مذكرة تخرج مقدمة في

جامعة محمد بوضياف – المسيلة

كلية الرياضيات والإعلام الآلي

قسم الإعلام الآلي

من أجل الاستيفاء الجزئي لمتطلبات شهادة

ماستر في الإعلام الآلي

من طرف

لقمة حفيظة

جعيل نسرين

عنوان المذكرة

تحسين استراتيجية التسويق عبر جمع البيانات من الويب
Technical Marketing Leader

تحت إشراف الأستاذ

سمية شيكوش

صلاح الدين بوحويتة قرمش

سميرة عميش

أعضاء لجنة المناقشة

محمود ابراهيمي	جامعة المسيلة	**رئيسا**
سمية شيكوش	جامعة المسيلة	**مقررا**
صلاح الدين بوحويتة قرمش	جامعة المسيلة	**مقررا**
سميرة عميش	جامعة المسيلة	**مقررا**
الطاهر مهني	جامعة المسيلة	**مناقشا**
صلاح قاسمية	جامعة المسيلة	**مناقشا**

جوان،2026

الملخص

تتناول هذه الدراسة إشكالية تحليل مشاعر النصوص المكتوبة باللهجة الجزائرية الدارجة، وهي لهجة تتميز بخصائص لغوية معقدة تشمل المزج بين العربية والفرنسية والإنجليزية، إضافة إلى غياب التوحيد الإملائي وقلة البيانات المصنفة.

اعتمدت الدراسة على بيانات حقيقية جُمعت من منصة YouTube باستخدام تقنية كشط الويب، مستهدفةً تعليقات المستخدمين حول منتجات العناية بالبشرة.

وقد تم اختبار نماذج تمثل المقاربات المعتمدة في الدراسة، شملت التعلم الآلي التقليدي، والتعلم العميق، ونماذج المحوّلات، حيث أظهرت النتائج تفوق نموذج MARBERT بدقة بلغت 79% وF1-Macro قدره 0.72، يليه نموذج SVM مع TF-IDF بدقة 77%.

وبناءً على هذه النتائج، تم اقتراح تصور أولي لمنصة ذكاء تسويقي موجهة للمؤسسات الجزائرية، مع دراسة جدواها الاقتصادية ضمن إطار مشروع شركة ناشئة.

الكلمات المفتاحية:

تحليل المشاعر، اللهجة الجزائرية، معالجة اللغة الطبيعية، MARBERT، التعلم العميق، ذكاء التسويق.

Abstract

This study addresses the problem of sentiment analysis in texts written in the Algerian dialect, a linguistic variety characterized by complex features such as code-switching between Arabic, French, and English, in addition to the absence of standardized spelling and the scarcity of annotated datasets.

The study relied on real data collected from YouTube using web scraping techniques, targeting users' comments related to skincare products.

Models representing the different approaches adopted in this study were evaluated, including traditional machine learning, deep learning, and transformer-based models. The results showed the superiority of the MARBERT model, which achieved an accuracy of 79%

and an F1-Macro score of 0.72, followed by the SVM model with TF-IDF, which achieved an accuracy of 77%.

Based on these results, an initial design of a marketing intelligence platform for Algerian enterprises was proposed, along with an economic feasibility study within the framework of a startup project.

Keywords:

sentiment analysis, Algerian dialect, natural language processing, MARBERT, deep learning, marketing intelligence.

Résumé

Cette étude traite du problème de l'analyse des sentiments des textes rédigés en dialecte algérien, une variété linguistique caractérisée par des phénomènes complexes tels que le mélange entre l'arabe, le français et l'anglais, ainsi que l'absence de normalisation orthographique et la rareté des données annotées.

L'étude s'appuie sur des données réelles collectées depuis YouTube à l'aide de techniques de web scraping, en ciblant les commentaires des utilisateurs liés aux produits de soins de la peau.

Des modèles représentant les différentes approches adoptées dans cette étude ont été évalués, notamment l'apprentissage automatique classique, l'apprentissage profond et les modèles basés sur les transformers. Les résultats ont montré la supériorité du modèle MARBERT, qui a atteint une précision de 79 % et un score F1-Macro de 0,72, suivi du modèle SVM avec TF-IDF ayant obtenu une précision de 77 %.

Sur la base de ces résultats, une première conception d'une plateforme d'intelligence marketing destinée aux entreprises algériennes a été proposée, accompagnée d'une étude de faisabilité économique dans le cadre d'un projet de startup.

Mots-clés :

analyse des sentiments, dialecte algérien, traitement automatique du langage naturel, MARBERT, apprentissage profond, intelligence marketing.

اهداء

"بسم الله الرحمن الرحيم ،الحمدُ للهِ الذي بنعمتهِ تتمُّ الصالحات، وبتوفيقهِ تُنجَزُ الغاياتُ.

إلى من أدينُ لهما بكلّ شيء، إلى والديّ الحبيبين، ينبوعِ الحنانِ الذي لا ينضب، وصخرتَي الراسختَينِ في كلّ عاصفة.

أنتما من زرعتُما في قلبي حبّ المعرفة، وعلّمتُماني أن الأحلام لا تتحقق إلا بالصبر والدعاء. لو أحصيتُ تضحياتكما ما وسعَها كتاب، فلتقبَلا هذا العمل هديةً متواضعةً من قلبٍ يعجز عن ردّ الجميل.

إلى رفقاءِ الدرب، إلى إخوتي الأعزّاء، من شاركوني ضحكاتِ النجاح ودموعَ الإرهاق، وآمنوا بي حين شككتُ في نفسي كانت كلمةُ تشجيعٍ منكم تُعيد إليّ روحَ الاستمرار، أنتم الجذور التي تمنحني الثبات.

إلى من شاركتني هذا العمل خطوة بخطوة، وكانت خير سند لي في إنجاز هذه المذكرة.

شكرًا لكِ على كل الدعم والمساعدة، وعلى وجودك الذي جعل الطريق أسهل وأجمل.

فهذا النجاح أيضًا لكِ.

كما أتوجّه بالامتنان إلى أدوات الذكاء الاصطناعي التي كانت رفيقًا علميًا أمينًا، أسهمت في صقل الأفكار وتوسيع آفاق البحث، وفي ذلك دليلٌ على أن العلم لا يتقدّم إلا بالتعاون والتكامل.

هذا العمل ثمرةُ دعائكم وحبّكم، فإن كان فيه توفيقٌ فمن الله وحده، وإن كان فيه قصورٌ فمن نفسي"

لقمة حفيظة

الحمد لله, اللهم لك الحمد حتى ترضى، ولك الحمد إذا رضيت، ولك الحمد بعد الرضا، فما كان هذا الإنجاز إلا بفضلك، وما كان هذا الطريق ليُقطع لولا توفيقك ورحمتك.

أهدي ثمرة جهدي وتعب سنوات طويلة من الدراسة والسهر والأمل إلى أغلى الناس على قلبي:

إلى **أبي الغالي**، إلى الرجل الذي حمل همومي قبل أن أنطق بها، والذي تحمل عناء الحياة ليمنحني فرصة التعلم والنجاح، إلى من كان سندي وقوتي ومصدر فخري واعتزازي، أسأل الله أن يجزيه عني خير الجزاء.

وإلى **أمي الحبيبة**، إلى القلب الذي لم يعرف إلا العطاء، وإلى اليد التي أمسكت بيدي منذ أول خطوة في حياتي، إلى صاحبة الدعوات التي كانت تسبقني إلى أبواب السماء، إلى من سهرت وتعبت وضحت دون انتظار مقابل، أمي... مهما كتبت فلن أوفيك حقك، فجزاك الله عني خير ما جزى أمًا عن ابنتها.

إلى جدتاي الغاليتان، **الزهراء وعياشة**، أطال الله عمرهما وأدام عليهما الصحة والعافية. والى إخوتي وأخواتي؛ **إسلام**، سندي الثاني بعد أبي، و**وصال**، وصال القلب والروح، و**سندس**، مشاكسة المنزل وبهجته، و**محمد**، صغير العائلة وغاليها، الذين كانوا جزءًا من رحلتي، يشاركوني فرحتي ويخففون عني تعبي، فوجودكم نعمة أحمد الله عليها دائمًا.

إلى صديقاتي العزيزات : **أشواق دباش ، شيماء، إيمان، سعاد، أريج، خلود، فاطمة الزهراء، نبيلة، نورهان ،هبة، إيمان العايد، ياسمين، شرقي صفاء، ديلمي شيماء ، وردة، سمية، سارة، خولة** وإلى كل من وقف إلى جانبي بكلمة طيبة، أو دعوة صادقة، أو دعم خفي لا يعلمه إلا الله، شكرًا لأنكن كنتن نورًا في الأيام الصعبة، وعونًا في لحظات التعب.

إلى أهلنا في **غزة**، إلى القلوب الصابرة التي علمت العالم معنى الثبات والإيمان، إلى من تحملوا ما تعجز الكلمات عن وصفه، نهديكم هذا العمل المتواضع، ونرفع أكف الدعاء أن يفرج الله كربكم وينصركم ويجبر قلوبكم

ورفيقاتي **شفاء**، التي كانت بحق شفاءً لقلبي وروحي كلما أثقلني التعب ،**ووفاء** التي جسدت معنى اسمها، فكانت نعم الصديقة ونعم الرفيقة،**ودعاء** التي كانت من أجمل الأقدار التي رزقني الله بها، وكأنها استجابة لدعوات الخير التي لطالما رفعتها أمي ، **و إيناس** مؤنسة الروح والقلب، ورفيقة اللحظات الجميلة، شكرًا لصدقكِ وطيب قلبكِ، **وهاجر** السند الذي لم يمل يومًا،والتي شاركتني لحظات الفرح والتعب، وكانت حاضرة في محطات كثيرة من حياتي، فكانت خير معين وخير رفيق و إلى رفيقة المذكرة، إليكِ **يا حفيظة**، يا من تقاسمنا معًا لحظات الحيرة والأمل، والنجاح والتحدي.

