة إلى إدراج ميزة جديدة مستخرجة من تصنيف المشاعر في نصوص الشكاوى. تم تطبيق نموذج DistilBERT لاستخراج التمثيلات العددية (Embeddings) من نصوص الشكاوى، ثم استخدام(LinearSVC) Linear Support Vector Classifier لتصنيف المشاعر إلى ثلاث فئات: إيجابية، سلبية، أو حيادية. أُدرجت نتائج التصنيف كعمود إضافي في جميع مجموعات الخصائص المستخدمة في التجارب (X1، X2، X3)، والتي اختلفت فيما بينها من حيث عدد ونوع الخصائص الجدولية المرافقة لميزة المشاعر.

تم اختبار خوارزميتي شجرة القرار (Decision Tree) والغابة العشوائية (Random Forest) للتنبؤ باعتراض المستهلك على هذه المجموعات الثلاث، وأظهرت النتائج تفوق نموذج الغابة العشوائية في جميع الحالات، محققًا أعلى دقة بلغت ٨٨,٥٠% على المجموعة الكاملة X3. كما حقق نموذج تحليل المشاعر باستخدام DistilBERT دقة ٧٧,٧٤%، مما عزز الأداء الكلي للنماذج التنبؤية. تؤكد هذه النتائج أهمية دمج تقنيات معالجة اللغة الطبيعية مع الخصائص الجدولية في تحسين دقة نماذج التنبؤ بسلوك المستهلك.

الكلمات المفتاحية: رضا العملاء؛ تعلم الآلة؛ الغابة العشوائية؛ SMOTE؛ CFPB؛ DistilBERT؛ LinearSVC.

*أستاذ مساعد في قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية، جامعة اللاذقية، سورية

**طالبة دكتوراه في قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية، جامعة اللاذقية، سورية

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (١٠) العدد (٣) ٢٠٢٦
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (10) No. (3) 2026

Enhancing Customer Satisfaction Prediction by Integrating Sentiment Analysis (DistilBERT) with Machine Learning Algorithms: A Case Study on CFPB Data

Dr. Kina Abou Kasem *
Yara Salman **

(Received 2/12/2025 . Accepted 27/1/2026)

□ **ABSTRACT** □

This study aims to develop an accurate predictive model for forecasting customer satisfaction based on consumer complaint data submitted to the Consumer Financial Protection Bureau (CFPB) using machine learning algorithms. The methodology involved processing tabular data, addressing class imbalance with the SMOTE technique, and incorporating a new feature derived from sentiment analysis of the complaint texts. The DistilBERT model was employed to extract numerical representations (embeddings) from the complaint texts, followed by a Linear Support Vector Classifier (LinearSVC) to categorize them into three sentiment classes: positive, negative, or neutral. The classification results were added as an additional feature to all feature sets used in the experiments (X1, X2, X3), which differed in the number and type of tabular attributes accompanying the sentiment feature.

Both Decision Tree and Random Forest classifiers were evaluated on these three feature sets. The results showed that the Random Forest model consistently outperformed Decision Tree, achieving the highest accuracy of 88.50% on the complete feature set X3. Furthermore, the sentiment analysis model using DistilBERT achieved an accuracy of 77.74%, contributing to the overall improvement of the predictive models. These findings highlight the importance of integrating natural language processing techniques with tabular features to enhance customer satisfaction prediction.

Keywords: Customer satisfaction; Machine learning; Random Forest; SMOTE; CFPB; DistilBERT; LinearSV

* Assistant Professor in the Department of Computer Engineering and Automatic Control, Faculty of Mechanical and Electrical Engineering, Latakia University, Syria
** PhD student in the Department of Computer Engineering and Automatic Control, Faculty of Mechanical and Electrical Engineering, Latakia University, Syria

المقدمة

شهدت السنوات الأخيرة تزايدًا ملحوظًا في استخدام تقنيات تعلم الآلة لتحسين جودة الخدمات المقدمة للعملاء، لا سيما في القطاعات المالية والخدمية، حيث أصبح تحليل شكاوى العملاء أداة محورية لفهم سلوك المستهلك وتحسين تجربة العميل وتعزيز سمعة المؤسسات. ومع النمو المتسارع للتفاعلات الرقمية، تعتمد المؤسسات بشكل متزايد على الأنظمة الآلية لاستخلاص رؤى ذات قيمة من البيانات غير المنظمة، وعلى رأسها النصوص السردية لشكاوى المستهلكين.

تُعد معالجة النصوص من أهم تطبيقات التعلم الآلي، إذ تهدف إلى استخراج الكلمات والعبارات التي تعبّر عن محتوى النصوص ودلالاتها. وفي مجالات مثل التجارة الإلكترونية والخدمات المالية، يمكن للمعرفة المستخرجة من شكاوى العملاء أن توفّر قيمة تحليلية عالية وتساهم في تحسين عملية اتخاذ القرار. إلا أن النصوص التي يُنشئها المستخدمون غالبًا ما تتسم بالتشويش والتعقيد، وتحتوي على معلومات جوهرية وأخرى غير ذات صلة، مما يجعل معالجتها آليًا تحديًا حقيقيًا.

يعتمد المسار التقليدي لتصنيف النصوص على مراحل متعددة تبدأ بتنظيف النصوص وتطبيعها، ثم تحويلها إلى تمثيلات عددية مناسبة للاستخدام في نماذج التعلم الآلي، سواء باستخدام أساليب إحصائية أو تضمينات سياقية متقدمة. ورغم التقدم الملحوظ في هذا المجال، فإن التعامل مع بيانات واقعية مثل شكاوى المستهلكين يفرض تحديات إضافية، أبرزها اختلال توازن الفئات وضرورة دمج البيانات النصية مع الخصائص الجدولية ضمن نموذج موحّد.

