تصميم نظام هجين لتصنيف المنتجات من نصوص الشكاوى باستخدام DistilBERT و Linear SVC

يارا سلمان * د. كندة أبو قاسم**

(الإيداع: 3 كانون الأول 2025، القبول: 13 كانون الثاني 2026)

الملخّص:

يُعتبر تصنيف نصوص الشكاوى من المهام الأساسية لتحسين جودة الخدمات وتعزيز رضا العملاء من خلال فهم طبيعة المشكلات المطروحة وتحليل شكاوي العملاء تجاه الخدمات المقدمة. تهدف هذه الدراسة إلى تصميم نظام هجين لتصنيف المنتجات من نصوص الشكاوى ضمن مجموعة بيانات شكاوى المستهلكين المقدمة إلى مكتب الحماية المالية للمستهلك في الولايات المتحدة الأمريكية CFPB (Consumer Financial Protection Bureau)، مؤلف من DistilBERT للتمثيل وLinear SVC للتنبؤ، ثم مقارنة أداء النموذج مع أداء نموذج التقليدي TF-IDF. تم استخراج التمثيلات العددية للنصوص باستخدامDistilBERT ، وتم تدريب نموذج LinearSVC مع ضبط معلمات النموذج عبر تقنية Grid Search لتحسين الأداء.

أظهرت النتائج التجريبية أن نموذجDistilBERT مع LinearSVC قدم أداءً قوياً في دقة التصنيف ومقاييس الدقة النوعية والاستدعاء ومقياس F1 بمتوسط قيم بلغ 78%، مما يعكس قدرة النموذج على التقاط السمات الدلالية العميقة من النصوص . تؤكد الدراسة أهمية اختيار النموذج الأمثل اعتماداً على خصائص البيانات، طبيعة المهمة، والموارد الحسابية المتاحة، كما تُبرز الدراسة فعالية الدمج بين التمثيلات السياقية لنموذج DistilBERT والخوارزميات الخطية في مهام التصنيف، وتؤكد أهمية استخدام النماذج الهجينة التي تجمع بين القوة التعبيرية للنماذج العميقة وكفاءة نماذج التعلّم التقليدية لتحقيق أداء متوازن وموثوق في تطبيقات معالجة اللغة الطبيعية.

الكلمات المفتاحية: تصنيف النصوص، DistilBERT، LinearSVC، شكاوى المستهلكين، CFPB، Grid Search.

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

**أستاذ في قسم هندسة الحاسبات والتحكم الآلي– كلية الهندسة الميكانيكية والكهربائية، جامعة تشرين

Designing a Hybrid System for Product Classification from Complaint Texts Using DistilBERT for Representation and Linear SVC for Prediction

Yara Salman, **Kina Abou Kasem,**

(Received: 3 December 2025, Accepted: 13 January 2026)

ABSTRACT:

Classifying complaint texts is a fundamental task for improving service quality and enhancing customer satisfaction by enabling organizations to understand the nature of reported issues and analyze consumers' experiences with the services provided. This study aims to design a hybrid system for classifying product categories from complaint texts in the Consumer Financial Protection Bureau (CFPB) dataset, combining DistilBERT for text representation with Linear SVC for prediction, and to compare its performance with that of the traditional TF-IDF model. Numerical text embeddings were extracted using DistilBERT, and the LinearSVC model was trained with hyperparameter optimization performed through Grid Search to enhance classification performance.

The experimental results showed that the DistilBERT model combined with LinearSVC performed strongly in classification accuracy, qualitative accuracy, recall, and F1, with average values of 78%, reflecting the model's ability to capture deep semantic features from texts. The study underscores the importance of selecting the optimal model based on data characteristics, task nature, and available computational resources. It also highlights the effectiveness of combining contextual representations of the DistilBERT model with linear algorithms in classification tasks and emphasizes the importance of using hybrid models that combine the expressive power of deep models with the efficiency of traditional learning models to achieve balanced and reliable performance in natural language processing applications.

Keywords: text classification, DistilBERT, LinearSVC, consumer complaints, CFPB, Grid Search.

*PhD student in the Department of Computer Engineering and Automatic Control, Faculty of Mechanical and Electrical Engineering, Tishreen University, Syria.

***Professor in the Department of Computer Engineering and Automatic Control, Faculty of Mechanical and Electrical Engineering, Tishreen University, Syria

المقدمة

يُعتبر تصنيف النصوص من المهام الجوهرية في مجال معالجة اللغة الطبيعية والتنقيب عن النصوص، حيث يُستخدم لتحويل النصوص غير المهيكلة إلى تمثيلات قابلة للتحليل تُسهم في دعم اتخاذ القرار في مجالات متعددة مثل الأعمال، والرعاية الصحية، والخدمات المالية (Minaee et al., 2020). وتُعد شكاوى العملاء من أبرز أشكال النصوص التي تحظى باهتمام الباحثين وصانعي القرار، إذ توفر رؤى مباشرة عن جودة الخدمات المقدمة وتوقعات المستهلكين، كما أن تحليل هذه النصوص يساعد المؤسسات على تحديد أنماط متكررة من المشاكل وتعزيز استراتيجيات الاستجابة بما ينعكس إيجابًا على رضا العملاء (Wen et al., 2024; Song et al., 2024).