وأخيرًا... إلى نفسي، التي تعثرت كثيرًا لكنها لم تستسلم، وواصلت السير حتى بلغت هذا اليوم. شكرًا لأنكِ صبرتِ واجتهدتِ وآمنتِ بحلمكِ.

نسرين جعيل .

تشكرات

الحمد لله الذي بنعمته تتم الصالحات، والذي وفقنا وأعاننا على إنجاز هذا العمل المتواضع.

نتقدم بخالص الشكر وعظيم الامتنان إلى أستاذتينا المشرفتين سمية شيكوش و سميرة عميش ، اللتين لم تبخلا علينا بتوجيهاتهما القيمة ونصائحهما السديدة، وكانتا خير معين لنا طوال مراحل إنجاز هذا المشروع، فلهما منا كل التقدير والاحترام.

ونتوجه بالشكر إلى الأستاذ صلاح الدين بوحويتة قرمش لموافقته على الإشراف على هذا العمل.

كما نتقدم بجزيل الشكر والتقدير إلى مديرة الحاضنة الأستاذة بن التومي سارة، وإلى مهندس الحاضنة الأستاذ جناوي عبد اللطيف، وإلى الأستاذة فصيلي سمية، على ما قدموه لنا من دعم وتوجيه وتشجيع، وكان لمساهمتهم أثر بالغ في إنجاز هذا العمل.

ولا يفوتنا أن نتقدم بالشكر إلى جميع الأساتذة عامة، وأساتذة قسم الإعلام الآلي خاصة، الذين ساهموا في تكويننا العلمي والمعرفي طيلة سنوات الدراسة.

ونتوجه بالشكر إلى كل من ساهم، من قريب أو بعيد، في إنجاز هذا العمل، سواء بالنصيحة أو المساعدة أو التشجيع.

وفي الأخير، نسأل الله أن يجعل هذا العمل خالصًا لوجهه الكريم، وأن ينفع به الجميع.

والحمد لله رب العالمين

الفهرس

قائمة المخططات والجداول والصور و المعادلات

المقدمة العامة 1

الفصل الأول: الإطار النظري للدراسة 4

1.1 المقدمة 4

2.1 التسويق الرقمي وذكاء الأعمال 5

1.2.1 مفهوم التسويق الرقمي وتطوره 5

2.2.1 دور وسائل التواصل الاجتماعي في فهم سلوك المستهلك 5

3.2.1 ذكاء الأعمال وذكاء التسويق 6

4.2.1 تحليل المنافسين عبر البيانات الرقمية 6

3.1 كشط الويب وجمع البيانات ((Web Scraping 7

1.3.1 مفهوم كشط الويب وآلية عمله 7

2.3.1 أنواع البيانات الرقمية 8

3.3.1 أدوات وتقنيات جمع البيانات 8

4.3.1 التحديات التقنية لجمع البيانات من يوتيوب 9

5.3.1 الجوانب القانونية والأخلاقية 9

4.1 معالجة اللغات الطبيعية ((NLP 10

1.4.1 تعريف معالجة اللغات الطبيعية ومجالات تطبيقها 10

2.4.1 المراحل الأساسية في معالجة النصوص 10

3.4.1 تمثيل النصوص: من الرمز إلى المعنى 11

4.4.1 أهمية NLP في تحليل البيانات النصية غير المهيكلة 12

5.1 تحليل المشاعر 12

1.5.1 تعريف تحليل المشاعر وموقعه من معالجة اللغات الطبيعية 12

2.5.1 أنواع تحليل المشاعر 13

3.5.1 مستويات تحليل المشاعر 14

4.5.1 تطبيقات تحليل المشاعر في التسويق ومراقبة العلامات التجارية 15

6.1 خصوصية اللهجة الجزائرية الدارجة 15

1.6.1 خصائص اللهجات العربية المحلية حاسوبيًا 15

2.6.1 ظاهرة المزج اللغوي 16

3.6.1 الإيموجي و الاختصارات و التعابير العامية.....16
4.6.1 نقص الموارد اللغوية والبيانات المصنّفة.....17
5.6.1 مقارنة اللهجة الجزائرية باللغة العربية الفصحى.....17
7.1 الخاتمة :.....19
الفصل الثاني : تقنيات تمثيل النصوص ونماذج تحليل المشاعر 20
1.2مقدمة :.....20
2.2 تمثيل النصوص.....20
1.2.2 مفهوم تمثيل النص وأهميته.....20
2.2.2 مفهوم التمثيل التقليدي القائم عن الاحصاء.....21
1.2.2.2 نموذج (TF-IDF) و (n-grams).....21
3.2.2 مفهوم التضمين الدلالي.....22
1.3.2.2 نموذج FastText.....22
4.2.2 مفهوم التمثيلات السياقية باستخدام نماذج المحوّلات.....22
5.2.2 الفرق بين التمثيل التقليدي والدلالي.....23
3.2 التعلم الآلي التقليدي في تصنيف النصوص.....23
1.3.2 مفهوم التصنيف النصي.....23
2.3.2 خوارزمية SVM و مبدأ عملها.....24
4.2 نماذج التعلم العميق للنصوص.....24
1.4.2 مفهوم التعلم العميق وأهميته في NLP.....24
2.4.2 الشبكات العصبية الالتفافية CNN و تطبيقها على النصوص.....24
5.2 نماذج المحولات.....25
1.5.2 مفهوم معمارية Transformer و آلية الانتباه.....25
2.5.2 نموذج MARBERT.....25
3.5.2 مفهوم Fine-Tuning و أهميته.....26
6.2 مقارنة عامة بين المقاربات الثلاث.....26
1.6.2 نقاط القوة والضعف لكل مقاربة.....26
2.6.2 معايير اختيار النموذج المناسب.....28

7.2 الخاتمة.....28
الفصل الثالث: الجانب التطبيقي والتجريبي 29
1.3المقدمة:.....29
2.3 وصف البيانات وجمعها.....29
1.2.3 مصدر البيانات و مبرر اختياره.....29
2.2.3 العلامات التجارية المستهدفة ومعايير اختيارها.....29
3.2.3 آلية جمع البيانات عبر YouTube Data API v3.....30
1.3.2.3 وصف الأداة المستخدمة.....30
2.3.2.3 منهجية الجمع البرمجي.....30
3.3.2.3 التحديات التقنية المواجَهة.....32
4.2.3 وصف مجموعة البيانات الخام والتصنيف اليدوي.....32
1.4.2.3 حجم البيانات و توزيعها.....32
2.4.2.3 منهجية التصنيف اليدوي.....33
3.3 المعالجة المسبقة للنصوص.....33
1.3.3 المرحلة الأولى: تصفية البيانات.....33
2.3.3 المرحلة الثانية: تنظيف البيانات.....34
3.3.3 المرحلة الثالثة: معالجة الرموز التعبيرية.....34
4.3.3 المرحلة الرابعة: تطبيع النصوص للهجة الجزائرية.....35
5.3.3 المرحلة الخامسة: معالجة التكرار النصي.....36
6.3.3 المرحلة السادسة: معالجة النفي.....36
7.3.3 المرحلة السابعة: التقطيع.....37
8.3.3 المرحلة الثامنة : إزالة كلمات الوقف (الحشو).....38
4.3 إعداد بيانات التدريب.....38
1.4.3 توزيع الفئات وتشخيص مشكلة عدم التوازن.....38
2.4.3 معالجة عدم توازن الفئات.....39
1.2.4.3 تقنية SMOTE.....39
2.2.4.3 تقنية Class Weights.....39
3.4.3 تقسيم البيانات.....39
5.3 تدريب النماذج وتقييم الأداء.....40

1.5.3 البيئة التقنية والمكتبات المستخدمة.....40
2.5.3 النماذج الكلاسيكية 41
1.2.5.3نموذج SVM مع TF-IDF.....41
3.5.3 نماذج التعلم العميق (Word Embeddings).....46
1.3.5.3نموذج CNN مع FastText 46
4.5.3 نماذج المحولات (Transformers).....52
1.4.5.3نموذج MARBERT (Fine-Tuning).....52
6.3 الاستنتاجات التحليلية للنماذج 59
7.3 الخاتمة.....60
الفصل الرابع: تجسيد وبناء النظام المقترح 61
1.4المقدمة:.....61
2.4 بيئة التطوير 61
1.2.4 العتاد والأجهزة (Hardware) 61
2.2.4 الأدوات البرمجية (Software Tools).....62
1.2.2.4 React.....62
2.2.2.4 Vite 62
3.2.2.4 FastAPI 62
4.2.2.4 Node.js 63
5.2.2.4 Firebase 63
6.2.2.4 Visual Studio Code.....63
3.4 التقنيات المدمجة في النظام (Implemented Technologies).....63
4.4 مكونات النظام الأساسية (Application Components) 64
5.4 واجهات التطبيق (Application Interfaces) 64
1.5.4 الواجهة الرئيسية للمنصة (Home Page Interface) 64
2.5.4 واجهات المصادقة وإدارة الحسابات 67
3.5.4 إدارة المستخدمين الخاصة بالمشرف 67
4.5.4 واجهات العميل (Client & Sentiment Analysis).....68
6.4 الخاتمة.....71
الفصل الخامس: الجانب الاقتصادي 72
1.5 المقدمة 72