بالرغم من فعالية نماذج التعلم الآلي في تحليل البيانات الجدولية، فإن الاقتصار على هذا النوع من البيانات قد يؤدي إلى إغفال مؤشرات دلالية مهمة تعكس مشاعر المستهلكين وسلوكهم الضمني. من هنا تبرز الحاجة إلى تصميم نموذج قادر على الاستفادة من كل من الخصائص الجدولية والمعلومات المستخرجة من النصوص في آن واحد، مع مراعاة مشكلة عدم توازن البيانات.

انطلاقًا من ذلك، تتمثل المشكلة البحثية في كيفية بناء نموذج تنبؤي قادر على التعامل مع بيانات غير متوازنة تجمع بين الخصائص الجدولية والنصية بهدف تحسين التنبؤ بسلوك المستهلك. وبناءً على هذه الإشكالية، يسعى هذا البحث إلى الإجابة عن الأسئلة البحثية التالية:

- هل يمكن لدمج نتائج تحليل المشاعر المستخرجة من نصوص الشكاوى تحسين دقة نموذج التنبؤ برضا العملاء؟
- ما مدى تأثير إدراج ميزة تحليل المشاعر على مقاييس الأداء مقارنة بالنماذج التي تعتمد على البيانات الجدولية فقط؟

الدراسات المرجعية

ركزت العديد من الدراسات السابقة على توظيف تقنيات تعلم الآلة ومعالجة اللغة الطبيعية لتحليل شكاوى العملاء وفهم سلوك المستهلك في القطاعات المالية والخدمية. فقد تناولت بعض الأعمال استخدام نماذج التعلم الآلي للتنبؤ برضا العملاء اعتمادًا على بيانات المراجعات والتفاعلات الرقمية، حيث أظهرت النماذج التجميعية مثل الغابة العشوائية أداءً متفوقًا مقارنة ببعض نماذج التعلم العميق في بيئات معينة(Zaghloul et al., 2024) .

كما أظهرت دراسات أُجريت باستخدام بيانات مكتب حماية المستهلك المالي فعالية نماذج التعلم الآلي في تحليل شكاوى المستهلكين والتنبؤ بسلوكهم استنادًا إلى الاستجابات السردية وإجراءات الشركات المتخذة لمعالجة الشكاوى(CFPB, 2024).

وفي السياق ذاته، تم استخدام نمذجة المواضيع بالاعتماد على تضمينات قائمة على BERT لاكتشاف الموضوعات الكامنة في الشكاوى المالية، مما ساهم في تقديم رؤى أعمق حول المشكلات النظامية المتكررة(Sangaraju et al., 2022).

من جهة أخرى، ركزت بعض الدراسات على تصنيف شكاوى المستهلكين وفقًا لطبيعة الردود وفعالية الشركات في التعامل معها، باستخدام تقنيات تعلم الآلة التقليدية(Vaishnav et al., 2024). كما استُخدمت خوارزميات مثل الانحدار اللوجستي وآلات متجهات الدعم لتحليل شكاوى العملاء والكشف المبكر عن الحالات الحرجة التي تتطلب تدخلاً سريعًا(Alarifi et al., 2023).

وفي مجال تصنيف النصوص، تناولت دراسات أخرى تحديد نوعية الشكوى بناءً على محتواها النصي بهدف دعم فهم نية المستهلك وتحسين استجابة الشركات، مما يؤكد أهمية الاستفادة من المعلومات الدلالية الكامنة في النصوص السردية(Arcipreste, 2023).

أهداف البحث

يهدف البحث إلى:

- بناء نموذج تنبؤي يعتمد على خوارزميات تعلم الآلة للتنبؤ باعتراض المستهلك استنادًا إلى بيانات شكاوى العملاء.
- مقارنة أداء خوارزميات مختلفة (Decision Tree وRandom Forest) على مجموعات ميزات متنوعة.
- دمج نتائج تصنيف المشاعر المستخرجة من نصوص الشكاوى ضمن إطار التنبؤ باعتراض المستهلك، وقياس أثر هذه الميزة على أداء النموذج.

مواد وطرق البحث

- **مجموعة البيانات المستخدمة:**

يقترح هذا البحث إطاراً متكاملاً ومرناً لبناء نموذج تنبؤي باستخدام تقنيات التعلم الآلي، موجه خصيصاً للتعامل مع مجموعات البيانات الجدولية التي تتسم بوجود قيم مفقودة واختلال في توازن الفئات.

اعتمد هذا البحث على مجموعة بيانات شكاوى المستهلكين المقدمة إلى مكتب الحماية المالية للمستهلك (Consumer Financial Protection Bureau – CFPB)، وهي من مجموعات البيانات المفتوحة المستخدمة على نطاق واسع في الأبحاث المتعلقة بسلوك المستهلك وتحليل الشكاوى. تتوفر هذه البيانات بصيغتي CSV و JSON، ويمكن الوصول إليها عبر الموقع الرسمي لمكتب CFPB أو من خلال منصات بيانات مفتوحة مثل Kaggle.