اعتمدت الأساليب التقليدية في تصنيف النصوص على تمثيلات إحصائية بسيطة مثل Bag–of–Words وTF–IDF، حيث يسمح الأخير بقياس أهمية الكلمات داخل النصوص عبر موازنة تكرارها المحلي والعالمي (Salton & Buckley, 1997). وقد أثبتت دراسات عديدة فعالية TF–IDF عند دمجه مع مصنفات خطية مثل SVM أو LinearSVC في تحقيق أداء قوي في مهام التصنيف، خاصة مع النصوص القصيرة أو مجموعات البيانات الصغيرة (Joachims, 1998). ورغم بساطته، ما يزال هذا الأسلوب منافسًا في بيئات تطبيقية متعددة، حيث بينت أبحاث لاحقة أنه يمكن أن يتفوق أحيانًا على النماذج العميقة عندما تكون البيانات محدودة أو الفئات غير متوازنة (Valmianski et al., 2019).

شهد المجال نقلة نوعية مع ظهور نماذج المحولات (Transformers) مثل BERT (Devlin et al., 2019)، التي أدخلت آليات الانتباه العميق لفهم السياق ثنائي الاتجاه، مما عزز دقة التمثيلات النصية بشكل كبير. إلا أن الحجم الضخم لـ BERT جعله مكلفًا من الناحية الحسابية، وهو ما دفع إلى تطوير نماذج مُصغرة مثل DistilBERT، الذي حافظ على جزء كبير من جودة التمثيل مع تقليل الحجم وزمن التدريب (Sanh et al., 2019). بالإضافة إلى ذلك، قدمت نماذج مشتقة من BERT مثل TinyBERT (Jiao et al., 2020) حلولاً متقدمة لأجهزة ذات موارد محدودة، مما يوسع نطاق استخدام هذه النماذج في التطبيقات العملية (Wang & Yang, 2020).

أظهرت دراسات مقارنة أن DistilBERT يتفوق غالبًا على الطرق التقليدية مثل TF–IDF، خصوصًا في المهام التي تتطلب فهمًا عميقًا للسياق (González–Carvajal & Garrido–Merchán, 2020؛ Dogança–Cetin, 2023). ومع ذلك، فإن هذا التفوق ليس مطلقًا، حيث أشارت بعض الدراسات إلى أن أداء DistilBERT قد يتراجع في مهام التصنيف متعددة الفئات أو عند التعامل مع بيانات محدودة الحجم (Ghadiridehkordi, 2024).

يُعتبر LinearSVC من بين المصنفات الأكثر شيوعًا في تصنيف النصوص عالية الأبعاد، نظرًا لفعاليته وكفاءته الحسابية، حيث أثبتت العديد من الدراسات أن أداءه قوي عند دمجه مع كل من التمثيلات التقليدية (مثل TF–IDF) والتمثيلات العميقة (مثل BERT وDistilBERT) (Jelodar et al., 2020). كما أن استخدام تقنيات تحسين الأداء مثل Grid Search يساعد في رفع دقة النماذج من خلال ضبط المعاملات المثلى (Pedregosa et al., 2011).

في سياق شكاوى المستهلك، استُخدمت بيانات مكتب الحماية المالية للمستهلك (CFPB) في الولايات المتحدة على نطاق واسع، حيث وفرت مجموعة بيانات غنية لدراسة أنماط الشكاوى وتحليلها باستخدام تقنيات تعلم الآلة. فقد بينت أبحاث حديثة أن معالجة نصوص هذه الشكاوى يُمكّن المؤسسات من التعرف على المنتجات أو القضايا الأكثر إثارة للمشاكل، مما يساعد على تحسين جودة الخدمات (Vaishnav et al., 2024; Raju et al., 2022). وقد دعمت هذه الدراسات أهمية دمج تقنيات التصنيف النصي المتقدمة مع قواعد بيانات العملاء من أجل بناء نماذج تنبؤية أكثر دقة.

بناء عليه، تُعتبر معالجة وتصنيف نصوص شكاوى العملاء من المهام الحيوية التي تساهم بشكل مباشر في تحسين جودة الخدمات وتعزيز رضا المستهلكين من خلال فهم المشاكل والاحتياجات الحقيقية للمستهلكين.

ورغم النجاحات التي حققتها النماذج العميقة مثل DistilBERT، لا تزال هناك فجوة بحثية واضحة فيما يتعلق بالمقارنات المباشرة مع النماذج التقليدية مثل TF-IDF + LinearSVC في مجال شكاوى المستهلك. فمعظم الدراسات ركزت إما على الأساليب التقليدية أو على الأساليب العميقة دون إجراء مقارنات منهجية، مما يبرز الحاجة لدراسات تقييمية شاملة تختبر أداء كلا النهجين في بيئات تطبيقية حقيقية، مع الأخذ بعين الاعتبار الكفاءة الحسابية، زمن التدريب، ودقة التصنيف (Alarifi et al., 2023; Lakatos et al., 2024).

