تصميم نموذج هجين يدمج العنقدة وتحليل المشاعر للتنبؤ باعتراض المستهلك لمجموعة بيانات CFPB

يارا سلمان* د. كندة أبو قاسم**

(الإيداع: 13 كانون الأول 2026 ، القبول: 19 شباط 2026)

الملخّص:

تشكّل نصوص شكاوى المستهلكين موردًا غنيًا لتحليل تجارب العملاء وتقييم جودة الخدمات المقدّمة في القطاع المالي، لما تحمله من مؤشرات مباشرة على مواطن الخلل ومصادر عدم الرضا. تهدف هذه الدراسة إلى تطوير إطار يعتمد على استخراج ميزات مشاعر قائمة على خوارزميات عنقدة غير موجّهة من نصوص شكاوى المستهلكين، ومن ثم دمج هذه الميزات ضمن نموذج تصنيف موجّه للتنبؤ بحالات اعتراض المستهلك، وذلك باستخدام مجموعة بيانات شكاوى المستهلكين التابعة لمكتب الحماية المالية للمستهلك في الولايات المتحدة الأمريكية (CFPB).

تم تمثيل نصوص الشكاوى باستخدام أسلوب TF–IDF، ثم تطبيق خوارزميتي KMeans والعنقدة الهرمية التراكمية (Agglomerative Clustering) لاستخلاص تمثيلات عنقودية تعكس البنية الدلالية للنصوص. بعد ذلك، جرى ربط كل عنقود بمؤشر مشاعر مستخرج باستخدام TextBlob، وتحويل ناتج العنقدة إلى ميزة عددية أُدمجت ضمن نموذج Random Forest للتصنيف. أظهرت النتائج التجريبية أن النموذج المعتمد على KMeans حقق دقة تصنيف بلغت 80.31%، في حين حقق النموذج المعتمد على Agglomerative Clustering دقة بلغت 80.57%. ونظرًا لعدم توازن البيانات بين فئتي الاعتراض وعدم الاعتراض، تم اعتماد مقياس F1–weighted لتقييم الأداء بشكل أكثر واقعية، حيث بلغ حوالي 77% في كلا النهجين. تؤكد النتائج فاعلية دمج ميزات المشاعر المعتمدة على العنقدة ضمن نماذج تصنيف موجّهة لمعالجة شكاوى المستهلكين في بيئات بيانات غير متوازنة. تتمثل المساهمة البحثية لهذا العمل في اقتراح إطار غير موجّه لاستخلاص ميزات مشاعر أكثر استقرارًا من النصوص غير المهيكلة، ودمجها ضمن نموذج تصنيف موجّه للتنبؤ باعتراض المستهلك دون الحاجة إلى بيانات موسومة مسبقًا.

الكلمات المفتاحية: تحليل شكاوى المستهلكين، تحليل المشاعر، العنقدة غير الموجّهة، KMeans، Agglomerative Clustering، Random Forest، البيانات غير المتوازنة، CFPB.

*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.

Design of a Hybrid Model Integrating Clustering and Sentiment Analysis for Consumer Dispute Prediction Using the CFPB Dataset

Yara Salman * Kina Abou Kasem,**

(Received: 13 January 2025, Accepted: 19 February 2026)

ABSTRACT:

Consumer complaint texts constitute a rich source for analyzing customer experiences and assessing service quality in the financial sector, as they provide direct indicators of service deficiencies and sources of dissatisfaction. This study aims to develop a framework that extracts sentiment features based on unsupervised clustering algorithms from consumer complaint texts and integrates these features into a supervised classification model to predict consumer dispute cases, using the Consumer Financial Protection Bureau (CFPB) complaint dataset from the United States.

Complaint texts were represented using the TF–IDF approach, after which KMeans and hierarchical Agglomerative Clustering were applied to extract cluster–based representations reflecting the underlying semantic structure of the texts. Each cluster was then associated with a sentiment indicator derived using TextBlob, and the resulting cluster–based sentiment was transformed into a numerical feature integrated into a Random Forest classifier. Experimental results showed that the KMeans–based model achieved a classification accuracy of 80.31%, while the Agglomerative Clustering–based model achieved an accuracy of 80.57%. Given the class imbalance between disputed and non–disputed cases, the F1–weighted metric was adopted to provide a more realistic performance evaluation, reaching approximately 77% for both approaches. The results confirm the effectiveness of integrating clustering–based sentiment features into supervised classification models for analyzing consumer complaints in imbalanced data environments. The research contribution of this work lies in proposing an unsupervised framework for extracting more stable sentiment features from unstructured texts and integrating them into a supervised classification model to predict consumer disputes without the need for labeled data..

Keywords: Consumer complaint analysis, sentiment analysis, unsupervised clustering, KMeans, Agglomerative Clustering, Random Forest, imbalanced data, CFPB.

---

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

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

المقدمة:

أصبحت البيانات النصية المولّدة من تفاعلات المستخدمين مع المؤسسات المالية مصدرًا غنيًا لاستخلاص المعرفة وبناء أنظمة دعم القرار، ولا سيما في مجال تحليل شكاوى المستهلكين. إذ تمثّل هذه الشكاوى انعكاسًا مباشرًا لتجربة العميل، كما تسهم دراستها في الكشف عن مصادر الخلل في الخدمات وتقييم مخاطر الفشل التشغيلي (Song et al., 2024). وفي هذا السياق، حظيت قاعدة بيانات شكاوى المستهلكين التابعة لمكتب الحماية المالية للمستهلك (CFPB) باهتمام واسع، نظرًا لحجمها الكبير وتنوّع نصوصها وارتباطها المباشر بالقطاع المالي (Consumer Financial Protection Bureau, 2024).

تعتمد معظم الدراسات السابقة في تحليل شكاوى CFPB على نماذج تصنيف موجّهة تهدف إلى التنبؤ بسلوك المستهلك أو تصنيف نوع الشكوى بالاعتماد على تمثيلات نصية تقليدية أو مشاعر مستخرجة بأساليب مباشرة (Vaishnav et al., 2024; Baskaran, 2023). ورغم فعالية هذه النماذج في العديد من السيناريوهات، إلا أن أداءها يتأثر بشكل واضح بطبيعة البيانات غير المتوازنة، حيث تميل فئة "عدم الاعتراض" إلى الهيمنة على مجموعة البيانات مقارنةً بفئة "الاعتراض"، مما يجعل الاعتماد على مقياس الدقة (Accuracy) وحده غير كافٍ لتقييم الأداء الحقيقي للنماذج (He & Garcia, 2009).