2.5 المحور الأول : تقديم المشروع 72
1.2.5 فكرة المشروع (الأبعاد الاستراتيجية والحل المقترح) 72
2.2.5 القيم المقترحة 74
3.2.5 فريق العمل 75
4.2.5 أهداف المشروع 76
5.2.5 الجدول الزمني لتحقيق المشروع 77
3.5 المحور الثاني : الجوانب الابتكارية 78
1.3.5 طبيعة الابتكارات 78
2.3.5 مجالات الابتكارات 78
1.2.3.5 ابتكار العمليات ((Process Innovation) 78
2.2.3.5 ابتكار الخدمات والخصائص (Service & Feature Innovation) 79
3.2.3.5 ابتكار نموذج العمل (Business Model Innovation) 79
4.5 المحور الثالث : التحليل الاستراتيجي للسوق 79
1.4.5 عرض القطاع السوقي 79
1.1.4.5 السوق المحتمل 79
2.1.4.5 السوق المستهدف 80
3.1.4.5 مبررات اختيار السوق المستهدف 81
2.4.5 قياس شدة المنافسة 82
1.2.4.5 المنافسون المباشرون 82
2.2.4.5 المنافسون غير المباشرين 82
3.4.5 الاستراتيجية التسويقية (Go-to-Market Strategy) 83
1.3.4.5 الاستراتيجية التنافسية والتسويقية 83
2.3.4.5 المزيج التسويقي للخدمات (Marketing Mix – 7Ps) 83
5.5 المحور الرابع : خطة الإنتاج والتنظيم 85
1.5.5 عملية الإنتاج (تطوير الخدمة) 85
1.1.5.5 الاكتشاف والتسجيل والتأهيل 85
2.1.5.5 تجميع البيانات وتحديد مصادر الطلب 86
3.1.5.5 المعالجة السحابية والتحليل الذكي 86
4.1.5.5 توليد المؤشرات والاستكشاف الإحصائي 87

5.1.5.5 العرض والتسليم النهائي للمنتج الرقمي 88
6.1.5.5 الدعم الفني والاستدامة التشغيلية 88
2.5.5 اليد العاملة 89
3.5.5 الشراكات الرئيسية 91
6.5 المحور الخامس : الخطة المالية 91
1.6.5 التكاليف الاستثمارية 91
2.6.5 التكاليف التشغيلية 92
1.2.6.5 المشتريات 92
2.2.6.5 أجور العمال 93
3.2.2.5 الأعباء الخارجية 93
3.6.5 تقدير رأس المال والإيرادات 94
1.3.6.5 تقدير راس المال 94
2.3.6.5 تقدير الإيرادات 94
7.5 المحور السادس : النموذج الأولي التجريبي 95
8.5 الخاتمة 96
الخاتمة العامة 97
المراجع
قائمة الملاحق

قوائم المخططات والجداول والصور و الاختصارات و المعادلات

1- قائمة المخططات

الفصل الاول: الاطار النظري للدراسة.....4

المخطط 1.1: مقاربة متعددة المستويات لتحليل المشاعر 13

المخطط 2.1: التحديات المشتركة بين معالجة اللهجة الدارجة الجزائرية وتحليل المشاعر في سياق المحتوى الرقمي 18

الفصل الثاني: تقنيات تمثيل النصوص ونماذج تحليل المشاعر 20

المخطط 1.2: توضح الصورة مستويات نماذج معالجة النصوص من التقليدي إلى المحوّلات 27

الفصل الخامس: الجانب الاقتصادي.....72

المخطط 1.5: مرحلة الاكتشاف والتسجيل والتأهيل 85

المخطط 2.5: مرحلة تجميع البيانات وتحديد مصادر الطلب 86

المخطط 3.5: مرحلة المعالجة السحابية والتحليل الذكي 87

المخطط 4.5: مرحلة توليد المؤشرات والاستكشاف الإحصائي 87

المخطط 5.5: مرحلة العرض والتسليم النهائي للمنتج الرقمي 88

المخطط 6.5: مرحلة الدعم الفني والاستدامة التشغيلية.....89

2- قائمة الجداول

الفصل الاول: الاطار النظري للدراسة.....4

الجدول 1.1:مقارنة بين العربية الفصحى والدارجة الجزائرية في المعالجة الحاسوبية 18

الفصل الثاني: تقنيات تمثيل النصوص ونماذج تحليل المشاعر 20

الجدول 1.2: يبين الجدول مقارنة بين أساليب تمثيل النصوص 23

الجدول 2.2: يوضح مقارنة تحليلية بين المقاربات الثلاث المعتمدة 27

الفصل الثالث : الجانب التطبيقي والتجريبي..... 29

الجدول 1.3: يبين توزيع التعليقات على الفئات الثلاث 33

الجدول 2.3: تقسيم البيانات المعالجة 40

الجدول 3.3: الأدوات والمكتبات البرمجية المستخدمة 40

الجدول 4.3: إعدادات TF-IDF المعتمدة مع نموذج SVM 42

الجدول 5.3: نتائج تقييم نماذج التعلم العميق 44

الجدول 6.3: معمارية نموذج CNN المعتمد في الدراسة 50

الجدول 7.3: نتائج تقييم نماذج التعلم العميق 51

الجدول 8.3: جدول إعدادات الضبط الدقيق (Hyperparameters) لنموذج MARBERT.....56

الجدول 9.3:نتائج نموذج الMARBERT.....57

الجدول 10.3: نتائج تقييم النماذج.....59

الفصل الرابع : تجسيد و بناء النظام المقترح.....61

الجدول1.4: العتاد المستخدم في التطوير 61

الفصل الخامس: الجانب الاقتصادي.....72

الجدول 1.5:القيم المقترحة للمشروع.....74

الجدول 2.5: فريق العمل و أدوارهم 75

الجدول 3.5:الجدول الزمني لتحقيق المشروع 77

الجدول 4.5: التكاليف الاستثمارية 92

الجدول 5.5:جدول المشتريات 92

الجدول 6.5: جدول أجور العمال.....93

الجدول 7.5: جدول أعباء الخارجية 93

الجدول 8.5: تقدير رأس المال.....94

الجدول 9.5:جدول الإيرادات.....94

3- قائمة الصور

الفصل الثالث : الجانب التطبيقي والتجريبي..... 29

الصورة 1.3: خوارزم عملية استخراج البيانات من منصة يوتيوب 31

الصورة 2.3 :عينة من التعليقات الخام المجمعة من منصة يوتيوب قبل المعالجة.....32

الصورة 3.3: خوارزم عملية معالجة الرموز التعبيرية.....35

الصورة 4.3: خوارزم عملية تطبيع النصوص 36

الصورة 5.3: خوارزم عملية معالجة النفي 37

الصورة 6.3: عينة من القاموس الناتج عن TF-IDF 43

الصورة 7.3: مصفوفة الارتباك لنموذج ال SVM 45

الصورة 8.3: عينة من القاموس الناتج عن FastText 49

الصورة 9.3: نتائج التشابه الدلالي باستخدام FastText 49

الصورة 10.3: مصفوفة الارتباك لنموذج ال CNN 51

الصورة 11.3: مخطط التوزيع الفضائي لتضمينات نموذج MARBERT 58

الفصل الرابع : تجسيد و بناء النظام المقترح.....61

الصورة 1.4: الواجهة الرئيسية للمنصة و قسم التعريف بخدمة تحليل المشاعر 65

الصورة 2.4 :قسم المميزات والخدمات المتاحة في المنصة.....65

الصورة 3.4: قسم كيف يعمل ؟ آلية عمل نظام تحليل المشاعر المقترح.....66

الصورة 4.4: خطط الاشتراك المتاحة في المنصة ومزاياها 66

الصورة 5.4: واجهة تسجيل الدخول وإنشاء الحساب واستعادة كلمة المرور.....67

الصورة 6.4: إدارة المستخدمين الخاصة بالمشرف 68

الصورة 7.4: واجهة العميل استخراج التعليقات وتحليلها 68

الصورة 8.4: جدول التعليقات والتحليل الآلي 69

الصورة 9.4: واجهة لوحة التحكم الإحصائية لتحليل اتجاهات.....70

الصورة 10.4: واجهة التقرير الذكي لملخص نقاط القوة والضعف المستخلصة آلياً من التعليقات.....70

4- قائمة الاختصارات

NLP: Natural Language Processing.

SVM: Support Vector Machine.

TF-IDF: Term Frequency – Inverse Document Frequency .

CNN: Convolutional Neural Network.

MARBERT: Moroccan Arabic Bidirectional Encoder Representations from Transformers.

UGC :User-Generated Content.

API:Application Programming Interface.

SEO/SEM :Search Engine Optimization / Search Engine Marketing.

MSA :Modern Standard Arabic.

ReLU :Rectified Linear Unit .

GELU :Gaussian Error Linear Unit .

SMOTE : Synthetic Minority Over-sampling Technique.

JSON :JavaScript Object Notation .

HTML :HyperText Markup Language.

OOV :Out-of-Vocabulary.

KPIs :Key Performance Indicators.

FFN :Feed-Forward Network.

GDPR :General Data Protection Regulation.

CSV:Comma-Separated Values.