تتضمن مجموعة البيانات معلومات جدولية متعددة مرتبطة بالشكاوى، مثل نوع المنتج المالي، الشركة المقدمة للخدمة، طريقة الرد، وتاريخ الشكوى، بالإضافة إلى النصوص السردية لشكاوى المستهلكين، والتي تمثل مصدراً غنيًا لاستخراج المؤشرات الدلالية والمشاعر الضمنية. كما تحتوي البيانات على متغير الهدف (Consumer disputed?) الذي يشير إلى ما إذا كان المستهلك قد اعترض على رد الشركة، مما يجعلها مناسبة لبناء نماذج تنبؤية لدراسة هذا السلوك.

تم اختيار مجموعة بيانات CFPB في هذا البحث لأنها تمثل بيانات واقعية من بيئة تطبيقية حقيقية، وتجمع بين البيانات الجدولية والنصية في آنٍ واحد، إضافةً إلى احتوائها على مشكلة اختلال في توازن الفئات، مما يجعلها مناسبة لتقييم فعالية تقنيات معالجة البيانات غير المتوازنة مثل SMOTE.

- **أدوات البحث**

تم في هذا البحث العمل على منصة Google Colab لإجراء المحاكاة من حيث تصميم واختبار النظام المقترح، باستخدام نسخة Python 3، مع ذاكرة وصول عشوائي ١٢,٧ GB ومعالج رسوميات NVIDIA Tesla T4 بذاكرة ١٥,٣ GB وحجم تخزين ٧٨,٢ GB.

بالإضافة إلى ذلك، تم استخدام مكتبات برمجية متخصصة منها:

- مكتبة Hugging Face Transformers لتطبيق نموذج DistilBERT واستخراج التمثيلات النصية وتحليل المشاعر،
- مكتبة Scikit-learn لتدريب نموذج Random Forest، وتقسيم البيانات، وتقييم أداء النموذج،
- مكتبة imbalanced-learn (IMB Learn) لتطبيق تقنية SMOTE ومعالجة اختلال توازن الفئات،
- مكتبات معالجة البيانات مثل Pandas وNumPy لتنظيم البيانات، وتحويلها، ومعالجتها قبل إدخالها للنماذج،
- مكتبة PyTorch لدعم تشغيل نموذج DistilBERT المبني على بنية Transformer.

١. **مراحل العمل**

- تحضير البيانات (Data Preparation)، وتتضمن: ١) تحميل البيانات من المصدر. ٢) معاينة القيم المفقودة. ٣) إزالة الأعمدة غير المهمة. ٤) ملء القيم المفقودة بقيم افتراضية.
- استخراج وتحويل الميزات (Feature Engineering)، وتتضمن: ١) استخراج ميزات زمنية (يوم، أسبوع، سنة). ٢) تحويل المتغيرات الفئوية إلى أرقام (Label Encoding). ٣) التحقق من أنواع البيانات وتحويلها عند الحاجة.
- تطبيق نموذج DistilBERT، وهو نسخة مصغرة من نموذج BERT مهيأ مسبقًا لمعالجة النصوص باللغة الإنجليزية (Buyukoz, 2023)، لاستخراج التمثيلات العددية (Embeddings) من نصوص الشكاوى- وتحديداً عمود cosumer complaint narrative. بعد ذلك، تم تدريب نموذج Linear Support Vector Classifier (LinearSVC) على هذه التمثيلات لتصنيف المشاعر النصية إلى ثلاث فئات رئيسية: إيجابية، سلبية، أو حيادية. أُدرجت نتائج هذا التصنيف كعمود إضافي في مجموعة

الميزات، بحيث يمكن للنماذج التنبؤية اللاحقة (مثل Random Forest) الاستفادة من الإشارات الشعورية المستخرجة من النصوص بجانب الخصائص الجدولية الأخرى.

- تقسيم البيانات (Data Splitting)، تم إنشاء ثلاث مجموعات ميزات مستقلة، صُممت بشكل تراكمي بهدف دراسة أثر زيادة عدد ونوع الخصائص على أداء النماذج وهي:
  - x1: تتضمن مجموعة محددة من الخصائص الجدولية الأساسية الأكثر ارتباطًا باعتراض المستهلك بالإضافة إلى ميزة تحليل المشاعر.
  - x2: تشمل مجموعة موسعة من الخصائص الجدولية مع ميزة تحليل المشاعر.
  - x3: تضم جميع الخصائص الجدولية بالإضافة إلى ميزة تحليل المشاعر، بهدف تقييم أثر دمج جميع المعلومات المتاحة على دقة النموذج.

حيث تم فصل البيانات إلى ميزات (X) وهدف (y). تم تقسيمها إلى مجموعة تدريب (٨٠%) ومجموعة اختبار (٢٠%).

يمثّل المتغير الهدف (y) في هذا البحث عمود ?Consumer disputed، وهو متغير ثنائي يشير إلى ما إذا كان المستهلك قد اعترض على رد الشركة (Yes/No) .