لمواجهة هذه الإشكالية، شددت دراسات حديثة على أهمية استخدام مقاييس تقييم أكثر ملاءمة للبيانات غير المتوازنة، مثل مقياس F1–weighted ومخططات Precision–Recall، التي تعكس أداء النموذج على كامل البيانات دون الانحياز للفئة الأكبر (Saito & Rehmsmeier, 2015). كما أظهرت الأبحاث أن تقنيات المعالجة المسبقة أو إعادة التوازن، مثل SMOTE، قد تسهم في تحسين أداء النماذج، إلا أنها لا تعالج جوهريًا مسألة تمثيل البنية الداخلية للنصوص (Chawla et al., 2002).

في موازاة ذلك، تُعد أساليب تحليل المشاعر التقليدية المعتمدة على القواميس اللغوية، مثل VADER وTextBlob، من أكثر الأساليب استخدامًا نظرًا لبساطتها وسهولة دمجها مع نماذج التصنيف (Hutto & Gilbert, 2014; Loria, 2020). غير أن هذه الأساليب تعتمد على قواعد ثابتة وقد تفشل في التقاط الأنماط الدلالية المعقّدة أو السياق الضمني للنصوص الرسمية، كما هو الحال في شكاوى CFPB.

ضمن هذا الإطار، تبرز خوارزميات العنقدة غير الموجّهة كبديل واعد لاستخلاص تمثيلات دلالية قائمة على البنية الكامنة للبيانات النصية دون الحاجة إلى تسميات مسبقة. وتُعد خوارزمية KMeans من أكثر خوارزميات العنقدة شيوعًا نظرًا لبساطتها وكفاءتها الحسابية (MacQueen, 1967; Jain, 2010)، في حين توفر خوارزميات العنقدة الهرمية، مثل Agglomerative Clustering، قدرة أفضل على تمثيل العلاقات المتدرجة بين النصوص واكتشاف التجمعات ذات البنية غير الخطية (Müllner, 2011). وقد بينت دراسات سابقة أن دمج نواتج العنقدة مع مهام تحليل المشاعر يمكن أن يساهم في تحسين جودة الميزات المستخرجة، خاصة في البيئات غير الموجّهة أو ضعيفة التوسيم (Steinbach et al., 2000; Bibi et al., 2022).

بناءً على ما سبق، يهدف هذا البحث إلى اقتراح إطار يعتمد على استخراج ميزات مشاعر مبنية على خوارزميات العنقدة غير الموجّهة (KMeans وAgglomerative Clustering) من نصوص شكاوى المستهلكين، ومن ثم دمج هذه الميزات ضمن نموذج تصنيف موجّه باستخدام خوارزمية Random Forest (Breiman, 2001). كما يسعى البحث إلى إجراء مقارنة تجريبية بين هذا النهج القائم على العنقدة ونهج تحليل المشاعر التقليدي، مع تقييم الأداء باستخدام مقياس F1–

weighted الملائم لطبيعة البيانات غير المتوازنة، بهدف تقديم معالجة أكثر واقعية ودقة لمشكلة التنبؤ باعتراض المستهلك.

دراسات ذات صلة (Related Work)

تناولت العديد من الدراسات السابقة تحليل شكاوى المستهلكين، ولا سيما بيانات شكاوى CFPB، باستخدام تقنيات تعلم الآلة والتنقيب النصي بهدف التنبؤ بسلوك المستهلك أو تصنيف الشكاوى. فقد اعتمد Vaishnav وآخرون (2024) و Baskaran (2023) على نماذج تصنيف موجّهة قائمة على تمثيلات نصية تقليدية وميزات مشتقة من النص، وبيّنوا أن هذه النماذج قادرة على تحقيق أداء مقبول، إلا أن فعاليتها تتأثر بشكل واضح بعدم توازن الفئات في البيانات. من جهة أخرى، ركّزت دراسات مثل Song وآخرون (2024) و Wen وآخرون (2024) على تحليل النصوص لاستخلاص مؤشرات تتعلق بمخاطر فشل الخدمة وتجربة العملاء، وأكدت أهمية التمثيل الدلالي للنصوص في فهم شكاوى المستهلكين.

في سياق تحليل المشاعر، استخدمت العديد من الأبحاث أساليب قائمة على القواميس اللغوية مثل VADER و TextBlob نظرًا لبساطتها وسهولة دمجها مع نماذج التصنيف (Hutto & Gilbert, 2014; Loria, 2020)، إلا أن هذه الأساليب تعتمد على قواعد ثابتة وقد تعجز عن التقاط الأنماط الدلالية المعقّدة في النصوص الرسمية. بالمقابل، أظهرت النماذج اللغوية العميقة المعتمدة على التعلّم الإشرافي، مثل DistilBERT، قدرة أعلى على تمثيل السياق الدلالي الكامل للنصوص وتحقيق أداء أفضل في مهام تصنيف النصوص (Minaee et al., 2020)، غير أنها تتطلب بيانات موسومة مسبقًا وموارد حسابية مرتفعة. أما فيما يتعلق بخوارزميات العنقدة غير الموجّهة، فقد أثبتت دراسات كلاسيكية وحديثة فاعلية خوارزمية KMeans في التقاط البنية العامة للبيانات النصية (MacQueen, 1967; Jain, 2010)، في حين أظهرت خوارزميات العنقدة الهرمية التراكمية قدرة أفضل على تمثيل العلاقات المتدرجة بين الوثائق النصية في بعض السيناريوهات (Müllner, 2011; Steinbach et al., 2000). كما أشارت دراسات حديثة مثل Bibi وآخرون (2022) إلى أن دمج تقنيات العنقدة مع مهام تحليل المشاعر يمكن أن يسهم في تحسين جودة الميزات المستخرجة، خاصة في البيئات غير الموجّهة أو ضعيفة التوسيم.