5- قائمة المعادلات

الفصل الثاني: تقنيات تمثيل النصوص ونماذج تحليل المشاعر 20

المعادلة (1): معادلة TF-IDF..... 21

المعادلة (2): معادلة Attention Mechanism..... 25

الفصل الثالث : الجانب التطبيقي والتجريبي..... 29

المعادلة (3): حساب Term Frequency 41

المعادلة (4): حساب Inverse Document Frequency.....41

المعادلة (5): حساب وزن TF-IDF 42

المعادلة (6): معادلة المستوى الفاصل في SVM..... 43

المعادلة (7): دالة تحسين SVM Objective Function..... 44

المعادلة (8): تجميع تمثيل العقدة في FastText..... 47

المعادلة (9): حساب درجة التصنيف 47

المعادلة (10): دالة Sigmoid 47

المعادلة (11): دالة الخسارة في Skip-Gram..... 48

المعادلة (12): تحديث الأوزان بالانتشار العكسي 48

المعادلة (13): تجميع تمثيل الكلمات الفرعية 48

المعادلة (14): تمثيل الكلمة في FastText 48

المعادلة (15): مصفوفة التضمين (Embedding Matrix) 48

المعادلة (16): تمثيل الإدخال في BERT 53

المعادلة (17): حساب مصفوفة Query 53

المعادلة (18): حساب مصفوفة Key 53

المعادلة (19): حساب مصفوفة Value 54

المعادلة (20): معادلة Residual Connection + Layer Normalization 54

المعادلة (21): دالة Feed Forward Network FFN 55

مقدمة عامة

شهد العالم الرقمي في السنوات الأخيرة تحولات جذرية في طبيعة التفاعل الإنساني مع المعلومات، إذ باتت منصات التواصل الاجتماعي والمواقع التفاعلية الفضاء الأول الذي يعبّر فيه الملايين من المستخدمين عن آرائهم وتجاربهم ومشاعرهم تجاه مختلف المنتجات والخدمات والعلامات التجارية. وقد أفرز هذا الواقع الجديد كميات هائلة من البيانات النصية غير المهيكلة، التي تتدفق يوميًا عبر التعليقات والمنشورات والمراجعات، وتحمل في طياتها ثروة معلوماتية حقيقية يصعب استثمارها دون اللجوء إلى أدوات تحليلية متطورة.

في مواجهة هذا التحدي، برزت تقنيات الذكاء الاصطناعي ومعالجة اللغات الطبيعية (Natural Language Processing) بوصفها الأداة الأكثر قدرة على استخراج المعرفة من هذا الكم الهائل من النصوص. ومن أبرز تطبيقات هذا المجال، تحليل المشاعر (Sentiment Analysis)، الذي يهدف إلى تصنيف النصوص وفق الحالة الوجدانية التي تعبّر عنها، سواء أكانت إيجابية أم سلبية أم محايدة. وقد أثبتت هذه التقنية فعاليتها الواسعة في مجالات عديدة، من بينها التسويق الرقمي، ومراقبة سمعة العلامات التجارية، وتحليل رضا الزبائن، ودعم اتخاذ القرار الاستراتيجي على مستوى المؤسسات.

غير أن فاعلية هذه التقنيات ترتبط ارتباطًا وثيقًا بطبيعة اللغة المستهدفة ودرجة توافر الموارد اللغوية الخاصة بها. ففي حين حققت نماذج تحليل المشاعر نتائج باهرة مع اللغة الإنجليزية وعدد من اللغات الأوروبية، لا تزال اللغة العربية بصفة عامة تعاني من ندرة نسبية في الأدوات والموارد المتخصصة، ويزداد الوضع تعقيدًا حين يتعلق الأمر باللهجات العربية المحلية وفي مقدمتها اللهجة الجزائرية الدارجة.

تتميز اللهجة الجزائرية بخصائص لغوية فريدة تجعل معالجتها حاسوبيًا تحديًا بالغ الدقة. فهي لهجة تفتقر إلى معايير إملائية موحدة، وتعرف تباينًا كبيرًا في طرق الكتابة والتعبير من منطقة إلى أخرى ومن مستخدم إلى آخر. فضلًا عن ذلك، تتسم بظاهرة المزج اللغوي (Code-Switching) بين العربية والفرنسية والإنجليزية في النص الواحد، وتعتمد اعتمادًا واسعًا على الاختصارات والرموز التعبيرية والتعابير العامية والتلقائية. يُضاف إلى ذلك شُح البيانات المصنّفة والموارد اللغوية المتخصصة بهذه اللهجة، مما يجعل بناء نماذج تحليل مشاعر دقيقة وقابلة للتعميم مهمةً بالغة التعقيد مقارنةً بما هو متاح للغات الأكثر توثيقًا.

وفي ظل هذه المعطيات، تندرج هذه المذكرة ضمن مسعى بحثي يجمع بين المتطلبات الأكاديمية والحاجات التطبيقية الفعلية للمؤسسات الاقتصادية الجزائرية. إذ تتمحور إشكاليتها الرئيسية حول التساؤل التالي: إلى أي مدى يمكن لنماذج التعلم الآلي والتعلم العميق ونماذج المحولات الحديثة أن تحقق أداءً فعالًا في تحليل مشاعر النصوص المكتوبة باللهجة الجزائرية الدارجة، وكيف يمكن توظيف هذه النتائج لبناء منصة ذكاء تسويقي تخدم القرار الاستراتيجي للمؤسسات الجزائرية؟

للإجابة عن هذا التساؤل، تم اتخاذ شركة "Massiliya Dermo-Cosmétiques" الجزائرية المتخصصة في منتجات شبه الصيدلانية والعناية بالبشرة إطارًا تطبيقيًا لهذه الدراسة. وقد اعتمد العمل على جمع بيانات حقيقية من منصة يوتيوب باستخدام تقنية كشط الويب (Web Scraping)، وذلك من خلال استخراج تعليقات المستخدمين المتعلقة بمنتجات العناية بالبشرة الخاصة بالعلامات التجارية المنافسة. وقد شكّلت هذه التعليقات قاعدة البيانات الخام التي انطلق منها البحث، بوصفها عيّنة حقيقية وممثّلة للخطاب الاستهلاكي الجزائري الرقمي.

على الصعيد التقني، اعتمدت الدراسة منهجية مقارنة شاملة تضم ثلاث نماذج موزعة على ثلاث مقاربات تكاملية. تمثّل المقاربة الأولى نماذج التعلم الآلي التقليدي، من خلال استخدام TF-IDF مقترنًا بخوارزمية SVM. وتضم المقاربة الثانية نماذج التعلم العميق، حيث تم توظيف تمثيل FastText مع معمارية CNN. أما المقاربة الثالثة فتوظّف نماذج المحولات الحديثة، من خلال إجراء Fine-Tuning على نموذج MARBERT المدرَّب خصيصًا على اللغة العربية. وقد رافقت هذه العملية التقنية مراحل معالجة مسبقة دقيقة للنصوص، تضمّنت تنظيف البيانات وتوحيدها، ومعالجة الرموز التعبيرية وأساليب النفي، والتعامل مع مشكلة عدم توازن الفئات باستخدام تقنيات Class Weights و SMOTE.

لا يقتصر هذا العمل على البُعد الأكاديمي المتعلق بمقارنة النماذج، بل يمتد ليشمل بُعدًا تطبيقيًا ذا قيمة عملية مباشرة. إذ يُقدّم تصورًا أوليًا لمنصة ذكاء تسويقي (Technical Marketing Leader Platform) مصمّمة لتمكين المؤسسات الجزائرية من رصد آراء المستهلكين في الوقت الفعلي، وتتبّع صورة المنافسين، وفهم توجهات السوق، وذلك انطلاقًا من تحليل تعليقات المستخدمين على وسائل التواصل الاجتماعي (اليوتيوب). وقد تم تعزيز هذه المنصة بواجهة توصيات ذكية تستند إلى Gemini API ، بهدف توليد تقارير وتوصيات تسويقية آلية تُسهم في دعم القرار وتحسين الاستراتيجيات التنافسية.

وتكتسب هذه الدراسة أهميتها من تقاطع اتجاهين بحثيين وتطبيقيين في آنٍ واحد؛ فمن الناحية العلمية تُسهم في إثراء مجال تحليل المشاعر الخاص باللهجة الجزائرية، وهو مجال لا يزال يعاني من شُح الأعمال المتخصصة، كما تتيح مقارنة موضوعية بين مختلف الأجيال التقنية في معالجة اللغات الطبيعية ضمن بيئة لغوية محلية. ومن الناحية التطبيقية، تُبيّن كيف يمكن تحويل البيانات الرقمية غير المهيكلة إلى أصول معرفية قابلة للاستثمار في صياغة الاستراتيجيات التسويقية وتعزيز القدرة التنافسية للمؤسسات الوطنية.

وتحقيقاً لهذا الهدف، جاءت الدراسة في خمس فصول متكاملة؛ تناول الفصل الأول الإطار النظري للدراسة من خلال استعراض مفاهيم التسويق الرقمي وذكاء الأعمال، وآليات كشط الويب وجمع البيانات، ومبادئ معالجة اللغات الطبيعية وتحليل المشاعر، مع التركيز على خصوصية اللهجة الجزائرية الدارجة وتحدياتها الحاسوبية.

أما الفصل الثاني فقد خُصِّص لتقنيات تمثيل النصوص ونماذج تحليل المشاعر، إذ استعرض المقاربات الثلاث المعتمدة في الدراسة، من التمثيل الإحصائي التقليدي كTF-IDF، مروراً بنموذج التضمين الدلالي FastText، وصولاً إلى نماذج المحوّلات السياقية كMARBERT، مع مقارنة تحليلية بين نقاط قوة كل مقاربة وحدودها.