يهدف هذا البحث إلى دراسة أداء نموذج DistilBERT مع مصنف LinearSVC، مع تحسين المعاملات عبر تقنية Grid Search، وتقديم تقييم شامل لمقاييس التصنيف المختلفة. كما يسعى البحث إلى مقارنة نتائج هذا النموذج مع أداء نموذج TF-IDF مع LinearSVC، وذلك بهدف توجيه اختيار النموذج الأمثل وفقاً لخصائص البيانات والموارد المتاحة في المؤسسات.

مواد وطرائق البحث

1. أدوات البحث

في هذه الدراسة، نعتمد نموذج DistilBERT، وهو نموذج تعلم عميق يستند إلى معمارية Transformer، لاستخراج تمثيلات نصية (embeddings) تأخذ بعين الاعتبار السياق الكامل للكلمات في النص، بدلاً من الاعتماد على التكرارات التقليدية للكلمات بغية استخراج تمثيلات النصوص. تم استخدام هذه التمثيلات كمدخلات لتدريب نموذج LinearSVC، مع تحسين أداء النموذج عبر ضبط المعاملات باستخدام تقنية Grid Search.

تم تطبيق هذا النهج على مجموعة بيانات شكاوى المستهلكين المقدمة إلى مكتب الحماية المالية للمستهلك (Consumer Financial Protection Bureau (CFPB)، حيث تتكون كل شكوى من سمات (ميزات) 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 Kaggle وموقع مكتب حماية المستهلك المالي (CFPB) على الانترنت.

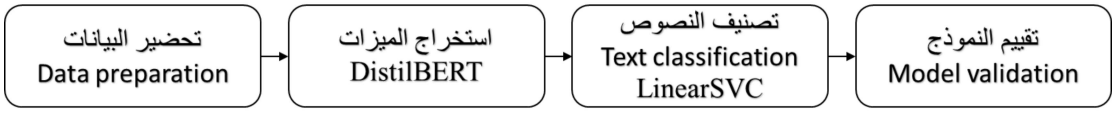
تم في هذا البحث العمل على منصة Google Colab لإجراء المحاكاة من حيث تصميم واختبار النظام المقترح، باستخدام نسخة Python 3، مع ذاكرة وصول عشوائي 12.7 GB ومعالج رسوميات NVIDIA Tesla T4 بذاكرة 15.3 GB وحجم تخزين GB78.2 .

بالإضافة إلى ذلك، تم استخدام مكتبات برمجية متخصصة منها:

- ✓ مكتبة Hugging Face Transformers لتطبيق نموذج DistilBERT،
- ✓ مكتبة Scikit-learn لتدريب نموذج LinearSVC وضبط المعاملات،
- ✓ مكتبات معالجة النصوص مثل NLTK وSpaCy لتنظيف وتحليل البيانات النصية،
- ✓ مكتبة Pandas لتنظيم وتحليل البيانات.

2. 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 رقم (1) : مخطط مراحل العمل

2-1. تحميل مجموعة البيانات (Dataset Loading)

حيث قمنا بتحميل مجموعة بيانات شكاوى المستهلكين من CFPB للفترة كانون الثاني 2025 حتى آذار 2025.

2-2. المعالجة الأولية للبيانات (Data Preprocessing)

وقد تمت هذه المرحلة على عدة خطوات؛ هي:

1. تنظيف البيانات Data Cleaning: وتتضمن إزالة السجلات الفارغة أو التي تحتوي على بيانات ناقصة في الأعمدة الأساسية، إزالة التكرارات لضمان جودة البيانات، تصحيح تنسيق النصوص، وتحويل جميع الحروف إلى أحرف صغيرة (Lowercasing).

2. تطبيع النص Text Normalization: تحويل النص إلى أحرف صغيرة لضمان التوحيد في المعالجة.

إزالة علامات الترقيم والرموز الخاصة التي لا تحمل معلومات دلالية.

بعد الانتهاء من خطوات المعالجة الأولية المذكورة أعلاه، تم الحصول على مجموعة بيانات نهائية مكوّنة من 5,433 سجلًا صالحًا للاستخدام .وقد شكّلت هذه البيانات الأساس الذي تم الاعتماد عليه في جميع التجارب اللاحقة.

2-3. تقسيم البيانات (Data Splitting)

تم تقسيم البيانات إلى جزءين: 75% تدريب، 25% اختبار.

2-4. استخراج الميزات (Feature Extraction)

تم استخراج الميزات النصية باستخدام نموذج DistilBERT، الذي يُعد نسخة مُصغرة ومُحسّنة من نموذج BERT القائم على تقنية الـ Transformer، وتتضمن خطوات استخراج الميزات ما يلي:

1) التقسيم إلى رموز (Tokenization):

يتم تقسيم النصوص إلى وحدات أصغر تُعرف بـ "الرموز tokens"، والتي يمكن أن تكون كلمات كاملة أو أجزاء من الكلمات (subwords). يستخدم DistilBERT أداة Tokenizer خاصة تعتمد على تقنية WordPiece لتقسيم النصوص، مما يساعد في التعامل مع الكلمات غير الشائعة والكلمات الجديدة.