بناءً على ما سبق، يتضح أن معظم الأعمال السابقة إما تعتمد على نماذج إشرافية تتطلب بيانات موسومة، أو تستخدم تحليل مشاعر مباشر قد لا يعكس البنية الدلالية الكامنة للنصوص. في هذا السياق، يأتي البحث الحالي ليقدّم إطارًا غير موجّه لاستخلاص ميزات مشاعر أكثر استقرارًا اعتمادًا على العنقدة، ودمجها ضمن نموذج تصنيف موجّه لمعالجة شكاوى المستهلكين في بيئات بيانات غير متوازنة، بما يحقق توازنًا بين الأداء، وقابلية التطبيق، ومتطلبات الموارد.

مواد وطرائق البحث

3. أدوات البحث

في هذه الدراسة، تم تطبيق إطار هجين يعتمد على خوارزميات عنقدة غير موجّهة لاستخلاص ميزات المشاعر من نصوص شكاوى المستهلكين، ودمج هذه الميزات ضمن نموذج تصنيف موجّه للتنبؤ باعتراض المستهلك. حيث تم تمثيل النصوص عددياً باستخدام أسلوب TF-IDF، ثم تطبيق خوارزميتي KMeans والعنقدة الهرمية التراكمية (Agglomerative Clustering) لاكتشاف البنية الدلالية الكامنة في نصوص الشكاوى دون الاعتماد على تسميات مسبقة.

لاستخراج مؤشر المشاعر، تم استخدام أداة TextBlob لربط كل عنقود بقيمة مشاعر عددية، ثم تحويل ناتج العنقدة إلى ميزة عددية أُدرجت ضمن نموذج تصنيف موجّه يعتمد على خوارزمية Random Forest للتنبؤ بحالات اعتراض المستهلك.تم تطبيق هذا النهج على مجموعة بيانات شكاوى المستهلكين المقدمة إلى مكتب الحماية المالية للمستهلك

(Consumer Financial Protection Bureau (CFPB))، وهي عبارة عن مجموعة من الشكاوى حول المنتجات والخدمات المالية للمستهلكين من كامل مساحة الولايات المتحدة الأمريكية. تتألف مجموعة البيانات هذه من 18 واصفة (أعمدة) ،من أهمها: Consumer complaint narrative الذي يتضمن النص الحر للشكوى، و Consumer disputed? الذي يمثل المتغير الهدف ويشير إلى ما إذا كان المستهلك قد اعترض على نتيجة معالجة الشكوى أم لا، إضافةً إلى حقول أخرى تصف نوع المنتج المالي، والشركة المقدَّم بحقها الشكوى، وطريقة تقديم الشكوى، وتاريخها. بعد المعالجة الأولية وحذف السجلات غير الصالحة، تم استخدام مجموعة البيانات الناتجة في تدريب وتقييم النماذج المقترحة.تتوفر البيانات بتنسيقات CSV وJSON ويمكن الوصول إليها من مصادر متعددة بما في ذلك Kaggle وموقع مكتب حماية المستهلك المالي (CFPB) على الانترنت.

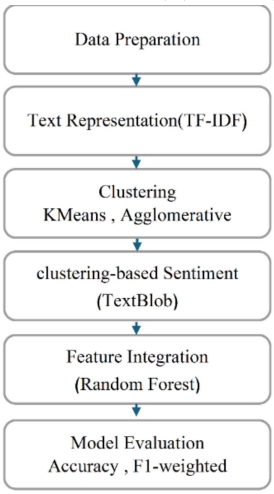
تم في هذا البحث العمل على منصة Google Colab لإجراء المحاكاة من حيث تصميم واختبار النظام المقترح، باستخدام نسخة Python 3، مع ذاكرة وصول عشوائي 12.7 GB ومعالج رسوميات NVIDIA Tesla T4 بذاكرة 15.3 GB وحجم تخزين GB78.2 .

بالإضافة إلى ذلك، تم استخدام مكتبات برمجية متخصصة منها:

- ✓ Scikit-learn لتطبيق خوارزميات TF-IDF،وخوارزميات العنقدة غير الموجهة، وبناء نموذج Random Forest وتقييم الأداء(Pedregosa et al., 2011) .
- ✓ TextBlob لاستخراج مؤشرات المشاعر من النصوص.
- ✓ Pandas وNumPy لمعالجة وتنظيم البيانات.
- ✓ Matplotlib لتمثيل النتائج بيانياً.

4. مخطط العمل

تتكوّن خطوات العمل في هذه الدراسة من مجموعة من المراحل المتسلسلة التي تم اعتمادها لبناء الإطار الهجين المقترح وتحقيق هدف البحث في التنبؤ باعتراض المستهلك ، وقد تم تنظيم هذه المراحل بشكل منهجي يبدأ بتحضير البيانات وينتهي بتقييم أداء النموذج، كما هو موضح في الشكل (1) الذي يبيّن مخطط مراحل العمل المعتمد في هذه الدراسة:



الشكل رقم (1): مخطط مراحل العمل

4-1. تحميل مجموعة البيانات (Dataset Loading)

حيث قمنا بتحميل مجموعة بيانات شكاوى المستهلكين من CFPB.

2-2. المعالجة الأولية للبيانات (Data Preprocessing)

وقد تمت هذه المرحلة على عدة خطوات؛ هي:

1. تنظيف البيانات Data Cleaning: وتتضمن تحميل بيانات شكاوى المستهلكين من ملف CSV، ثم معالجة العمود الهدف Consumer disputed? عبر توحيد القيم وتحويلها من (Yes/No) إلى ترميز ثنائي (0/1)، مع حذف السجلات التي تحتوي على قيم مفقودة في هذا العمود لضمان صلاحية البيانات للتدريب والتقييم. كما تم فحص القيم المفقودة في بقية الأعمدة الوصفية وتحديد طبيعتها ضمن مرحلة الاستكشاف الأولي.

2. تطبيع النص Text Normalization: تمت معالجة نصوص الشكاوى ضمن العمود Consumer complaint narrative عبر تحويل النص إلى أحرف صغيرة (Lowercasing) لتوحيد الصياغة وتقليل اختلافات الكتابة، بالإضافة إلى تعويض القيم المفقودة في هذا العمود بسلسلة فارغة (Empty String) لتجنب أخطاء المعالجة أثناء استخراج الميزات وتمثيل النصوص.