لم يتم إنشاء مجموعة تحقق مستقلة، حيث تم الاعتماد على التحقق المتقاطع (Cross-Validation) ضمن مرحلة ضبط المعاملات باستخدام GridSearchCV، حيث تُستخدم أجزاء من بيانات التدريب داخليًا كمجموعات تحقق لاختيار أفضل الإعدادات، مع الإبقاء على مجموعة الاختبار مستقلة للتقييم النهائي.

```
X1 = X_new[['Product', 'Issue', 'Company', 'State', 'Submitted via', 'Timely response?', 'Time interval', 'Sentiment']].values
X2 = X_new[['Product', 'Sub-product', 'Issue', 'Sub-issue', 'Company', 'State', 'Submitted via',
            'Company response to consumer', 'Timely response?', 'Time interval', 'Sentiment']].values
X3 = X_new[predictors].values
```

الشكل ١. مجموعات الميزات المستخدمة

- معالجة اختلال التوازن (Class Imbalance Handling)، وتتضمن تطبيق تقنية SMOTE لتوليد عينات اصطناعية للفئة الأقل تمثيلاً.
- تجريب النماذج (Model Training and Evaluation) من خلال:

تم اختيار Decision Tree كنموذج أولي Baseline نظرًا لبساطته وسهولة تفسيره، وقدرته على التعامل مع الميزات النصية بعد تحويلها. أما Random Forest فقد تم استخدامه كنموذج أكثر قوة واستقرارًا، كونه يقلل من مشكلة الـ overfitting، وقدرة على التعامل مع عدد كبير من الخصائص الفئوية الموجودة في بيانات الشكاوى، مما جعله نموذجًا مناسبًا لمقارنة الأداء مع النماذج الأخرى في الدراسة (Wright & Ziegler, 2017; González-Carvajal & Garrido-Merchán, 2020).

١. تدريب نموذج Decision Tree ومقارنة أدائه عبر ٣ مجموعات خصائص.
٢. تدريب نموذج Random Forest على نفس المجموعات ومقارنة الأداء.

٣. تقييم النماذج باستخدام: الدقة، F1، الاستدعاء، مصفوفة الارتباك.

- تضمّنت مرحلة بناء النموذج النهائي عدّة خطوات مترابطة، بدأت أولاً بـ تحليل أهمية الميزات بهدف تحديد المتغيرات الأكثر تأثيرًا على عملية التنبؤ وتحسين اختيار المدخلات المستخدمة في النمذجة. تلا ذلك إجراء ضبط للمعاملات (Hyperparameter Tuning) باستخدام أساليب مثل GridSearchCV للوصول إلى أفضل إعدادات تحقق أعلى أداء للنموذج. وفي الخطوة الأخيرة تم تدريب النموذج النهائي وتقييمه باستخدام بيانات الاختبار لقياس الدقة والأداء العام.

الشكل ١. مخطط تدفقي لمراحل العمل

٢. مراحل تقييم النموذج والتجارب التطبيقية

تم تصميم الإطار المقترح في هذا البحث ليكون مرنًا وقابلًا للتطبيق على البيانات التي تتضمن خصائص جدولية ونصية مع وجود اختلال في توازن الفئات. نقترح السيناريوهات العملية التالية:

المرحلة الأولى: التقييم الآلي لمشاعر العملاء

الوصف: استخدام نموذج DistilBERT لاستخراج التمثيلات العددية من نصوص الشكاوى، تلاه تدريب نموذج LinearSVC لتصنيف المشاعر النصية إلى ثلاث فئات: إيجابية، سلبية، وحيادية.

النتيجة: تم إدراج مخرجات تصنيف المشاعر كميزة إضافية ضمن مجموعات الخصائص، وحقق نموذج تصنيف المشاعر دقة بلغت ٧٧,٧٤%.

المرحلة الثانية: اختيار النموذج الأفضل أداء في التنبؤ برضا العملاء

الوصف: مقارنة أداء Decision Tree و Random Forest على مجموعات الخصائص المختلفة.

النتيجة: Random Forest تفوق على Decision Tree، وحقق أعلى دقة ٨٨,٥% مع المجموعة الكاملة.

المرحلة الثالثة: تحسين النموذج

الوصف: بعد اختيار الخوارزمية الأفضل أداء(Random Forest) ، تم تحسين النموذج عبر خطوتين رئيسيتين:

١. تحليل أهمية الميزات لتحديد الخصائص الأكثر تأثيرًا في التنبؤ باعتراض المستهلك، حيث تم استبعاد الميزات ذات الأهمية المنخفضة.

٢. ضبط معاملات النموذج باستخدام Grid Search لتحديد العدد الأمثل للأشجار.

النتيجة: أسهمت هذه الخطوات في تبسيط بنية النموذج وتحسين استقراره، حيث حقق أفضل أداء عند استخدام ٢٦٠ شجرة، مع متوسط دقة تحقق متقاطع بلغ ٩١,٣٢%.

المرحلة الرابعة: تقييم النموذج النهائي

الوصف: تم اختبار النموذج النهائي المحسّن على مجموعة الاختبار المستقلة لتقييم أدائه الحقيقي.

النتيجة: أظهر النموذج النهائي أداءً قويًا، حيث بلغت الدقة ٨٧,٩٧%، وارتفع Precision إلى ٩٧,٥٩%، بينما كان Recall أقل نسبيًا (٧٧,٨٥%)، ما يشير إلى قدرة النموذج على تقليل الإيجابيات الكاذبة مع بعض الحاجة لتحسين الاستدعاء.

المقارنة مع نهج قائم على العنقدة وتحليل المشاعر

تم في هذا البحث إجراء مقارنة منهجية مع نهج سابق قائم على دمج العنقدة غير الموجّهة وتحليل المشاعر لاستخلاص ميزة داعمة للتنبؤ باعتراض المستهلك. ففي النهج السابق، تم أولًا تمثيل نصوص شكاوى المستهلكين باستخدام أسلوب TF–IDF، ثم تطبيق خوارزميتي العنقدة غير الموجّهة K–Means و Agglomerative Clustering لتقسيم البيانات إلى ثلاث عناقيد متجانسة. بعد ذلك، تم ربط كل عنقود باتجاه شعوري عام من خلال استخدام أدوات تحليل مشاعر تقليدية (TextBlob)، ومن ثم إدراج ناتج هذا الربط كميزة إضافية ضمن نموذج Random Forest المستخدم للتنبؤ باعتراض المستهلك، حيث تمت المقارنة بين النهجين على نفس مجموعة البيانات، كما تم اعتماد نفس نموذج التصنيف النهائي Random Forest (n_estimators=260).