2) إضافة رموز التحكم (Special Tokens):

يتم إدخال رموز خاصة مثل [CLS] التي تمثل بداية النص، و[SEP] لفصل الجمل أو الفقرات، وذلك لتوجيه النموذج نحو فهم البنية العامة للنص.

3) تضمين المواضع (Positional Embeddings):

لأن نماذج Transformer لا تفهم ترتيب الكلمات تلقائياً، تُضاف معلومات الموضع لكل رمز بغية تعريف ترتيب الكلمات في النص.

4) تمثيل النص العددي (Embedding Extraction):

يقوم النموذج بتمثيل كل رمز كمتجه عددي عالي الأبعاد يعكس السياق والمعنى الدلالي، عبر الطبقات المتعددة للنموذج.

5) التجميع (Pooling)

لتلخيص النص الكامل إلى تمثيل عددي موحد (embedding)، يتم استخدام طرق مثل التجميع المتوسط (mean pooling) على تمثيلات الرموز للحصول على متجه واحد لكل نص.

6) إخراج تمثيلات الميزات:

بعد الخطوات السابقة، يُنتج نموذج DistilBERT تمثيلات عددية (vectors) لكل نص يمكن استخدامها مباشرة كمدخلات لنموذج التصنيف مثل LinearSVC.

2-5. إنشاء الموديل SVM

تم استخدام خوارزمية Linear Support Vector Classifier (LinearSVC)، وهي نموذج شائع في تصنيف النصوص لما يتميز بالكفاءة والسرعة في التعامل مع البيانات ذات الأبعاد العالية. في هذا البحث، استُخدمت التمثيلات العددية (embeddings) المستخرجة من نموذج DistilBERT كنقاط إدخال إلى نموذج LinearSVC.

ولتحسين أداء نموذج LinearSVC، تم تطبيق تقنية البحث الشبكي (Grid Search) لضبط معاملات النموذج الأساسية، وخاصة معامل التنظيم C، الذي يتحكم في توازن النموذج بين تعقيد البيانات واحتمالية الخطأ. استخدمنا GridSearchCV مع تقاطع بيانات ثلاثي الطيات (3-fold cross-validation) لاختيار أفضل قيمة لمعامل **C** من بين مجموعة قيم مدروسة [0.01,0.1,1,10].

2-6. تقييم النموذج (Model Evaluation)

بعد الانتهاء من تدريب النموذج وضبط معاملاته، تم تقييم أداء النموذج من خلال حساب مؤشرات الأداء: الدقة (Accuracy)، الدقة النوعية (Precision)، الاستدعاء (Recall)، ومقياس F1 (F1-score)، كما تم رسم مصفوفة الالتباس (Confusion Matrix).

2-6. اختبار سيناريوهات التنبؤ (Prediction Scenarios)

هدفت هذه الدراسة إلى استخراج استنتاجات تتعلق بطبيعة شكاوى المستهلكين وتحديد المجالات المحتملة للتحسين في المنتجات والخدمات المالية، وذلك من خلال تحليل بيانات الشكاوى باستخدام تمثيلات نصية سياقية مستخرجة من نموذج DistilBERT، تليها عملية تصنيف باستخدام خوارزمية LinearSVC، وذلك على محتوى الشكاوى المسجّلة في قاعدة بيانات مكتب الحماية المالية للمستهلكين CFPB.

السيناريو الأول: التنبؤ بالمنتج قيد الشكوى كما هو موضح في الشكل (2).

```
[68] new_complaint = """The bank obtained the property through a foreclosure.
     embedding = get_single_embedding(new_complaint)
     predicted_product = loaded_model.predict(embedding)[0]
     print(f"[{get_product_name(predicted_product)}]\n")

[Mortgage]
```

الشكل رقم (2): السيناريو الأول

عند تطبيق السيناريو الأول وتوليد التمثيلات السياقية للنصوص باستخدام نموذج DistilBERT، ثم تمريرها إلى نموذج SVMللتصنيف النصي، نجح النموذج في التنبؤ بالمنتج المرتبط بالشكوى، حيث تم تحديده بشكل صحيح على أنه "Mortgage".

السيناريو الثاني: كتابة الموضوع الفرعي للشكوى كما هو موضح في الشكل (3).

```
[69] new_complaint = "Impersonated an attorney or official"
     embedding = get_single_embedding(new_complaint)
     predicted_product = loaded_model.predict(embedding)[0]
     print(f"[{get_product_name(predicted_product)}]\n")

[Debt collection]
```

الشكل رقم (3): السيناريو الثاني

عند تطبيق السيناريو الثاني وتوليد التمثيلات العددية للنصوص باستخدام نموذج DistilBERT، ثم إدخالها إلى نموذج التصنيف SVM، تمكن النموذج من تصنيف الموضوع الفرعي للشكوى بنجاح، حيث تم تحديد التصنيف الصحيح على أنه "Debt collection".