2-3. تقسيم البيانات (Data Splitting)

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

2-4. تمثيل النصوص (Text Representation)

تم تمثيل نصوص شكاوى المستهلكين باستخدام TF-IDF لتحويل النصوص غير المهيكلة إلى متجهات عددية تعبّر عن أهمية الكلمات داخل النصوص. يعتمد هذا الأسلوب على موازنة تكرار الكلمة داخل الشكوى مع تكرارها على مستوى مجموعة البيانات كاملة. تم اختيار أسلوب TF-IDF نظرًا لبساطته وكفاءته في تمثيل النصوص الرسمية القصيرة والمتوسطة الطول، مثل نصوص شكاوى CFPB، إضافةً إلى ملاءمته لخوارزميات العنقدة والتصنيف الإحصائية المستخدمة في هذا البحث.

تم تطبيق TF-IDF على العمود Consumer complaint narrative مع إزالة كلمات التوقف وتحديد عدد الميزات، لتوفير تمثيل عددي فعّال يشكّل الأساس لتطبيق خوارزميات العنقدة واستخدامه كمدخل مباشر في نموذج التصنيف.

2-5. استخلاص ميزات المشاعر المعتمدة على العنقدة غير الموجّهة(clustering-based Sentiment)

تم تطبيق خوارزميتي KMeans والعنقدة الهرمية التراكمية (Agglomerative Clustering) بوصفهما مرحلة وسيطة قبل استخراج المشاعر، بهدف تجميع النصوص المتشابهة دلاليًا ضمن عدد محدود من العناقيد (k=3). لا تُستخدم العنقدة في هذا الإطار بوصفها أداة لاستخلاص المشاعر بشكل مباشر، وإنما كوسيلة لتنظيم النصوص وفق تشابهها الدلالي العام، بما يساهم في تقليل الضجيج الناتج عن التباين الفردي في صياغة الشكاوى النصية الرسمية.

إن الاعتماد المباشر على قيم القطبية المستخرجة على مستوى النص الفردي قد يؤدي إلى تذبذب ملحوظ في تمثيل المشاعر، نتيجة اختلاف الأسلوب اللغوي واستخدام المصطلحات، حتى بين نصوص تعبّر عن تجارب متشابهة. لذلك، يتيح تجميع النصوص المتقاربة دلاليًا أولًا، ثم استخراج مؤشر مشاعر على مستوى العنقود، الحصول على تمثيل أكثر استقرارًا يعكس الاتجاه العام للمشاعر، بدلًا من الاعتماد على قيم قطبية فردية قد تكون متأثرة بعوامل لغوية سطحية.

بعد الحصول على العناقيد، تم استخراج قيم المشاعر لكل نص باستخدام مكتبة TextBlob بالاعتماد على مؤشر القطبية (polarity). ثم حُسب متوسط القطبية لكل عنقود، وأُستخدم هذا المتوسط لتوصيف العناقيد إلى ثلاث فئات: عنقود ذو مشاعر سلبية، وعنقود ذو مشاعر محايدة، وعنقود ذو مشاعر إيجابية. تم اعتماد حدود قطبية محددة مسبقًا لضمان الاتساق وقابلية إعادة الإنتاج، حيث اعتُبر العنقود سلبيًا إذا كان متوسط القطبية أقل من -0.1، ومحايدًا إذا وقع ضمن المجال [-0.1، 0.1]، وإيجابيًا إذا تجاوز 0.1. وقد تم اختيار هذه الحدود بالاستناد إلى الممارسات الشائعة في تحليل

المشاعر باستخدام TextBlob، بهدف عزل التذبذبات الطفيفة حول الصفر ومنع تأثير الفروق اللغوية الهامشية على توصيف المشاعر.

لتقييم ملاءمة العناقيد الناتجة للمنهجية المتبعة، تم تحليل التوزيع العددي للنصوص داخل كل عنقود، إضافة إلى دراسة توزيع قيم القطبية للنصوص ضمن كل عنقود قبل حساب المتوسط. أظهرت هذه التحليلات أن القيم داخل كل عنقود تتسم بدرجة مقبولة من التجانس، مما يدعم استخدام متوسط القطبية بوصفه تمثيلًا مناسبًا للاتجاه العام للمشاعر داخل العنقود. وبناءً على ذلك، تم ربط كل نص بقيمة مشاعر عددية واحدة تمثل العنقود الذي ينتمي إليه، وتحويل ناتج العنقدة إلى ميزة رقمية منفردة أُدرجت ضمن نموذج التصنيف الموجّه، بما يعزز استقرار الميزات وقابليتها للتفسير مقارنةً بتحليل المشاعر المباشر على مستوى الشكوى الفردية.

2-6. دمج الميزات (Feature Integration)

تم دمج ميزات المشاعر المستخرجة اعتمادًا على خوارزميات العنقدة غير الموجّهة مع الميزات النصية الممثلة بأسلوب TF-IDF، إضافةً إلى الميزات الوصفية للشكوى والتي تمثل الحقول غير النصّية المرافقة لكل سجل في مجموعة بيانات CFPB، مثل نوع المنتج المالي، الشركة المقدَّم بحقها الشكوى، وطريقة تقديم الشكوى، والتي تسهم في توصيف السياق العام للشكوى.. جرى تحويل الميزات الاسمية باستخدام One-Hot Encoding، ثم تجميع جميع الميزات ضمن متجه موحّد أُدخل إلى نموذج Random Forest للتنبؤ بحالات اعتراض المستهلك.

2-7. تقييم النموذج (Model Evaluation)

تم تقييم أداء النماذج المقترحة باستخدام مجموعة من مقاييس التصنيف الشائعة، شملت الدقة (Accuracy)، الدقة النوعية (Precision)، الاستدعاء (Recall)، ومقياس F1-weighted. وعليه، تم تعريف المقاييس المستخدمة كما يلي:

دقة التصنيف (Accuracy):

$$\frac{TP + TN}{TP + TN + FP + FN} = Accuracy$$

الدقة النوعية (Precision):

$$\frac{TP}{TP + FP} = \downarrow recision$$

الاستدعاء (Recall):

$$\frac{TP}{TP + FN} = Recall$$

مقياس F1:

$$\frac{Precision \times Recall}{Precision + Recall} \times 2 = F1$$

ونظرًا لعدم توازن البيانات بين فئتي الاعتراض وعدم الاعتراض، تم اعتماد مقياس F1-weighted بوصفه مؤشرًا أكثر ملاءمة لتقييم الأداء الكلي للنموذج.