النتائج ومناقشتها

١. نتائج تحليل المشاعر باستخدام DistilBERT

تم تطبيق نموذج DistilBERT لاستخراج التمثيلات العددية (Embeddings) من نصوص الشكاوى، تلاه استخدام LinearSVC لتصنيف النصوص إلى ثلاث فئات: إيجابية، سلبية، أو حيادية. ويوضح الجدول (١) نتائج تطبيق هذا النموذج، حيث حقق هذا النموذج دقة تصنيف بلغت ٧٧,٧٤%.

الجدول ١. نتائج دقة نموذج DistilBERT لتحليل المشاعر

Accuracy	Precision	Recall	F1-score
77.74	76.54	77.74	76.51

تعكس هذه النتائج قدرة النموذج على استخلاص مؤشرات شعورية مهمة من النصوص، حتى مع تنوع محتواها، وتم استخدام مخرجات هذا التحليل كميزة إضافية في بيانات التدريب لنموذج التنبؤ برضا العملاء، حيث تمت إضافة عمود الـ Sentiment الناتج عن نموذج DistilBERT إلى البيانات الجدولية ليُستخدم كميزة إضافية في تدريب نماذج التنبؤ برضا العملاء.

٢. مقارنة أداء الخوارزميات على مجموعات الخصائص

تم تقييم أداء كل من شجرة القرار (Decision Tree) والغابة العشوائية (Random Forest) على مجموعات الخصائص الثلاثة.

يوضح الجدول (٢) القيم المحققة لكل من الدقة (Accuracy)، والدقة الإيجابية (Precision)، والاستدعاء (Recall)، ودرجة F1.

الجدول ٢. نتائج اختبار دقة خوارزمية شجرة القرار والغابة العشوائية

الخوارزمية	مجموعة الخصائص	Accuracy	Precision	Recall	F1-score
Decision Tree	x1	0.8121	0.817	0.8044	0.8106
	x2	0.8153	0.8158	0.8144	0.8151

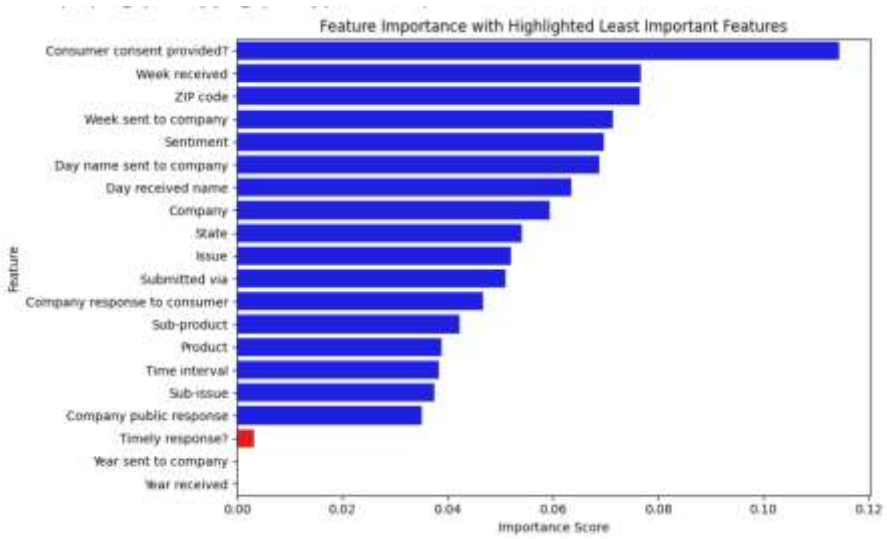
0.8138	0.8226	0.8052	0.8118	x3	
0.8466	0.7994	0.8997	0.8552	x1	**Random Forest**
0.8565	0.8026	0.9183	0.8656	x2	
0.8733	0.7927	0.9722	0.885	x3	

نلاحظ من الجدول (٢) أن خوارزمية Random Forest أظهرت تفوقًا واضحًا على Decision Tree في جميع المجموعات من حيث الدقة ودرجة F1، حيث بلغت أعلى دقة له ٨٨,٥٠% على المجموعة الكاملة، مقارنة بـ ٨١,١٨% لأفضل أداء لشجرة القرار، مما يعكس قدرة النموذج التجميعي على التقاط أنماط أكثر تعقيدًا.

ارتفعت دقة Random Forest من ٨٥,٥٢% (x1) إلى ٨٨,٥٠% (x3)، أي بزيادة قدرها ٢,٩٨%، وهو ما يعكس أهمية دمج جميع الخصائص بما فيها نتائج تحليل المشاعر في تحسين التنبؤ برضا العملاء. من جهة أخرى، لوحظ أن قيم Precision ارتفعت بشكل كبير في Random Forest عند الانتقال إلى المجموعة الكاملة (٩٧,٢٢%)، مما يشير إلى تحسن قدرة النموذج على تقليل الإيجابيات الكاذبة. بالنسبة لخوارزمية Decision Tree، كانت التحسينات أقل وضوحًا، وهو متوقع نظرًا لبنية النموذج البسيطة مقارنة بالغابة العشوائية.