وتضمّن الفصل الثالث الجانب التطبيقي والتجريبي، حيث وُصِفت منهجية جمع البيانات من منصة يوتيوب وتصنيفها يدوياً، ثم مراحل المعالجة المسبقة للنصوص المكتوبة بالدارجة الجزائرية، وإعداد بيانات التدريب، وتدريب مجموعة من النماذج الكلاسيكية والعميقة وتقييم أدائها المقارن.

أما الفصل الرابع فقد تناول الجانب التطبيقي للمشروع، حيث يتم عرض مراحل تجسيد الحل المقترح، مع تقديم الأدوات والتقنيات المستخدمة، وهندسة النظام، وأهم واجهات التطبيق المطورة.

في حين انصرف الفصل الخامس إلى الجانب الاقتصادي، مُجسِّداً إمكانية تحويل النتائج البحثية إلى منتج رقمي قابل للتسويق، من خلال تقديم المشروع بوصفه شركةً ناشئةً، وتحليل سوقه وبيئته التنافسية، ووضع خطة إنتاج وتنظيم وخطة مالية تُثبت جدواه الاقتصادية.

الفصل الأول: الإطار النظري للدراسة

1.1 المقدمة

يُشكّل هذا الفصل الركيزة النظرية التي يرتكز عليها البحث بأكمله، إذ يهدف إلى بناء إطار مفاهيمي متكامل يجمع بين عدة أبعاد مترابطة، تمتد من مجال التسويق الرقمي وصولًا إلى تقنيات الذكاء الاصطناعي ومعالجة اللغات الطبيعية.

وقد اقتضت طبيعة الدراسة تنظيم هذا الفصل وفق تسلسل منطقي متدرج؛ إذ يتناول في بدايته السياق التسويقي والاقتصادي للدراسة، من خلال استعراض مفهوم التسويق الرقمي وتطوره، وإبراز دور وسائل التواصل الاجتماعي في فهم سلوك المستهلك، بالإضافة إلى التطرق لمفهوم ذكاء التسويق وأهميته بالنسبة للمؤسسات في ظل التحول الرقمي.

ثم ينتقل إلى تقنية كشط الويب باعتبارها الأداة الأساسية المعتمدة في جمع البيانات، مع عرض آليات عملها، وأهم الأدوات المستخدمة فيها، فضلًا عن التحديات التقنية والقانونية المرتبطة بها. ليتوقف بعدها عند مجال معالجة اللغات الطبيعية، من خلال تقديم مفاهيمه الأساسية ومراحله المختلفة، بما يُمهّد لفهم التقنيات المستخدمة لاحقًا في معالجة البيانات النصية.

وينتقل بعد ذلك إلى مفهوم تحليل المشاعر بوصفه المحور الرئيسي لهذه الدراسة، مع استعراض أنواعه ومستوياته وأبرز تطبيقاته في المجال التسويقي، وصولًا إلى الخصوصية اللغوية للهجة الجزائرية الدارجة، والتحديات الحاسوبية المرتبطة بمعالجتها وتحليلها.

ولا يهدف هذا الفصل إلى تقديم عرض موسوعي شامل لهذه المجالات، وإنما يسعى إلى توفير الأسس النظرية والمفاهيم الجوهرية التي يحتاجها القارئ لفهم الاختيارات المنهجية والتقنية المعتمدة في الجانب التطبيقي. ومن هذا المنطلق، يُمثّل هذا الفصل حلقة وصل بين الإشكالية البحثية والأدوات والنماذج التجريبية التي سيتم تناولها في الفصول اللاحقة.

2.1 التسويق الرقمي وذكاء الأعمال

1.2.1 مفهوم التسويق الرقمي وتطوره

يُعد التسويق الرقمي توظيفًا استراتيجيًا لمنصات الإنترنت والتقنيات الرقمية المتنوعة بهدف تحقيق أهداف تسويقية وبناء علاقات مستدامة مع العملاء. ويشمل هذا المجال إدارة الحضور الرقمي للمؤسسة عبر المواقع الإلكترونية والتطبيقات ووسائل التواصل الاجتماعي، مما يساعد على الوصول إلى العملاء وتحقيق أهداف تسويقية فعالة في آنٍ واحد [1].

وقد مرّ التسويق الرقمي بمراحل تطور متعاقبة، إذ انطلق في مرحلته الأولى مع بداية انتشار الإنترنت في منتصف تسعينيات القرن الماضي، حيث اقتصر على بناء المواقع الإلكترونية وإرسال البريد الإلكتروني التسويقي. ثم تطور لاحقًا مع ظهور محركات البحث وأدوات تحسين الظهور (SEO/SEM)، قبل أن يدخل مرحلته الثالثة والأكثر نضجًا مع انتشار منصات التواصل الاجتماعي وأدوات التحليل والبيانات الضخمة، مما أعاد رسم ملامح العلاقة بين المؤسسات وعملائها[1].

ويتميّز التسويق الرقمي عن نظيره التقليدي بقدرته على التخصيص الفائق (Hyper-personalization) وقياس الأثر في الوقت الفعلي، إضافة إلى إتاحة تفاعل مباشر ومتبادل بين العلامة التجارية والمستهلك، وهو ما كان شبه مستحيل في الأطر التسويقية الكلاسيكية [1].

2.2.1 دور وسائل التواصل الاجتماعي في فهم سلوك المستهلك

تُعرَّف وسائل التواصل الاجتماعي بأنها مجموعة من تطبيقات الإنترنت المبنية على الأسس الأيديولوجية والتقنية للجيل الثاني من الويب (Web 2.0)، والتي تتيح إنشاء المحتوى الذي يولّده المستخدمون وتبادله [2]. وقد أحدث هذا التحول نقلة نوعية في طبيعة العلاقة بين المستهلك والعلامة التجارية، إذ لم يعد المستهلك متلقيًا سلبيًا للرسائل التسويقية، بل أصبح طرفًا فاعلًا في إنتاج المحتوى والتأثير في الصورة الذهنية للمنتجات والخدمات.

وتتجلى أهمية وسائل التواصل الاجتماعي في فهم سلوك المستهلك من خلال ما تولّده من محتوى حقيقي وتلقائي يعكس الآراء الحقيقية للمستخدمين، وهو ما يُعرف بمحتوى المستخدمين (User-Generated Content - UGC) . ويُعتبر

هذا النوع من المحتوى مصدرًا بيانيًا لا يُعوَّض لفهم توجهات المستهلكين وتفضيلاتهم وسلوكياتهم الشرائية، نظرًا لكونه يصدر بصورة تلقائية بعيدًا عن التأثير التسويقي المباشر [2].

ومن الناحية التطبيقية، تعتمد المؤسسات على بيانات وسائل التواصل الاجتماعي في متابعة التفاعل كالإعجابات والتعليقات والمشاركات، وتحليل الاتجاهات السائدة (Trending Topics)، ورصد ردود أفعال المستهلكين تجاه الحملات التسويقية والمنتجات الجديدة. كما تُشير الدراسات إلى أن المستهلكين باتوا يعتمدون بشكل متزايد على تقييمات وآراء الآخرين على هذه المنصات عند اتخاذ قراراتهم الشرائية، مما يجعل مراقبة هذا المحتوى أمرًا استراتيجيًا لا ترفًا [1].

3.2.1 ذكاء الأعمال وذكاء التسويق

يُقصد بذكاء الأعمال (Business Intelligence) منظومة العمليات والأدوات والتقنيات التي تساعد على جمع البيانات الخام وتحويلها إلى معلومات منظمة وقابلة للتفسير، بهدف دعم اتخاذ القرار على مختلف المستويات الإستراتيجية والتكتيكية والتشغيلية داخل المؤسسة. وفي السياق التسويقي تحديدًا، فيُستخدم مفهوم أكثر تخصصًا يُعرف بذكاء التسويق (Marketing Intelligence)، الذي يهتم بجمع المعلومات المتعلقة بالسوق والمنافسين والمستهلكين بصورة منهجية ومستمرة، وتحليلها لاستخلاص رؤى تُوجّه القرارات التسويقية [1].

ويرتكز ذكاء التسويق في البيئة الرقمية على ثلاثة محاور متكاملة، أولها تحليل بيانات العملاء لفهم سلوكهم وتوقع احتياجاتهم، وثانيها رصد أداء المنافسين وتتبع تحركاتهم في الفضاء الرقمي، وثالثها تقييم أداء الحملات التسويقية باستخدام مؤشرات الأداء (KPIs). وبذلك لم يعد ذكاء التسويق مجرد أداة تحليلية استرجاعية، بل تحوّل إلى ركيزة استشرافية تُمكّن المؤسسات من الاستجابة الاستباقية لمتغيرات السوق [1].

4.2.1 تحليل المنافسين عبر البيانات الرقمية

يُعدّ تحليل المنافسين ركيزة أساسية في منظومة التخطيط الاستراتيجي للمؤسسات، وقد اكتسب هذا التحليل في البيئة الرقمية أبعادًا جديدة وأدوات أكثر دقة وفعالية. إذ أصبح بمقدور المؤسسات اليوم رصد الحضور الرقمي للمنافسين بشكل شبه آني، ومتابعة ردود أفعال المستهلكين تجاه منتجاتهم وخدماتهم، واستخلاص دروس إستراتيجية قيّمة من تجاربهم [1].

وفي هذا الإطار، تُعتبر البيانات المستخرجة من منصات التواصل الاجتماعي مصدرا مهما لتحليل المنافسين .لأنها تعكس تصورات المستهلكين الفعلية تجاه العلامات التجارية المنافسة بصورة تلقائية وغير مُصطنعة. مما تساعد على فهم موقع العلامة التجارية التنافسي في أذهان المستهلكين [2].