إضافةً إلى ذلك، تم تطبيق تقاطع التحقق الطبقي (Stratified K-Fold Cross-Validation) بخمس طيات لضمان استقرار النتائج وتقليل أثر تقسيم البيانات. كما تم استخدام مصفوفة الالتباس (Confusion Matrix) لتحليل سلوك النموذج وفهم قدرته على التمييز بين الفئتين.

2-8. اختبار مراحل التنبؤ (Prediction Stages)

تم تصميم الإطار المقترح في هذا البحث ليكون مرنًا وقابلًا للتطبيق على بيانات نصية غير متوازنة، مع الاستفادة من خوارزميات عنقدة غير موجّهة لاستخلاص ميزات مشاعر أكثر تعبيرًا. وفيما يلي السيناريوهات العملية المقترحة لتوظيف الإطار المقترح:

المرحلة الأولى: تمثيل نصوص الشكاوى باستخدام TF-IDF.

الوصف: استخدام أسلوب TF-IDF لتمثيل نصوص شكاوى المستهلكين واستخراج الكلمات الأكثر أهمية داخل النصوص، اعتمادًا على تكرارها المحلي والعالمي.

النتيجة: أتاح هذا التمثيل تحويل النصوص غير المهيكلة إلى تمثيلات عددية فعّالة، شكّلت أساسًا مناسبًا لتطبيق خوارزميات العنقدة والتحليل اللاحق.

المرحلة الثانية: استخراج المشاعر اعتمادًا على العنقدة غير الموجّهة.

الوصف: تطبيق خوارزميتي KMeans والعنقدة الهرمية التراكمية (Agglomerative Clustering) على تمثيلات النصوص باستخدام TF-IDF بهدف تجميع الشكاوى المتشابهة دلاليًا دون الحاجة إلى تسميات مسبقة. بعد ذلك، جرى ربط كل عنقود ناتج بمؤشر مشاعر مستخرج باستخدام TextBlob، وذلك من خلال حساب متوسط القطبية داخل كل عنقود وتحويله إلى قيمة عددية تمثل المشاعر (سلبية، محايدة، إيجابية).

النتيجة: أسهم هذا النهج في الكشف عن أنماط دلالية كامنة داخل البيانات، مع اشتقاق ميزات مشاعر أكثر استقرارًا وتمثيلًا للبنية العامة للنصوص مقارنةً بتحليل المشاعر المباشر على مستوى الشكوى الفردية.

المرحلة الثالثة: دمج ميزات المشاعر العنقودية ضمن نموذج Random Forest.

الوصف: تم في هذا السيناريو دمج ميزات المشاعر المستخرجة بالاعتماد على العنقدة مع الميزات الوصفية الأخرى ضمن نموذج Random Forest للتنبؤ بحالات اعتراض المستهلك.

النتيجة: حقق النموذج المعتمد على ميزات مشاعر KMeans دقة بلغت 80.31%، في حين حقق النموذج المعتمد على Agglomerative Clustering دقة بلغت 80.57%.

المرحلة الرابعة: تقييم الأداء على بيانات غير متوازنة.

الوصف: نظرًا لعدم توازن البيانات بين فئتي الاعتراض وعدم الاعتراض، تم تقييم أداء النماذج باستخدام مقياس F1-weighted بوصفه مؤشرًا أكثر ملاءمة من الدقة وحدها.

النتيجة: أظهرت النتائج أن كلا النهجين حقق قيمة F1-weighted بلغت حوالي 77%، مما يعكس قدرة الإطار المقترح على تقديم أداء متوازن على كامل البيانات.

2-9. المقارنة مع نهج تصنيف المشاعر الإشرافي باستخدام DistilBERT

تم في هذا البحث إجراء مقارنة مع نهج سابق قائم على تصنيف المشاعر الإشرافي باستخدام تمثيلات سياقية مستخرجة بواسطة نموذج DistilBERT ومصنف Linear SVC، كما هو موضح في دراسة Minaee وآخرون (2020). يعتمد هذا النهج على الاستفادة من النماذج اللغوية العميقة المدربة مسبقًا لاستخلاص تمثيلات سياقية غنية للنصوص، ثم استخدام هذه التمثيلات في مهام تصنيف المشاعر اعتمادًا على بيانات موسومة مسبقًا. في إطار هذه الدراسة، تم استخدام مؤشر المشاعر المستخرج إشرافيًا بوصفه ميزة مشتقة، ودمجه ضمن نموذج Random Forest للتنبؤ باعتراض المستهلك، وذلك بهدف توفير خط أساس (Baseline) مرجعي لمقارنة أداء الإطار غير الموجّه المقترح. حيث تمت

المقارنة بين النهجين على نفس مجموعة البيانات، كما تم اعتماد نفس نموذج التصنيف النهائي Random Forest (n_estimators=260).

النتائج والمناقشة:

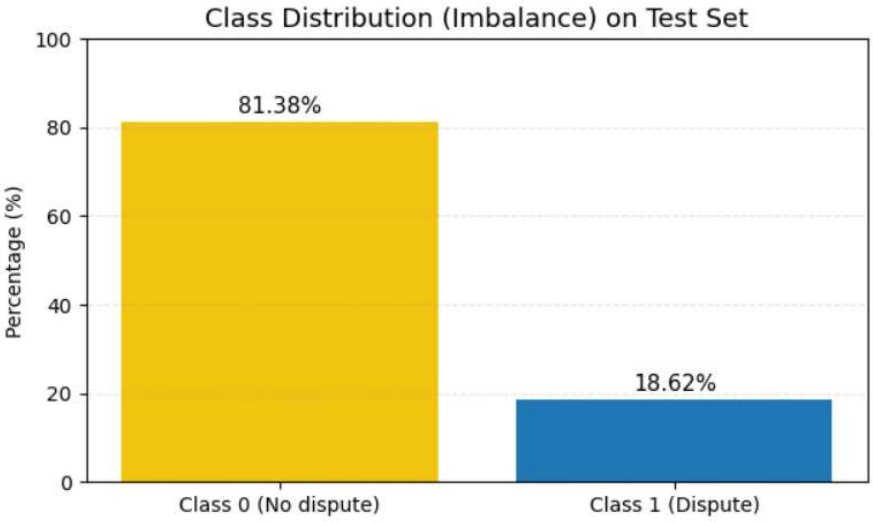
يبيّن الجدول (1) نتائج تقييم نماذج العنقدة غير الموجّهة باستخدام خوارزميتي KMeans و Agglomerative Clustering، إضافة إلى نموذج المشاعر المعتمد على DistilBERT.