السيناريو الثالث: إدخال الموضوع الفرعي للشكوى كما هو موضح في الشكل (4).

```
[70] new_complaint = "Not given enough info to verify debt"
     embedding = get_single_embedding(new_complaint)
     predicted_product = loaded_model.predict(embedding)[0]
     print(f"[{get_product_name(predicted_product)}]\n")

[Credit reporting]
```

الشكل رقم (4): السيناريو الثالث

عند تطبيق السيناريو الثالث، وبعد إدخال الموضوع الفرعي للشكوى وتمثيله باستخدام نموذج DistilBERT، ثم تمريره إلى نموذج التصنيف SVM، نجح النموذج في تصنيف البيانات بدقة وفقاً للموضوع الفرعي المحدد.

السيناريو الرابع: اكتشاف المنتج قيد الشكوى وموضوع المشكلة كما هو موضح في الشكل (5).

```
[71] new_complaint = "PayPal is randomly sending PayPal account holders debit cards without permission
     embedding = get_single_embedding(new_complaint)
     predicted_product = loaded_model.predict(embedding)[0]
     print(f"[{get_product_name(predicted_product)}]\n")

[Credit card]
```

الشكل رقم (5): .السيناريو الرابع

عند تطبيق السيناريو الرابع، وبعد تمثيل بيانات الشكوى باستخدام نموذج DistilBERT وتمريرها إلى نموذج التصنيفSVM ، نجح النموذج في التنبؤ بالمنتج قيد الشكوى بدقة، حيث تم تحديده على أنه"Credit card" .

2-7. المقارنة مع TF-IDF

تم إجراء تجربة على مجموعة البيانات نفسها، ومقارنة نتائج النموذج المعتمد في الدراسة /DistilBERT وخوارزمية التصنيف LinearSVC/ مع نتائج النموذج التقليدي الذي يستخدم تمثيل النصوص بطريقة TF-IDF مع خوارزمية التصنيف SVM، حيث تم تطبيق خوارزمية التصنيف النصي ضمن موديل SVM للتنقيب عن البيانات ضمن قواعد البيانات الضخمة في قاعدة البيانات CFPB بنفس شروط النموذج المدروس.

النتائج والمناقشة

أظهر النموذج المعتمد في هذه الدراسة، والمبني على توليد تمثيلات نصية سياقية باستخدام نموذجDistilBERT متبوعاً بخوارزمية التصنيف LinearSVC، أداءً مقبولاً في تصنيف شكاوى المستهلكين ضمن قاعدة بيانات CFPB.

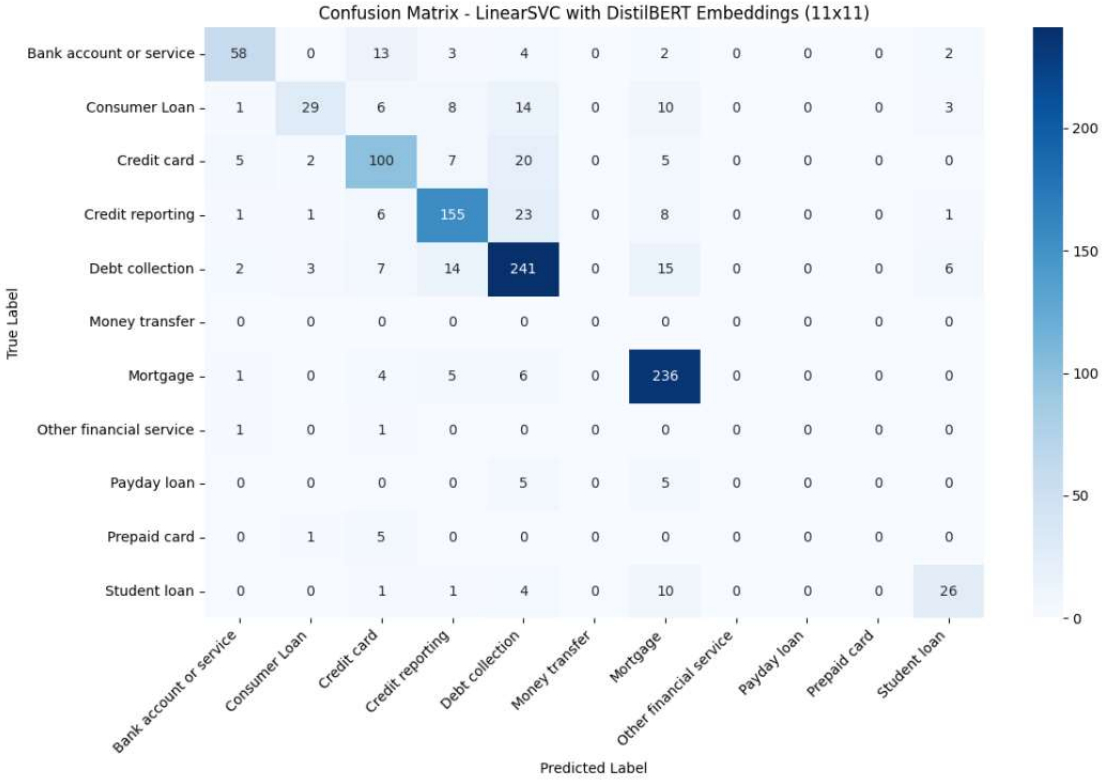
بلغت دقة النموذج (Accuracy) ما نسبته 0.7774، مما يشير إلى أن حوالي 77.7% من العينات تم تصنيفها بشكل صحيح. كما حقق النموذج دقة نوعية (Precision) قدرها 0.7654، واستدعاء (Recall) بنسبة 0.7774، وهو ما يدل على قدرة النموذج على استرجاع نسبة كبيرة من الفئات الصحيحة.