وتشمل أبرز محاور تحليل المنافسين رقميًا رصد التعليقات والمراجعات المتعلقة بمنتجات المنافسين، وتحليل مشاعر المستهلكين تجاههم، ومتابعة الاتجاهات الجديدة في قطاع النشاط، وتحديد الفجوات والفرص التي يمكن للمؤسسة استثمارها تنافسيًا. وهذا بالضبط ما تسعى إليه هذه الدراسة من خلال تحليل تعليقات المستهلكين المتعلقة بمنتجات العناية بالبشرة المنافسة لشركة "ماسيليا" على منصة يوتيوب.

3.1 كشط الويب وجمع البيانات (Web Scraping)

1.3.1 مفهوم كشط الويب وآلية عمله

يُعرَّف كشط الويب (Web Scraping) بأنه العملية الآلية لاستخراج البيانات من صفحات الويب عبر برامج حاسوبية مخصصة، تعمل على محاكاة سلوك المتصفح البشري في الوصول إلى المحتوى الرقمي وتحليله واستخلاص المعلومات المرغوبة منه. وتتميز هذه التقنية بقدرتها على معالجة كميات ضخمة من البيانات في وقت قصير، مما يجعلها أداةً لا غنى عنها في مشاريع علم البيانات و البحث الأكاديمي التي تستلزم جمع بيانات حقيقية من الفضاء الرقمي[3] .

وتسير عملية كشط الويب وفق مراحل متسلسلة ومحددة، تبدأ بإرسال طلب HTTP إلى الخادم المستهدف (HTTP Request)، فيستجيب الخادم بإرسال شيفرة HTML الخاصة بالصفحة المطلوبة. بعد ذلك تتولى أداة الكشط تحليل هذه الشيفرة (HTML Parsing) للتعرف على بنيتها واستخراج العناصر المطلوبة كالنصوص والروابط والصور والتعليقات. وتُخزَّن البيانات المستخرجة في النهاية في صيغ قابلة للمعالجة كملفات CSV أو JSON أو قواعد البيانات[3]. كما يمكن التمييز بين نوعين رئيسيين من كشط الويب، الكشط الثابت (Static Web Scraping) الذي يتعامل مع الصفحات ذات المحتوى الثابت، والكشط الديناميكي (Dynamic Web Scraping) الذي يستهدف الصفحات التي تعتمد على JavaScript لتحميل محتواها، كما هو الحال في منصة يوتيوب.

2.3.1 أنواع البيانات الرقمية

تنقسم البيانات الرقمية التي يمكن استخراجها عبر تقنيات كشط الويب إلى فئتين رئيسيتين متمايزتين من حيث البنية و أساليب المعالجة :

- البيانات المهيكلة (Structured Data): وهي البيانات المنظمة في جداول وحقول محددة المعنى، كقواعد البيانات العلائقية والجداول الإلكترونية، إذ يسهل معالجتها واستعلامها مباشرةً؛
- البيانات غير المهيكلة (Unstructured Data): وهي البيانات التي تفتقر إلى تنظيم مسبق محدد، كالنصوص الحرة والتعليقات ومقاطع الفيديو والصور ومنشورات وسائل التواصل الاجتماعي.

وتُمثّل التعليقات المستخرجة من منصة يوتيوب في هذه الدراسة نموذجًا للبيانات غير المهيكلة، إذ تتسم بتنوع كبير في الأسلوب والمحتوى واللغة، وتستلزم مراحل معالجة مسبقة دقيقة قبل توظيفها في تدريب نماذج تحليل المشاعر. وتشير الإحصاءات إلى أن نسبة كبيرة من بيانات العالم الرقمي تقع ضمن هذه الفئة غير المهيكلة، مما يجعل أدوات معالجتها من أكثر المجالات البحثية نشاطًا وتطورًا[3].

3.3.1 أدوات وتقنيات جمع البيانات

تتعدد الأدوات والمكتبات المتاحة لتنفيذ عمليات كشط الويب، وتتباين فيما بينها من حيث الأسلوب التقني ومجال التطبيق.

تُعدّ مكتب BeautifulSoup من أكثر أدوات تحليل HTML شيوعًا في بيئة Python ، إذ تتيح التنقل عبر شجرة عناصر الصفحة (DOM Tree) واستخراج البيانات المستهدفة بيُسر وبساطة. وتتميز بسهولة التعلم وملاءمتها للصفحات الثابتة البسيطة [3].

أما مكتبة Selenium فتُوظَّف أساسًا للتعامل مع الصفحات الديناميكية التي تعتمد على JavaScript، إذ تعمل عبر محاكاة متصفح حقيقي يُنفّذ الشيفرة البرمجية ويتفاعل مع الصفحة كما يفعل المستخدم البشري [3].

وفي سياق هذه الدراسة تحديدًا، تم الاعتماد على YouTube Data API v3 بوصفها الأداة الرسمية التي توفرها Google للوصول البرمجي المنظم إلى بيانات المنصة. وتتيح هذه الواجهة استخراج التعليقات والبيانات الوصفية للمقاطع

عبر طلبات HTTP منظمة تعيد البيانات بصيغة JSON، كما تُقدّم ضمانات أعلى من حيث الاستقرار والموثوقية مقارنةً بأساليب الكشط المباشر[4].

4.3.1 التحديات التقنية لجمع البيانات من يوتيوب

يطرح جمع البيانات من منصة يوتيوب جملةً من التحديات التقنية التي ينبغي أخذها بعين الاعتبار عند تصميم أي مشروع بحثي يعتمد على بياناتها.

أبرز هذه التحديات قيود الحصص اليومية (API Quotas) التي تفرضها Google على استخدام YouTube Data API، حيث تُحدَّد الطلبات المسموح بها يوميًا لكل مشروع، مما قد يُعيق جمع كميات كبيرة من البيانات دفعةً واحدة ويتطلب تخطيطًا مسبقًا دقيقًا [4]. كما نواجه التعددية اللغوية في التعليقات، إذ قد يحتوي المقطع الواحد تعليقات بلغات ولهجات متعددة، مما يستوجب عملية تصفية وتصنيف لغوي قبل الشروع في المعالجة. إضافة إلى وجود الإيموجي والروابط والمحتوى غير النصي مصدرًا إضافيًا للضجيج البياني الذي ينبغي معالجته.

ومن الناحية التقنية أيضا، قد يتعرض الكشط المباشر للحجب من قِبَل المنصة عبر آليات مكافحة الروبوتات (Anti-bot mechanisms) ، مما يجعل استخدام الواجهة الرسمية خيارًا أكثر استقرارًا و أمانًا [3].

5.3.1 الجوانب القانونية والأخلاقية

يستدعي جمع البيانات من الفضاء الرقمي الوقوف عند جملة من الاعتبارات القانونية والأخلاقية التي باتت تحظى باهتمام متزايد في أوساط البحث العلمي. وتُشكّل هذه الاعتبارات إطارًا ضروريًا يضمن أن تسير عملية جمع البيانات ضمن الحدود المشروعة من الناحيتين القانونية والأخلاقية [5].

من الناحية القانونية، تفرض منصة يوتيوب شروط استخدام (Terms of Service) صريحة تحكم طريقة الوصول إلى بياناتها، إذ تُجيز الوصول عبر الواجهة البرمجية الرسمية (API) وتحظر الكشط المباشر غير المصرح به. كما نشير إلى أن هذه الدراسة اعتمدت على YouTube Data API v3 الرسمية بما يضمن الامتثال الكامل لهذه الشروط [4]. وعلى الصعيد الأوسع، تفرض تشريعات حماية البيانات كاللائحة الأوروبية العامة لحماية البيانات (GDPR) قيودًا صارمة على جمع البيانات الشخصية ومعالجتها وتخزينها دون موافقة صريحة من أصحابها [5].

أما على الصعيد الأخلاقي، فقد التزمت هذه الدراسة بمجموعة من المبادئ الجوهرية، إذ اقتصر جمع البيانات على التعليقات العامة المنشورة دون المساس بأي معلومات شخصية حساسة. وقد استُخدمت البيانات المجموعة حصريًا لأغراض البحث العلمي الأكاديمي، مع عدم توظيفها في أي استخدام تجاري أو غرض آخر غير متعلق بالدراسة.

4.1 معالجة اللغات الطبيعية (NLP)

1.4.1 تعريف معالجة اللغات الطبيعية ومجالات تطبيقها

تُعرَّف معالجة اللغات الطبيعية (Natural Language Processing - NLP) بأنها فرع من فروع الذكاء الاصطناعي الذي يهدف الى تمكين الحواسيب من فهم اللغة البشرية وتحليلها وتوليدها. ويقع هذا المجال عند تقاطع علم اللغويات والحوسبة والذكاء الاصطناعي، حيث يسعى إلى تطوير نماذج حاسوبية قادرة على التعامل مع اللغة في صورها المكتوبة والمنطوقة بطريقة أقرب إلى الفهم البشري.

وتمتد تطبيقات هذا المجال لتشمل مجالات واسعة ومتنوعة، ففي مجال الترجمة الآلية باتت أنظمة كـ Google Translate قادرة على الترجمة بين مئات اللغات بجودة مقبولة. وفي مجال استرجاع المعلومات تعتمد محركات البحث اعتمادًا جوهريًا على تقنيات NLP لفهم نية المستخدم وتقديم النتائج الأكثر صلة. كما تتوسع تطبيقاته لتشمل الإجابة الآلية عن الأسئلة، وتلخيص النصوص، وتحليل المشاعر، والتعرف على الكيانات المسماة (Named Entity Recognition)، وأنظمة الحوار الآلي (Chatbots)، وهو ما يجعله من أكثر مجالات الذكاء الاصطناعي حضورًا في التطبيقات اليومية والحديثة [6].