الجدول رقم (1): مقارنة أداء نماذج KMeans وAgglomerative Clustering ونموذج المشاعر الإشرافي

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

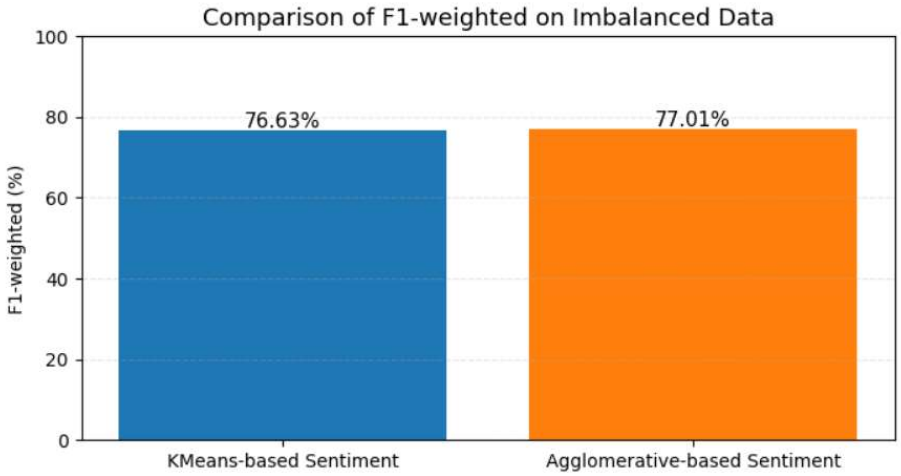
حقق نموذج Agglomerative Clustering دقة تصنيف (Accuracy) بلغت 0.8057، متفوقًا بشكل طفيف على KMeans الذي حقق دقة 0.8031. كما أظهر النموذجان قيمًا متقاربة لمقاييس Precision وRecall وF1.

تشير هذه النتائج إلى أن العنقدة الهرمية التراكمية قادرة على التقاط البنية الدلالية للنصوص بشكل أفضل نسبيًا من KMeans، إلا أن الفارق في الأداء يبقى محدودًا، مما يعكس تقارب الخوارزميتين في تمثيل العلاقات النصية عند الاعتماد على تمثيلTF–IDF . ومع ذلك، لا يمكن الاعتماد على مقياس الدقة (Accuracy) وحده في تقييم الأداء، نظرًا لوجود عدم توازن واضح في توزيع فئات البيانات بين حالات الاعتراض وعدم الاعتراض كما هو مبين بالشكل(2).



الشكل رقم (2): توزيع فئات البيانات في مجموعة الاختبار وبيان عدم توازن الفئات

ففي مثل هذه الحالات، قد تعكس الدقة المرتفعة قدرة النموذج على التنبؤ بالفئة المهيمنة أكثر من قدرته على التمييز الحقيقي بين الفئات. لذلك تم اعتماد مقياس F1–weighted بوصفه مؤشرًا أكثر ملاءمة لتقييم الأداء الكلي للنماذج، كونه يوازن بين الدقة والاستدعاء مع مراعاة وزن كل فئة وفق حجمها في البيانات. حقق نموذج Agglomerative Clustering قيمة F1–weighted بلغت 77.01%، متفوقًا بشكل طفيف على نموذج KMeans الذي حقق 76.63% كما هو موضح بالشكل(3).



الشكل3. مقارنة مقياس F1–weighted لنهجي KMeans وAgglomerative Clustering

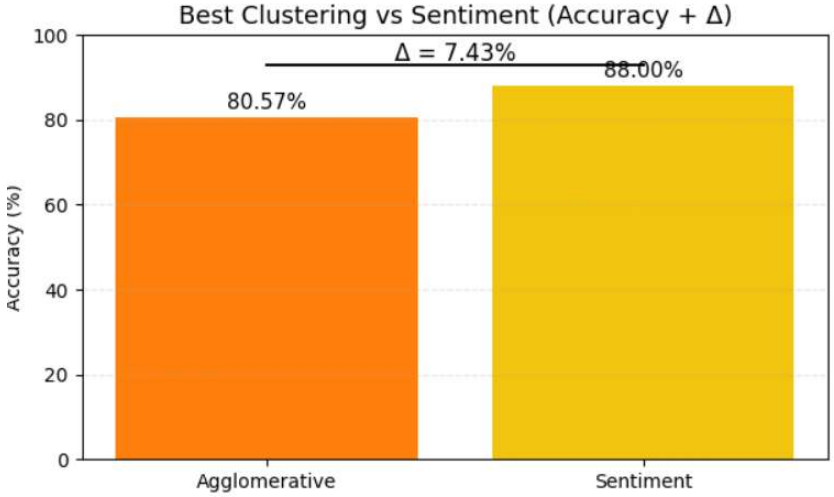
لإيضاح تأثير دمج ميزات المشاعر المعتمدة على العنقدة، تمت مقارنة أداء نموذج Random Forest في حالتين: الأولى بالاعتماد على الميزات النصية والوصفية فقط، والثانية بعد إضافة ميزة المشاعر المستخرجة اعتمادًا على العنقدة غير الموجّهة. أظهرت النتائج أن النموذج دون إضافة ميزات العنقدة حقق دقة تصنيف أعلى، وهو أمر متوقع نظرًا لاعتماد النموذج على عدد أكبر من الميزات التفصيلية المباشرة. في المقابل، أظهر النموذج المعزَّز بميزات المشاعر المعتمدة على العنقدة أداءً أكثر توازنًا عند تقييمه باستخدام مقياس F1–weighted، مما يشير إلى أن ميزات العنقدة تسهم في تحسين استقرار التنبؤ وتقليل الانحياز نحو الفئة المهيمنة في البيانات غير المتوازنة. وعليه، فإن دور العنقدة في الإطار المقترح لا يتمثل في تعظيم الدقة العددية، وإنما في دعم جودة التمثيل الدلالي للمشاعر وتقديم ميزات أكثر تعميمًا وقابلية للتفسير ضمن نموذج التصنيف. عند مقارنة الأساليب المعتمدة على العنقدة مع النموذج التنبؤي القائم على إضافة ميزة مشاعر مستخرجة إشرافيًا باستخدام DistilBERT، تبيّن أن نموذج المشاعر الإشرافي حقق أعلى دقة تصنيف بلغت 88.00%، متفوقًا على نماذج المشاعر المعتمدة على العنقدة غير الموجّهة. وبالمقارنة مع أفضل نموذج عنقدة (Agglomerative Clustering) الذي حقق دقة 80.57%، سجّل نموذج المشاعر الإشرافي تحسنًا قدره 7.43% في الدقة كما هو موضح بالشكل(4) . ويُعزى هذا التفوّق إلى قدرة نموذج DistilBERT على تمثيل السياق الدلالي الكامل للنصوص والتقاط تدفّق المعنى داخل الشكوى الواحدة، في حين تعتمد خوارزميات العنقدة على تمثيلات إحصائية عامة (TF–IDF) وتجميع النصوص على مستوى العناقيد، مما قد يؤدي إلى فقدان بعض التفاصيل الدلالية الدقيقة داخل كل عنقود.