٣. تحليل أهمية الخصائص (Feature Importance)

تم إجراء تحليل لأهمية الخصائص باستخدام نموذج الغابة العشوائية (Random Forest) بهدف تحديد الميزات الأكثر تأثيرًا في عملية التنبؤ برضا العملاء، ويبين الشكل (٢) تمثيلاً بيانياً لأهمية الميزات مع الإشارة إلى الميزات الأقل أهمية.



الشكل ٢. درجة أهمية الميزات

أظهر تحليل الأهمية باستخدام Random Forest أن ميزة "موافقة المستهلك Consumer consent provided" كانت الأكثر تأثيرًا (٠,١١٤٥)، تليها الميزات الزمنية مثل "أسبوع الاستلام Week received" و"أسبوع الإرسال للشركة Week sent to company"، إضافة إلى ميزة "تحليل المشاعر Sentiment" التي أُدرجت من خلال نموذج DistilBERT + LinearSVC، حققت وزنًا ملحوظًا (٠,٠٦٩٧)، مما يبرز مساهمتها في تحسين دقة النموذج.

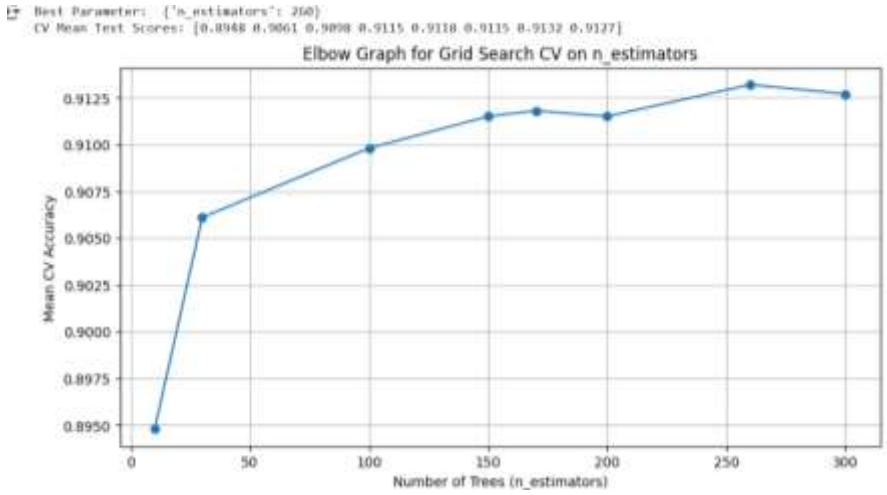
بناءً على هذا التحليل، تم حذف الخصائص الأقل تأثيرًا والتي سجلت أوزانًا قريبة من الصفر، وهي:

Timely response بأهمية (٠,٠٠٣٣)، Year sent to company بأهمية (٠,٠٠٠٠٠٠٩)، وYear received بأهمية معدومة. يؤدي استبعاد هذه الخصائص إلى تبسيط بنية النموذج وتقليل التعقيد الحسابي دون التأثير سلبًا على الأداء.

٤. ضبط معاملات نموذج الغابة العشوائية

باستخدام البحث الشبكي (Grid Search) على قيم n_estimators، تم تحديد العدد الأمثل للأشجار في نموذج Random Forest عند ٢٦٠، حيث حقق متوسط دقة عبر التحقق المتقاطع (CV Mean Test Score) بلغ ٩١,٣٢% (الشكل ٣).

```
Best Parameter:  {'n_estimators': 260}
CV Mean Test Scores: [0.8948 0.9061 0.9098 0.9115 0.9118 0.9115 0.9132 0.9127]
```

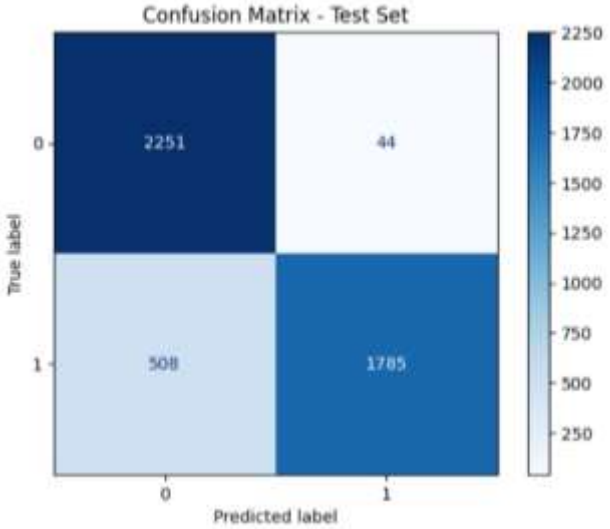


الشكل ٣. تقدير عدد أشجار النموذج

بعد تدريب النموذج النهائي باستخدام هذا العدد، حقق النموذج النهائي على مجموعة الاختبار دقة بلغت ٨٨,٠٠%، مع Precision مرتفع (٩٨,٠٠%) واستدعاء اقل نسبياً (٧٨,٠٠%)، مما يشير قوة النموذج في تقليل الإيجابيات الكاذبة، لكنه يحتاج إلى تحسين إضافي لالتقاط جميع الحالات الإيجابية الحقيقية.

```
Final model
Training Accuracy: 0.9998
Evaluation on held-out test set:
Accuracy: 0.88
Precision: 0.98
Recall: 0.78
F1 Score: 0.87
Confusion Matrix:
 [[2251   44]
 [ 508 1785]]
```

توضح مصفوفة الارتباك في الشكل (٤) توزيع التنبؤات الصحيحة والخاطئة للفئتين المستهدفتين (اعترض المستهلك و لم يعترض المستهلك). وتشير القيم إلى أن النموذج حقق أداءً قويًا في التعرف على حالات عدم الاعتراض، بينما لا يزال هناك مجال لتحسين قدرة النموذج على اكتشاف جميع حالات الاعتراض.