أما مقياس F1 (F1-score)، الذي يمثل المتوسط التوافقي للدقة والاستدعاء، فقد بلغ 0.7651، ما يعكس توازناً جيداً بين مؤشري الاسترجاع والدقة.

الجدول 1. نتائج تقييم النموذج

الدقة Accuracy	0.7774
الدقة النوعية Precision	0.7654
الاستدعاء Recall	0.7774
مقياس F1 score	0.7651

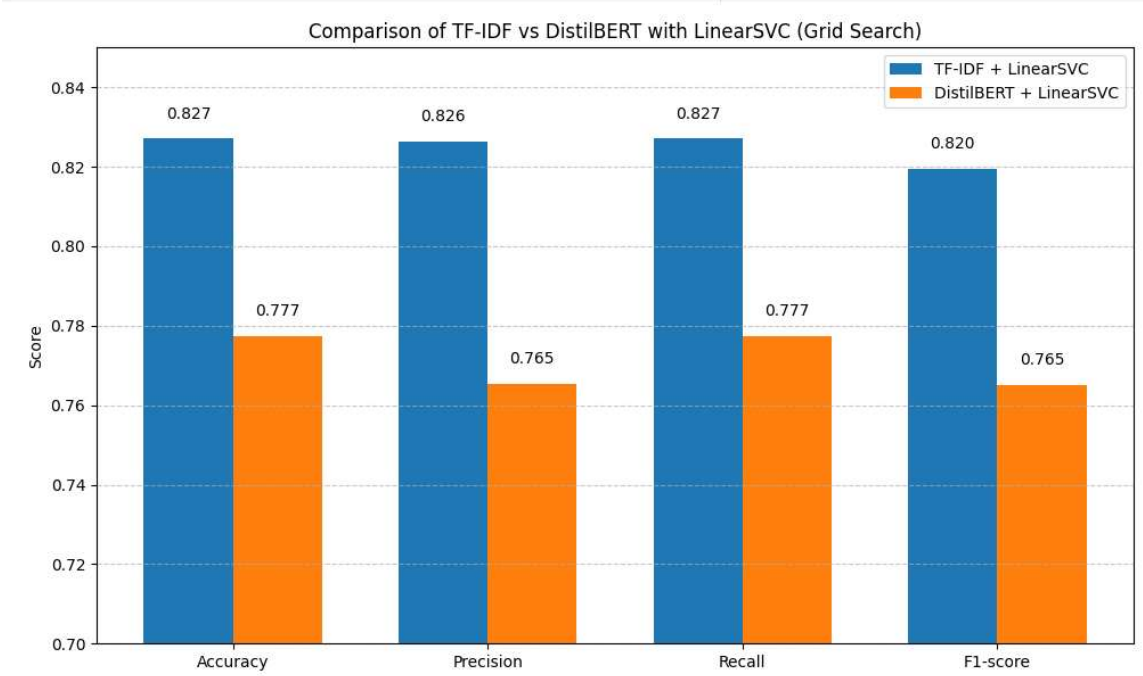
يبين الشكل (6) مصفوفة الارتباك التي تم حسابها من خلال اختبارات الموديل.



الشكل رقم(6): مصفوفة الارتباك

تشير هذه النتائج إلى أن استخدام التمثيلات النصية العميقة المستخرجة من DistilBERT يساهم في تحسين فهم النموذج للسياق اللغوي للشكاوى، وهو ما انعكس إيجاباً على دقة التصنيف. ومع ذلك، يبقى المجال مفتوحاً لتحقيق أداء أعلى، خصوصاً في حالات التصنيف متعددة الفئات التي تتسم بتوزيع غير متوازن للفئات وتنوع دلالي كبير في النصوص. وقد دعمت دراسة Ghadiridehkordi (2024) هذه النتيجة، حيث أظهرت أن دمج DistilBERT مع مصنفات خطية مثل SVM يحسن دقة التصنيف الكلي لشكاوى العملاء، لكنه ما زال يواجه تحديات عند التعامل مع تعدد الفئات وتنوعها الدلالي..

لدى مقارنة النموذج المعتمد في هذه الدراسة مع نتائج نموذج TD–IDF أظهرت النتائج أن نموذج TF–IDF تفوّق من حيث الأداء الكمي كما هو موضح في الشكل (7).



الشكل رقم (7): نتائج مقارنة نموذجي Distilbert وTF–IDF

يبين الجدول (2) مقارنة دقة نموذجي Distilbert وTF–IDF، حيث حقق نموذج TF–IDF دقة (Accuracy) بلغت 0.8272 مقابل 0.7774 في نموذجDistilBERT ، كما تفوّق أيضاً في مؤشرات الدقة النوعية (Precision= 0.8264) والاستدعاء (Recall= 0.8272)، ومقياس F1 (F1–score= 0.8196) مقارنة بالقيم المقابلة في نموذج DistilBERTوالتي بلغت (0.7654، 0.7774، 0.7651) على التوالي.