على الرغم من أن نهج تصنيف المشاعر الإشرافي المعتمد على نموذج DistilBERT قد حقق دقة تصنيف أعلى مقارنةً بالنماذج المعتمدة على العنقدة غير الموجّهة، فإن هذا التحسّن العددي يأتي على حساب الاعتماد على بيانات موسومة مسبقًا وارتفاع التعقيد الحسابي خلال مرحلتي التدريب والاستدلال. ففي التطبيقات العملية واسعة النطاق، ولا سيما في تحليل شكاوى المستهلكين، يُعدّ توفير بيانات عالية الجودة وموسومة يدويًا عملية مكلفة زمنيًا وماديًا، وقد تكون غير متاحة في العديد من البيئات الواقعية أو محدودة الموارد.

لذلك، لا يُنظر إلى نموذج DistilBERT بوصفه بديلًا مباشرًا للإطار المقترح، بل كنقطة مرجعية (Upper Bound) للأداء عند توفر بيانات موسومة مسبقًا وموارد حسابية كافية.

في المقابل، يعمل الإطار المقترح القائم على استخلاص ميزات المشاعر اعتمادًا على العنقدة غير الموجّهة دون الحاجة إلى أي عملية توصيف يدوي للبيانات، مع المحافظة على أداء تنافسي عند تقييمه باستخدام مقياس F1–weighted الملائم لطبيعة البيانات غير المتوازنة. ويمنح هذا النهج ميزة إضافية تتمثل في خفّة الحمل الحسابي وقابلية التوسّع، مما يجعله مناسبًا للتطبيق في السيناريوهات التي تكون فيها البيانات الموسومة نادرة أو غير متوافرة.

وعليه، فإن الهدف من هذه المقارنة لا يتمثل في التفوق على النماذج اللغوية السياقية العميقة، وإنما في إبراز فعالية الإطار المقترح بوصفه حلًا عمليًا وخفيفًا وغير إشرافي للتنبؤ باعتراض المستهلك في بيئات البيانات النصية غير المتوازنة، حيث تُعدّ البساطة وقابلية التطبيق والتفسير عوامل ذات أولوية.



الشكل رقم (4) : مقارنة أداء أفضل نموذج عنقدة مع نموذج المشاعر الإشرافي من حيث الدقة

تتسق النتائج التي توصلت إليها هذه الدراسة مع ما أشار إليه Vaishnav وآخرون (2024) وBaskaran (2023)، حيث بيّنت دراساتهم أن نماذج التنبؤ باعتراض المستهلك المعتمدة على تمثيلات نصية مباشرة أو ميزات مشتقة من النصوص يمكن أن تحقق أداءً مقبولًا عند التعامل مع بيانات شكاوى CFPB، إلا أن هذا الأداء يتأثر بشكل واضح بطبيعة الميزات المستخدمة وآلية استخراجها. كما تدعم هذه النتائج ما خلص إليه Song وآخرون (2024) وWen وآخرون (2024) حول أهمية تمثيل البعد الدلالي للنصوص في تحليل شكاوى العملاء، خصوصًا في السياقات التي تتضمن نصوصًا رسمية ومباشرة.

فيما يتعلق بخوارزميات العنقدة غير الموجّهة، تتوافق النتائج مع ما عرضه MacQueen (1967) وJain (2010) حول فعالية خوارزمية KMeans في التقاط البنية العامة للبيانات، ومع ما أوضحه Müllner (2011) وSteinbach وآخرون (2000) بأن خوارزميات العنقدة الهرمية التراكمية قادرة على تمثيل العلاقات المتدرجة بين الوثائق النصية بشكل أفضل في بعض الحالات. وقد انعكس ذلك في تفوق نموذج Agglomerative Clustering بشكل طفيف على KMeans من حيث الدقة وF1–weighted، إلا أن الفارق بقي محدودًا، مما يشير إلى أن كلا الخوارزميتين تعتمدان بدرجة كبيرة على التمثيل الإحصائي للنصوص (TF–IDF) ولا تلتقطان السياق الدلالي العميق داخل كل شكوى.

من جهة أخرى، يتوافق التفوق الواضح للنموذج التنبؤي المعتمد على إضافة ميزة المشاعر المستخرجة إشرافيًا باستخدام DistilBERT مع ما ذكره Minaee وآخرون (2020) حول قدرة النماذج اللغوية العميقة على تمثيل السياق والمعنى الضمني للنصوص بشكل أدق من الأساليب الإحصائية التقليدية. ويؤكد هذا التفوق أن استخراج المشاعر اعتمادًا على

تمثيلات سياقية عميقة يوفر ميزات أكثر ثراءً عند دمجها ضمن نماذج تصنيف موجّهة مثل Random Forest، مقارنةً بميزات المشاعر المعتمدة على تجميع النصوص داخل عناقيد غير موجّهة.

كما تدعم هذه النتائج ما أشار إليه Saito & Rehmsmeier (2015) و He & Garcia (2009) بخصوص محدودية الاعتماد على مقياس الدقة (Accuracy) في البيئات غير المتوازنة، وضرورة استخدام مقاييس أكثر ملاءمة مثل F1–weighted. فقد أظهرت النتائج أن الفروق بين نماذج العنقدة تصبح أكثر وضوحًا عند تقييم الأداء باستخدام F1–weighted، مما يعكس قدرة هذا المقياس على تقديم تقييم أكثر واقعية لأداء النماذج في ظل اختلال توزيع الفئات.