2.4.1 المراحل الأساسية في معالجة النصوص

يمر النص الخام عادةً بسلسلة من المراحل المتتابعة حتى يصبح قابل للتحليل الحاسوبي. وقد عُرضت هذه المراحل ضمن إطار منهجي يُعدّ مرجعًا راسخًا في هذا المجال [6].

أولًا: التقطيع (Tokenization): تُمثّل عملية التقطيع الخطوة الأولى في أي خط معالجة نصية، وتقوم على تجزئة النص المتواصل إلى وحدات أصغر تُسمى الرموز (Tokens). وعلى الرغم من بساطتها الظاهرة، تطرح هذه العملية

تحديات حقيقية في اللغات التي تفتقر إلى فواصل واضحة بين الكلمات، أو في النصوص التي تتضمن أرقامًا وعلامات ترقيم ورموزًا خاصة؛

ثانيًا: إزالة كلمات التوقف (Stop Words Removal): كلمات التوقف هي الكلمات الوظيفية الشائعة التي تتكرر بكثافة في اللغة كحروف الجر وأدوات التعريف وبعض الضمائر. ولا تحمل معنى دلاليًا مهمًا، لذلك يتم حذفها لتقليل حجم البيانات والتركيز على الكلمات الأكثر أهمية؛

ثالثًا: التجذير والإرجاع إلى الأصل (Stemming & Lemmatization) : يهدف التجذير (Stemming) إلى اختزال الكلمة إلى جذرها الصرفي بحذف اللواحق والسوابق. أما الإرجاع إلى الأصل (Lemmatization) فهو يعيد الكلمة إلى صورتها القاموسية الأصلية (Lemma) اعتمادًا على السياق، مما يجعله أكثر دقة وأكثر كلفة من الناحية الحاسوبية؛

رابعًا: تحليل الأصناف النحوية (Part-of-Speech Tagging): تهدف هذه المرحلة إلى تصنيف كل كلمة في الجملة وفق وظيفتها النحوية، كالاسم والفعل والصفة والحرف. وتُعدّ هذه المعلومة ضرورية للمراحل التحليلية ؛

خامسًا: التحليل النحوي (Parsing): يسعى التحليل النحوي إلى استيعاب البنية التركيبية للجملة وتمثيلها في صورة شجرة نحوية تُبيّن العلاقات الهرمية بين مكوناتها. ويُوفّر هذا التحليل أساسًا لفهم المعنى الكامل للجملة بما يتجاوز حدود الكلمة المفردة.

3.4.1 تمثيل النصوص: من الرمز إلى المعنى

تُعدّ مسألة تمثيل النصوص في صورة رقمية قابلة للمعالجة الحاسوبية من أكثر المسائل أهميةً وتأثيرًا في أداء نماذج NLP. ويُعتبر تحويل الرموز اللغوية إلى تمثيلات رياضية هو ما يحدد في نهاية المطاف مدى قدرة النموذج على فهم المعنى.

في البداية، كانت الأنظمة تعتمد على تمثيلات متفرقة (Sparse Representations) كنموذج Bag of Words الذي يُختزل فيه النص إلى تكرارات كلماته دون مراعاة للترتيب أو السياق، ونموذج TF-IDF الذي يُطوّر هذا التمثيل بموازنة تكرار الكلمة في الوثيقة مقابل ندرتها عبر المجموعة الكاملة من الوثائق. وعلى الرغم من فعاليتهما في مهام كثيرة، يظل قصورهما الرئيسي في إهمال العلاقات الدلالية بين الكلمات وعدم قدرتهما على التقاط المعنى السياقي .

لاحقًا ظهر مفهوم تمثيلات الكلمات الكثيفة (Dense Word Representations) أو ما يُعرف بـ Word Embeddings وهو نقلةً نوعية في هذا المجال، إذ باتت الكلمات تُمثَّل في فضاء متجهي مستمر تنعكس فيه العلاقات الدلالية بين الكلمات على شكل قرب هندسي بين المتجهات. وقد أتاح هذا التطور التقاط التشابه الدلالي والعلاقات التناظرية بين الكلمات بصورة لم تكن ممكنة مع التمثيلات التقليدية [6].

4.4.1 أهمية NLP في تحليل البيانات النصية غير المهيكلة

تكتسب معالجة اللغات الطبيعية أهميةً استثنائية في سياق البيانات غير المهيكلة، التي أصبحت تُشكّل الجزء الأكبر من البيانات المتداولة في الفضاء الرقمي. فالتعليقات والمنشورات والمراجعات والمحادثات النصية هي في جوهرها بيانات لغوية خام تحمل معلومات ذات قيمة عالية، لكنها تحتاج إلى أدوات تحليلية متخصصة .

وفي السياق التسويقي تحديدًا، تُمكّن تقنيات NLP المؤسساتِ من استخراج رؤى إستراتيجية من ملايين التعليقات والمراجعات بأسلوب آلي وفي وقت قياسي، بدلًا من الاعتماد على التحليل اليدوي المحدود والمكلف. وتظهر أهميتها بصورة خاصة في تحليل مشاعر المستهلكين، وتصنيف الشكاوى وطلبات الدعم، إضافة إلى تتبع الاتجاهات السائدة في النقاشات المتعلقة بالعلامات التجارية، وهي جميعًا تُشكّل مهام أساسية لبناء منصة الذكاء التسويقي التي تسعى إليها هذه الدراسة [6].

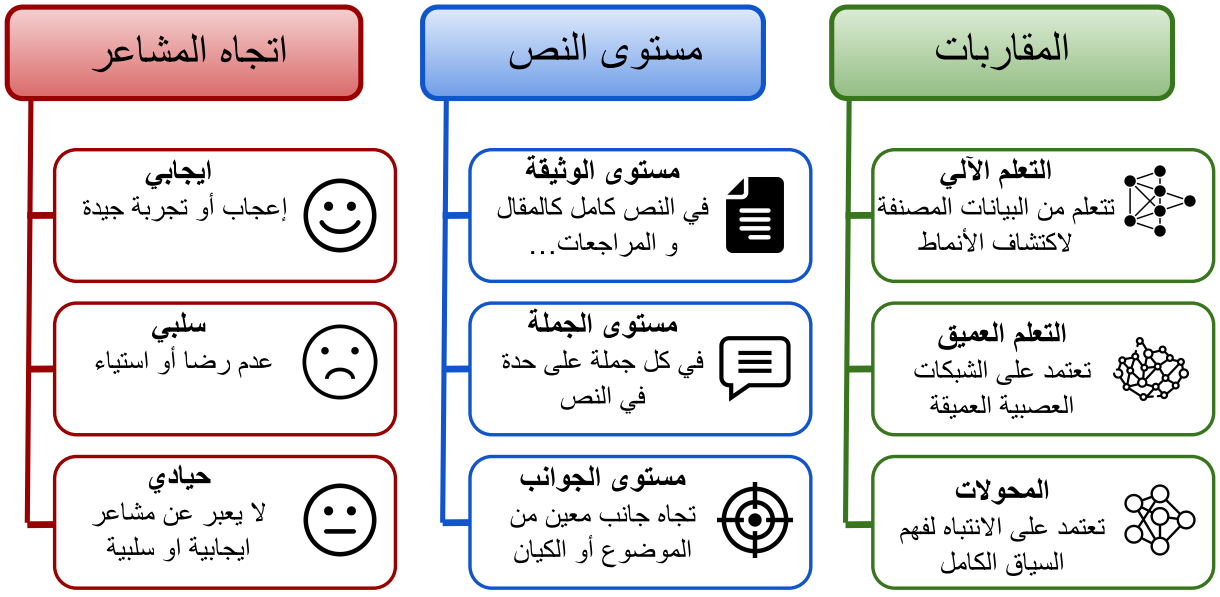
5.1 تحليل المشاعر

1.5.1 تعريف تحليل المشاعر وموقعه من معالجة اللغات الطبيعية

يُعدّ تحليل المشاعر (Sentiment Analysis) أو ما يُعرف كذلك بالتعدين عن الآراء (Opinion Mining) من أبرز مجالات معالجة اللغات الطبيعية وأكثرها توظيفًا في التطبيقات العملية. ويُعرَّف بأنه عملية تحليل الآراء والمشاعر والتقييمات والمواقف والعواطف التي يُعبّر عنها الناس تجاه المنتجات والخدمات والأفراد والمؤسسات والأحداث [7]. ويتميز هذا المجال عن باقي مهام NLP في كونه لا يسعى إلى استخراج المعلومات الموضوعية فقط، بل إلى الكشف عن البُعد الذاتي والوجداني الكامن فيها.

ضمن معالجة اللغة الطبيعية، يُصنَّف تحليل المشاعر ضمن المستوى الدلالي والتداولي، إذ يتجاوز فهم معاني الكلمات المفردة أو بنية الجمل ليصل إلى فهم الموقف الذي يحمله النص والنية الكامنة وراءه. ولهذا فهو يستلزم تضافر عدة مستويات تحليلية تشمل معالجة المفردات والبنية النحوية والسياق الدلالي والمعطيات التداولية في آنٍ واحد، مما يجعله من أكثر مهام NLP تعقيدًا وثراءً [6] .