الجدول 2. مقارنة دقة نموذجي Distilbirt وTD–IDF

Metric	TF-IDF	DistilBERT	Difference
Accuracy	0.8272	0.7774	0.0498
Precision	0.8264	0.7654	0.0610
Recall	0.8272	0.7774	0.0498
F1-Score	0.8196	0.7651	0.0545

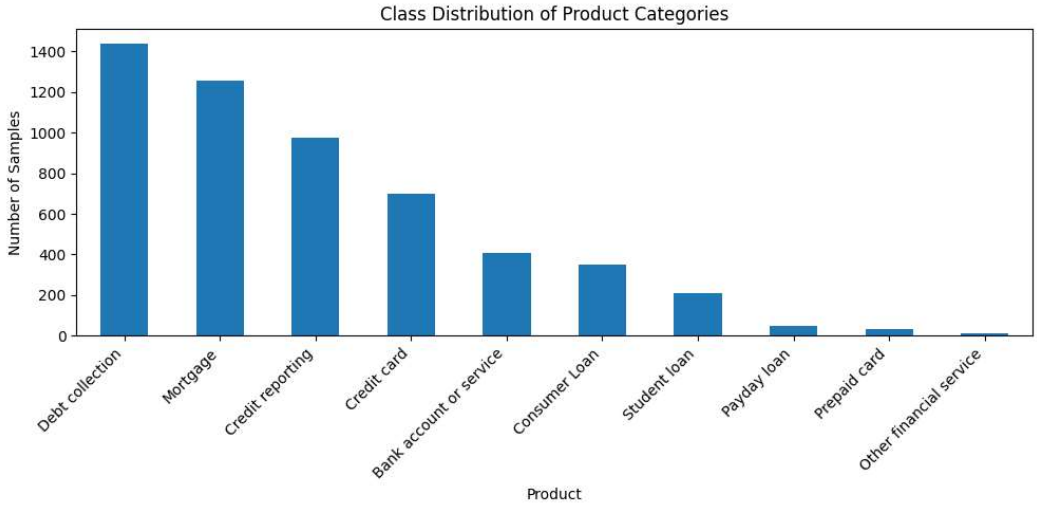
تتسق هذه النتائج مع ما ذكره Minaee وآخرون (2020)، حيث أوضحوا أن النماذج التقليدية مثل TF–IDF ما زالت تحقق أداءً تنافسيًا، خصوصًا في المهام البسيطة أو عند التعامل مع مجموعات بيانات محدودة الحجم أو قليلة التعقيد. كما بينت دراسة González–Carvajal & Garrido–Merchán (2020) أن تفوق نماذج المحولات مثل BERT وDistilBERT ليس مطلقًا، إذ يمكن أن تتراجع دقتها في حالات التصنيف متعدد الفئات أو عند وجود اختلال في توازن الفئات، وهي خصائص تنطبق جزئيًا على بيانات هذه الدراسة. وفي المقابل، أظهرت دراسات أخرى مثل Ghadiridehkordi (2024) أن DistilBERT يتفوق على الأساليب التقليدية عند توافر بيانات أكبر وأكثر تنوعًا، مما يشير إلى أن اختيار النموذج الأمثل يظل معتمدًا على خصائص البيانات وطبيعة المهمة البحثية.

على الرغم من أن نموذج DistilBERT يُعدّ أكثر تطوراً من حيث فهم السياق اللغوي العميق للنصوص، إلا أن النموذج التقليدي المعتمد على TF-IDF قدّم نتائج أفضل في هذا التطبيق العملي على بيانات شكاوى المستهلكين. ويُعزى ذلك إلى طبيعة النصوص القصيرة والمباشرة في البيانات، والتي قد لا تستفيد كثيراً من التمثيلات السياقية العميقة، مما يجعل تمثيل TF-IDF كافياً في هذه الحالة.

توضح هذه المقارنة أهمية اختيار طريقة تمثيل النصوص بعناية حسب طبيعة البيانات والنموذج المستهدف، وتفتح المجال لاستخدام نماذج هجينة أو تقنيات Fine-Tuning على DistilBERT لتحقيق أداء أعلى.

إضافةً إلى فروقات الأداء، أظهرت المقارنة بين النموذجين اختلافًا واضحًا من حيث زمن التدريب والكلفة الحسابية. فقد تميز نموذج TF-IDF مع LinearSVC بزمن تنفيذ قصير واعتماده على المعالج المركزي فقط، مما يجعله خيارًا عمليًا في البيئات ذات الموارد المحدودة. في المقابل، تطلّب نموذج DistilBERT زمنًا أطول لاستخراج التمثيلات السياقية، مع حاجة إلى موارد حسابية أعلى رغم عدم تطبيق Fine-tuning. تؤكد هذه النتائج أن اختيار النموذج الأمثل لا يعتمد فقط على دقة التصنيف، بل يتأثر أيضًا بطبيعة البيانات والموارد الحسابية المتاحة.