بصورة عامة، تؤكد هذه المقارنة أن اختيار آلية استخراج المشاعر والميزات النصية يجب أن يتم بعناية تبعًا لطبيعة البيانات والهدف من التحليل. ففي حين توفر خوارزميات العنقدة غير الموجّهة حلاً مناسبًا في البيئات ضعيفة التوسيم، فإن دمج ميزات مشاعر مستخرجة إشرافيًا من نماذج لغوية سياقية عميقة مثل DistilBERT يحقق أداءً أعلى وأكثر استقرارًا في مهام التنبؤ باعتراض المستهلك، خاصة عند التعامل مع بيانات نصية غير متوازنة كما هو الحال في مجموعة بيانات CFPB. عند مقارنة نتائج هذا البحث مع دراسات أخرى اعتمدت على نهج غير إشرافي أو ضعيف الإشراف في تحليل شكاوى المستهلكين ضمن مجموعة بيانات CFPB، يلاحظ أن معظم هذه الدراسات ركّزت على اكتشاف الأنماط العامة أو تصنيف الشكاوى دون الاعتماد على بيانات موسومة مسبقًا. فقد بيّنت أعمال سابقة أن استخدام تقنيات غير إشرافية مثل العنقدة أو تحليل التوزيعات الدلالية يمكن أن يوفر تمثيلات عامة للنصوص تسهم في فهم بنية الشكاوى واتجاهاتها، إلا أن هذه الدراسات غالبًا ما واجهت تحديات تتعلق باستقرار الأداء ودقته عند التعامل مع بيانات غير متوازنة. بالمقارنة مع تلك الأعمال، يقدّم الإطار المقترح في هذه الدراسة نهجًا غير إشرافي لاستخلاص ميزات مشاعر مستقرة اعتمادًا على العنقدة، مع دمجها ضمن نموذج تصنيف موجّه للتنبؤ باعتراض المستهلك. وقد أظهرت النتائج أن هذا الدمج يحقق أداءً تنافسيًا عند تقييمه باستخدام مقياس F1–weighted، وهو ما يتماشى مع ما أشارت إليه الدراسات غير الإشرافية السابقة من حيث أهمية تقليل الاعتماد على البيانات الموسومة، مع تقديم معالجة أكثر توازنًا لطبيعة بيانات CFPB غير المتوازنة.

الاستنتاجات

توصّلت هذه الدراسة إلى أن دمج ميزات مشاعر مستخرجة من نصوص شكاوى المستهلكين ضمن نموذج تصنيف موجّه يُعدّ نهجًا فعّالًا للتنبؤ باعتراض المستهلك في بيئات البيانات النصية غير المتوازنة. فقد تم اقتراح إطار هجين يعتمد على استخراج ميزات مشاعر قائمة على خوارزميات عنقدة غير موجّهة (KMeans و Agglomerative Clustering) وربطها بمؤشر مشاعر مستخرج باستخدام TextBlob، ثم دمج هذه الميزات مع تمثيل TF–IDF والميزات الوصفية ضمن نموذج Random Forest.

أظهرت النتائج التجريبية أن نموذج Agglomerative Clustering حقق أداءً أفضل بشكل طفيف من KMeans، سواء من حيث الدقة أو مقياس F1–weighted، مما يشير إلى قدرة العنقدة الهرمية التراكمية على تمثيل العلاقات الدلالية المتدرجة بين النصوص بصورة أفضل نسبيًا. ومع ذلك، بقي الفارق بين خوارزميتي العنقدة محدودًا، وهو ما يعكس اعتماد كلتيهما على تمثيلات إحصائية عامة للنصوص وعدم قدرتهما على التقاط السياق الدلالي العميق داخل الشكوى الواحدة.

في المقابل، أظهر النموذج التنبؤي المعتمد على إضافة ميزة مشاعر مستخرجة إشرافيًا باستخدام DistilBERT تفوقًا واضحًا، حيث حقق أعلى دقة تصنيف بلغت 88.00%، متجاوزًا أفضل نموذج عنقدة بنسبة 7.43%. ويُعزى هذا التفوق إلى قدرة DistilBERT على تمثيل السياق الدلالي الكامل وتدفّق المعنى داخل النصوص، مما يوفر ميزات أكثر ثراءً عند دمجها ضمن نموذج Random Forest مقارنةً بميزات المشاعر المعتمدة على العنقدة غير الموجّهة.

كما أكدت النتائج أهمية استخدام مقاييس تقييم ملائمة لطبيعة البيانات غير المتوازنة، حيث أظهر مقياس F1-weighted قدرة أكبر على عكس الأداء الحقيقي للنماذج مقارنةً بالاعتماد على الدقة (Accuracy) وحدها. وقد ساهم ذلك في تقديم تقييم أكثر واقعية للفروق بين النماذج المقترحة.

بناءً على النتائج، نوصي بما يلي:

- استكشاف خوارزميات عنقدة نصية أكثر تطورًا أو تمثيلات دلالية أعمق (مثل embeddings السياقية) لتحسين جودة ميزات المشاعر المعتمدة على العنقدة.
- دراسة تأثير دمج ميزات مشاعر متعددة المصادر (إشرافية وغير موجّهة) ضمن إطار هجين موحّد لتعزيز الأداء.
- إجراء تجارب مستقبلية على مجموعات بيانات أكبر وأكثر تنوعًا للتحقق من قابلية تعميم النتائج.

المراجع:

20. Vaishnav, D., Neethinayagam, M., Khaire, A., & Woo, J. (2024). Predictive analysis of CFPB consumer complaints using machine learning. arXiv preprint.
21. Consumer Financial Protection Bureau (CFPB). CFPB website, Available at: https://www.consumerfinance.gov/ (Accessed on 1/3/2024)
22. Baskaran, P. K. (2023). Enhancing customer complaint classification in banking. Doctoral dissertation.
23. 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.
24. 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).
25. 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.
26. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.
27. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.
28. 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.
29. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32.

30. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
31. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of ICWSM, 8, 216–225.
32. Loria, S. (2020). TextBlob documentation (Release 0.16.0+).
33. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24, 513–523.
34. Bibi, M., Abbas, G., Azam, M., & Shahzad, M. (2022). A novel unsupervised ensemble framework using concept-based and hierarchical clustering for Twitter sentiment analysis. Pattern Recognition Letters.
35. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.
36. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666.
37. Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. arXiv preprint arXiv:1109.2378.
38. Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 109–118).