الشكل ٤. مصفوفة الارتباك لمجموعة الاختبار

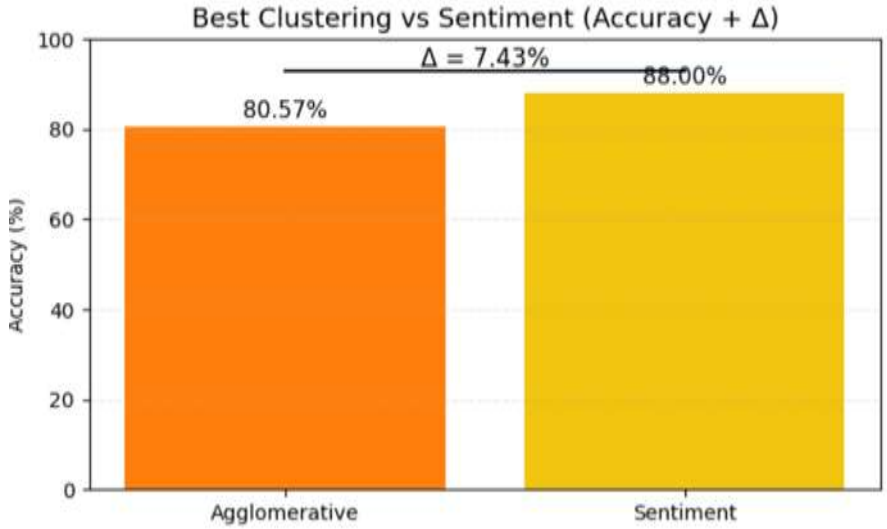
٥. نتائج مقارنة أداء النموذج المقترح مع نهج قائم على العنقدة وتحليل المشاعر

يوضح الجدول (٣) مقارنة أداء النموذج الحالي مع نهج سابق قائم على دمج العنقدة غير الموجّهة وتحليل المشاعر. أظهر نموذج Agglomerative Clustering دقة تصنيف بلغت ٠,٨٠٥٧، متفوقًا بشكل طفيف على نموذج KMeans الذي حقق دقة ٠,٨٠٣١، مع تقارب ملحوظ في قيم مقاييس Precision و Recall و F1-score بين النموذجين. وتشير هذه النتائج إلى أن العنقدة الهرمية التراكمية قادرة على التقاط البنية الدلالية للنصوص بصورة أفضل نسبيًا، إلا أن الفارق في الأداء يبقى محدودًا، وهو ما يعكس تقارب الخوارزميتين عند الاعتماد على تمثيل النصوص باستخدام TF-IDF.

الجدول 3. مقارنة مع أداء نماذج KMeans وAgglomerative Clustering

Model	Accuracy	Precision	Recall	F1
KMeans	0.8031	0.7570	0.8031	0.7663
Agglomerative	0.8057	0.7620	0.8057	0.7701
Sentiment	0.8800	0.9800	0.7800	0.8700

عند مقارنة الأساليب القائمة على العنقدة مع النموذج الحالي المعتمد على إضافة ميزة مشاعر مستخرجة باستخدام تمثيلات سياقية عبر DistilBERT، يتضح أن النموذج الحالي حقق أعلى دقة تصنيف بلغت ٨٨,٠٠%، متفوقًا بشكل واضح على نماذج المشاعر المعتمدة على العنقدة غير الموجّهة. وبالمقارنة مع أفضل نموذج عنقدة (Agglomerative Clustering) الذي حقق دقة ٨٠,٥٧%، سجّل النموذج الحالي تحسنًا قدره ٧,٤٣% في الدقة، كما هو موضح في الشكل (٥).



الشكل٥. مقارنة أداء أفضل نموذج عنقدة مع نموذج المشاعر الإشرافي من حيث الدقة

يُعزى هذا التفوق إلى قدرة نموذج DistilBERT على تمثيل السياق الدلالي الكامل للنصوص والتقاط تدفّق المعنى داخل الشكوى الواحدة، مما يسمح باستخلاص إشارات شعورية أدق على مستوى كل شكوى. في حين تعتمد خوارزميات العنقدة على تمثيلات إحصائية عامة مثل TF-IDF وتجميع النصوص على مستوى العناقيد، مما قد يؤدي إلى فقدان بعض التفاصيل الدلالية الدقيقة داخل كل عنقود. وتؤكد هذه النتائج أن جودة ميزة المشاعر المستخرجة تلعب دورًا جوهريًا في تحسين أداء نماذج التنبؤ باعتراض المستهلك.

الاستنتاجات والتوصيات

وفق نتائج هذه الدراسة، نستنتج ما يلي:

١. تفوق نموذج الغابة العشوائية على شجرة القرار: أظهرت المقارنة بين الخوارزميتين على مجموعات الخصائص الثلاث (X1، X2، X3) تفوقًا واضحًا للغابة العشوائية، حيث حققت أعلى دقة ٨٨,٥٠% و–F1 score بلغت ٠,٨٧٣٣ على المجموعة الكاملة، مقابل أفضل دقة لشجرة القرار بلغت ٨١,٥٣%.