ويُظهر تحليل توزيع فئات المنتجات في مجموعة البيانات وجود عدم توازن واضح، حيث تتركز غالبية السجلات في عدد محدود من الفئات مقابل تمثيل ضعيف لفئات أخرى. وقد يؤثر هذا التوزيع غير المتوازن على أداء النماذج، ويسهم جزئيًا في تفسير الفروق الملحوظة بين أداء النماذج المدروسة.



الشكل رقم (8): التوزيع العددي لفئات المنتجات بعد المعالجة الأولية

الاستنتاجات

في هذا البحث، تم استكشاف استخدام نموذج DistilBERT لاستخراج التمثيلات العددية من النصوص، وتطبيق مصنف LinearSVC عليها بعد تطبيق تقنية البحث الشبكي (Grid Search) لضبط معاملات النموذج الأساسية، مما أدى إلى تحسين أداء نموذج LinearSVC.

بلغت دقة النموذج (Accuracy) ما نسبته 0.7774، مما يشير إلى أن حوالي 77.7% من العينات تم تصنيفها بشكل صحيح. كما حقق النموذج دقة نوعية (Precision) قدرها 0.7654، واستدعاء (Recall) بنسبة 0.7774، وهو ما يدل على قدرة النموذج على استرجاع نسبة كبيرة من الفئات الصحيحة.

أما مقياس F1 (F1-score)، الذي يمثل المتوسط التوافقي للدقة والاستدعاء، فقد بلغ 0.7651، ما يعكس توازناً جيداً بين مؤشري الاسترجاع والدقة.

أظهرت النتائج أن هذه الطريقة حققت دقة وصلت حتى 78% تقريباً في مؤشرات الأداء مما يؤكد فاعليتها في معالجة نصوص شكاوى المستهلكين ذات الطابع الرسمي والمحدد. لكنه بالمقابل، يترك الباب مفتوحاً لمزيد من العمل من أجل تحسين الأداء مثل عمليات Fine-tuning واستخدام بيانات أكبر لتحقيق تفوقها.

قد يكون تفوق TF-IDF في بعض المقاييس بسبب بساطة النصوص وتكرار الكلمات الدالة عليها، مما يجعل التمثيل القائم على التردد فعالاً جداً. أما DistilBERT، كنموذج يعتمد على التعلم العميق وفهم السياق، فقد يستفيد أكثر في حالات النصوص المعقدة أو المتنوعة التي تتطلب فهماً أعمق.

بناءً على النتائج، نوصي بما يلي:

- استكشاف طرق Fine-tuning لنماذج DistilBERT وغيرها من نماذج التعلم العميق لتحسين أدائها في تطبيقات تصنيف النصوص.
- إجراء دراسات مستقبلية باستخدام مجموعات بيانات أكبر وأكثر تنوعاً، لتقييم إمكانيات النماذج العميقة في استخراج ميزات السياق والمعنى العميق.
- استثمار تقنيات تفسير النماذج لتسهيل فهم وتفسير نتائج نماذج التعلم العميق في تطبيقات تصنيف النصوص.

المراجع

1. 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.
2. Consumer Financial Protection Bureau (CFPB). CFPB website, Available at: https://www.consumerfinance.gov/ (Accessed on 1/3/2024)
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.
4. Doganca-Cetin, S. (2023). Exploring the effectiveness of BERT using different pooling strategies on SVM for news classification. Doctoral dissertation, Tilburg university.
5. Ghadiridehkordi, A. (2024). Leveraging sentiment analysis via text mining to improve customer satisfaction: A DistilBERT and SVM application in banking. University of Birmingham. DOI: 10.1108/IJBM-05-2024-0323
6. González-Carvajal, S., & Garrido-Merchán, E. (2020). Comparing BERT against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*.
7. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78(11), 15169-15211.

8. Jiao, X., Sun, Z., Shao, J., & Huo, J. (2020). TinyBERT: Distilling BERT for natural language understanding. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 419–428).
9. Joachims, T. (1998, April). 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.
10. Lakatos, R., Bogacsovics, G., Harangi, B., Lakatos, I., Tiba, A., Tóth, J., ... & Hajdu, A. (2024). A machine learning-based pipeline for the extraction of insights from customer reviews. *Big Data and Cognitive Computing*, 8(3), 20.
11. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:*2004.03705.
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.
13. 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.
14. Salton, G., & Buckley, C. (1997). Term-weighting approaches in automatic text retrieval. *Readings in information retrieval*, 323, 328.
15. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
16. Song, W., Rong, W., & Tang, Y. (2024). Quantifying risk of service failure in customer complaints: A textual analysis-based approach. *Advanced Engineering Informatics*, 60, 102377.
17. Vaishnav, D., Neethinayagam, M., Khaire, A., & Woo, J. (2024). Predictive Analysis of CFPB Consumer Complaints Using Machine Learning. *arXiv preprint arXiv*:2407.06399.
18. Wang, S., & Yang, Y. (2020). A survey of BERT-based models for text classification. *arXiv preprint arXiv:2004.03705*.
19. Wen, Z., Chen, Y., Liu, H., & Liang, Z. (2024). Text Mining Based Approach for Customer Sentiment and Product Competitiveness Using Composite Online Review Data. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(3), 1776-1792.