٢. أهمية الخصائص الجدولية والنصية معًا: بيّن تحليل الأهمية أن ميزة "موافقة المستهلك" كانت الأعلى تأثيرًا (٠,١١٤٥)، تلتها الميزات الزمنية وميزة "تحليل المشاعر" (٠,٠٦٩٧)، مما يؤكد على القيمة المضافة لدمج البيانات النصية والجدولية في تحسين التنبؤ.

٣. تأثير تبسيط النموذج: أدى حذف الميزات منخفضة الأهمية مثل "Timely response?" وحقول السنوات إلى تقليل التعقيد الحسابي دون التأثير سلبًا على الأداء، مما يعكس فعالية استراتيجيات اختيار الميزات.

٤. أهمية ضبط المعاملات لتحسين الأداء: أسفر استخدام Grid Search عن تحديد العدد الأمثل للأشجار عند ٢٦٠، محققًا متوسط دقة عبر التحقق المتقاطع ٩١,٣٢% ودقة اختبار بلغت ٨٧,٩٧%، مع Precision مرتفع (٩٧,٥٩%) و Recall أقل نسبيًا (٧٧,٨٥%)، مما يشير إلى قوة النموذج في تقليل الإيجابيات الكاذبة ووجود مجال لتحسين القدرة على اكتشاف جميع الحالات الإيجابية..

بناء على نتائج الدراسة، نوصي بما يلي:

- تحسين التوازن بين Precision وRecall من خلال معايرة عتبة القرار (Decision Threshold Calibration) أو استخدام استراتيجيات تعديل أوزان الفئات، بما يساهم في التقاط عدد أكبر من الحالات الإيجابية الحقيقية دون التضحية بالدقة.
- تجربة خوارزميات تعزيزية متقدمة مثل XGBoost وLightGBM التي تتميز بقدرتها على التعامل مع البيانات غير المتوازنة وتحقيق أداء أعلى في مهام التصنيف المعقدة.
- دمج التحليل الزمني للمشاعر من خلال تتبع تطور شعور العميل عبر سلسلة التفاعلات، بهدف التنبؤ المسبق بتغير مستوى الرضا والتدخل الاستباقي عند رصد مؤشرات انخفاضه.
- تطبيق الإطار المقترح في بيئات واقعية كنظم خدمة العملاء المباشرة أو أنظمة إدارة التذاكر، وقياس الأثر الفعلي على مؤشرات الأداء الرئيسية ((Key Performance Indicators (KPIs) مثل زمن حل المشكلة ونسبة الرضا بعد التحسينات.

المراجع

[1] Abdelhamid, M., & Desai, A. (2024). Balancing the scales: A comprehensive study on tackling class imbalance in binary classification. arXiv preprint arXiv:2409.19751.

[2] Alarifi, G., Rahman, M. F., & Hossain, M. S. (2023). Prediction and analysis of customer complaints using machine learning techniques. International Journal of E-Business Research (IJEBR), 19(1), 1-25.

[3] Arcipreste, B. M. (2023). Product Complaint Understanding using NLP Techniques (Master's thesis, Universidade do Porto (Portugal)).

[4] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[5] Buyukoz, B. (2023). Analyzing the generalizability of deep contextualized language representations for text classification. arXiv preprint arXiv:2303.12936.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[7] Consumer Financial Protection Bureau (CFPB). CFPB website, Available at: https://www.consumerfinance.gov/ (Accessed on 1/12/2024)

[8] Feldman, R., & Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.

[9] Gandhi, J. N., Guru, K. V., Kannan, A. R., Sudhan, R. A., Kumar, S. A., & Bharathvaj, M. (2025, May). Efficient Sentiment Classification using DistilBERT for Enhanced NLP Performance. In International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024) (pp. 1500-1511). Atlantis Press.

[10] González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

[11] Lakatos, R.; Bogacsovics, G.; Harangi, B.; Lakatos, I.; Tiba, A.; Tóth, J.; Szabó, M.; Hajdu, A. A Machine Learning-Based Pipeline for the Extraction of

Insights from Customer Reviews. Big Data Cogn. Comput. 2024, 8, 20. https://doi.org/10.3390/bdcc8030020

[12] Mikalef, P., Krogstie, J., Pappas, I. O., & Pavlou, P. A. (2021). Big Data and Business Analytics: A Research Agenda for Realizing Business Value. Information & Management, 58(3), 103407. https://doi.org/10.1016/j.im.2021.103407

[13] Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81–106. https://doi.org/10.1023/A:1022643204877

[14] Raju, S. V., Bolla, B. K., Nayak, D. K., & Kh, J. (2022, April). Topic modelling on consumer financial protection bureau data: An approach using BERT based embeddings. In 2022 IEEE 7th International Conference for Convergence in Technology (I2CT) (pp. 1–6). IEEE. https://doi.org/10.1109/I2CT54291.2022.9824476

[15] Sangaraju, V. R., Bolla, B. K., Nayak, D. K., & Kh, J. (2022). Topic modelling on consumer financial protection bureau data: an approach using BERT based embeddings. arXiv preprint arXiv:2205.07259.

[16] Vaishnav, D., Neethinayagam, M., Khaire, A., & Woo, J. (2024). Predictive Analysis of CFPB Consumer Complaints Using Machine Learning. arXiv preprint arXiv:2407.06399.

[17] Wright, M. N., & König, I. R. (2019). Splitting on categorical predictors in random forests. PeerJ, 7, e6339.

[18] Zaghloul, M., Barakat, S., & Rezk, A. (2024). Predicting E-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches. Journal of Retailing and Consumer Services, 79, 103865.

[19] Zhang, Y., & Wallace, B. C. (2017). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:1510.03820.