وقد شهد هذا المجال نموًا متسارعًا في السنوات الأخيرة مدفوعًا بعاملين رئيسيين، أولهما الانتشار الهائل للمحتوى النصي على الإنترنت ووسائل التواصل الاجتماعي، وثانيهما التقدم الكبير في تقنيات التعلم الآلي والتعلم العميق التي وفّرت أدوات أكثر قدرة على فهم الفروق الدقيقة في اللغة البشرية [7].



المخطط 1.1 : مقاربة متعددة المستويات لتحليل المشاعر

2.5.1 أنواع تحليل المشاعر

صُنِّف تحليل المشاعر وفق عدة تصنيفات، أبرزها التصنيف القائم على الاتجاه العاطفي (Sentiment Polarity) الذي يُميّز بين ثلاثة أنواع رئيسية [7] :

يُمثّل المشاعر الإيجابية (Positive) وهي النصوصُ التي تعبّر عن رضا أو تأييد أو إعجاب تجاه موضوع ما، كتعليق مستهلك يُشيد بجودة منتج أو يُعبّر عن رضاه عن تجربة شرائية. في المقابل، تشير المشاعر السلبية (Negative) إلى النقد أو الرفض أو الاستياء، كتعليق يُعبّر عن خيبة أمل تجاه منتج أو خدمة. أما المشاعر المحايدة (Neutral) فتُمثّل

النصوص التي تخلو من الحمولة الوجدانية الواضحة، كالتعليقات الإخبارية أو الاستفسارات الوصفية التي لا تُعبّر عن موقف محدد.

وتجدر الإشارة إلى أن هذا التصنيف الثلاثي، رغم بساطته، يواجه تحديات حقيقية في التطبيق العملي، إذ كثيرًا ما تحمل النصوص مشاعر مختلطة ومتداخلة أو أساليب بلاغية كالتهكم والسخرية التي تجعل المعنى الحقيقي مختلفا عن الظاهر [7]. ويزداد هذا التحدي تعقيدًا في اللهجة الجزائرية الدارجة نظرًا لغنى تعابيرها العامية وكثرة استخدام السخرية فيها.

3.5.1 مستويات تحليل المشاعر

هناك ثلاثة مستويات رئيسية لتحليل المشاعر تتدرج من الأعم إلى الأكثر تخصصًا [7] :

أولًا: المستوى الوثائقي (Document Level): يهتم بالنص ككل، ويسعى إلى تحديد التوجه العاطفي العام الذي يحمله. ويفترض ضمنيًا أن الوثيقة بأكملها تتحدث عن موضوع واحد وتحمل موقفًا متسقًا منه، وهو افتراض يناسب التعليقات القصيرة والمراجعات الموجزة، لكنه يُظهر محدوديته مع النصوص الأطول التي قد تتضمن آراءً متعددة ومتباينة؛

ثانيًا: المستوى الجملي (Sentence Level): يعمل هذا المستوى على تحليل كل جملة أو وحدة نصية قصيرة (مثل تعليقات منصات التواصل الاجتماعي) بشكل منفصل، بهدف تحديد التوجه العاطفي الذي تعبر عنه. وقد يتضمن في بعض التطبيقات التمييز بين الجمل الذاتية التي تعكس آراءً ومشاعر، والجمل الموضوعية التي تعرض حقائق أو معلومات. ويتيح هذا المستوى معالجة أكثر دقة للنصوص، كما يناسب طبيعة التعليقات والمراجعات الموجزة؛

ثالثًا: المستوى القائم على الجوانب (Aspect Level): وهو الأكثر دقة، إذ لا يكتفي بتحديد المشاعر بشكل عام، بل يقوم بربط كل المشاعر بالجانب المحدد الذي تستهدفه. فمراجعة منتج قد تحمل إعجابًا بتصميمه وانتقادًا لسعره في الوقت ذاته، وهو ما لا يستطيع المستوى الوثائقي أو الجملي التقاطه بدقة [7] . ويُعتبر هذا المستوى من أكثر الاتجاهات تطورًا في أبحاث تحليل المشاعر بسبب تعقيده وثرائه المعلوماتي[6].

4.5.1 تطبيقات تحليل المشاعر في التسويق ومر اقبة العلامات التجارية

تتعدد مجالات توظيف تحليل المشاعر في السياق التسويقي وتغطي جوانب استراتيجية مهمة. وقد كانت المؤسسات التجارية من أوائل المستفيدين من هذه التقنية، نظرًا لما توفره من قدرة على رصد الرأي العام تجاه علاماتها التجارية ومنتجاتها بصورة آلية وفي الوقت الفعلي [7] .

وفي مجال مراقبة سمعة العلامة التجارية، تُمكّن أنظمة تحليل المشاعر المؤسساتِ من تتبع ما يُقال عنها في الفضاء الرقمي لحظةً بلحظة، و الاكتشاف المبكر لأي موجة انتقادية قبل أن تتحول إلى أزمة علاقات عامة. وعلى صعيد تطوير المنتجات، تُوفّر تعليقات المستخدمين وتحليل مشاعرها ملاحظات حقيقية تساعد في تحسين المنتجات بدلا من الاعتماد الاستبيانات التقليدية المكلفة [7] .

أما في مجال تحليل المنافسين فيُتيح تحليل مشاعر المستهلكين تجاه المنتجات المنافسة استخلاص نقاط القوة والضعف في عروض المنافسين، واكتشاف الفجوات التي يمكن للمؤسسة استثمارها تنافسيًا. وتُشير الدراسات إلى أن المؤسسات التي توظّف هذه التقنيات في قراراتها الإستراتيجية قادرة على التكيف مع تغيرات السوق مقارنةً بتلك التي تعتمد على أساليب جمع المعلومات التقليدية [7] .

وبناءً على ذلك، تبرز أهمية المنصة التي تسعى هذه الدراسة إلى تصورها، والتي تجمع بين تحليل المشاعر وذكاء التسويق لتُقدّم للمؤسسات الجزائرية أداةً متكاملة تُحوّل التعليقات الخام إلى رؤى إستراتيجية قابلة للتوظيف المباشر في القرار التسويقي.

6.1 خصوصية اللهجة الجز ائرية الدارجة

1.6.1 خصائص اللهجات العربية المحلية حاسوبيًا

تُشكّل اللهجات العربية المحلية تحديًا حقيقيًا أمام أنظمة معالجة اللغات الطبيعية، وذلك بسبب جملة من الخصائص البنيوية التي تُميّزها جذريًا عن اللغة العربية الفصحى. وتنتمي اللغة العربية بشكل عام إلى فئة اللغات عالية الاشتقاقية (Morphologically Rich Languages)، مما يعني أن كلمةً واحدة قد تحمل معلومات صرفية ودلالية معقدة

تتطلب أدوات تحليلية متخصصة [8]. وتزداد هذه التعقيدات حدةً في اللهجات المحلية التي تتباعد في أغلب الأحيان عن المعيار الفصيح على المستويات الصوتية والصرفية والمعجمية والتركيبية.

وتتسم اللهجات العربية المحلية عمومًا بغياب معيار كتابي موحد، إذ تُكتب في الغالب بالطريقة التي يراها كل مستخدم مناسبة، مما يُؤدي إلى تعدد كبير في أشكال الكلمة الواحدة داخل النص ذاته. فضلًا عن ذلك، تختلف هذه اللهجات من منطقة إلى أخرى حتى داخل البلد الواحد، مما يجعل بناء نموذج موحد قادر على التعامل مع هذا التنوع مهمةً بالغة التعقيد. وتُعد اللهجة الجزائرية الدارجة مثالًا واضحًا على ذلك، نظرًا لتأثرها بعوامل تاريخية وجغرافية وثقافية أفرزت خصائص لغوية فريدة تزيد من صعوبة معالجتها حاسوبيًا.

2.6.1 ظاهرة المزج اللغوي

تُعدّ ظاهرة المزج اللغوي (Code-Switching) من أبرز السمات المميزة للخطاب الرقمي الجزائري، كما تُعد من العوامل التي تعقّد أداء أنظمة تحليل المشاعر. وتتجلى هذه الظاهرة من خلال الانتقال بين العربية الدارجة والفرنسية والإنجليزية داخل النص الواحد وأحيانًا داخل الجملة الواحدة [9]. ويرتبط ذلك بالطبيعة اللغوية المتعددة للمجتمع الجزائري الذي يحمل إرثًا تاريخيًا من التعدد اللغوي الناتج عن الحقبة الاستعمارية الفرنسية وانتشار المحتوى الرقمي الإنجليزي بين فئة الشباب.

ومن الناحية الحاسوبية، يسبب هذا التداخل عدة مشكلات لأنظمة معالجة اللغات الطبيعية. فعلى مستوى التعرف على اللغة (Language Identification)، تجد الأنظمة صعوبة في تحديد اللغة الأساسية للنص عندما تتداخل عدة لغات. وعلى مستوى التحليل الصرفي، تتطلب كل لغة قواعد مختلفة قد تتعارض مع بعضها داخل النص ذاته. أما على مستوى تمثيل المعنى، فإن نماذج الكلمات المدرّبة على لغة تجد صعوبة في فهم الدلالات الكاملة لنص متعدد اللغات [9]. إضافة إلى ذلك، كثيرًا ما تتضمن التعليقات الجزائرية كلمات فرنسية مكتوبة بأحرف عربية أو كلمات عربية مكتوبة بالحروف اللاتينية، وهو ما يزيد من تعقيد المعالجة الآلية.

3.6.1 الإيموجي و الاختصارات و التعابير العامية

لا تقتصر تعقيدات النص الجزائري الرقمي على المزج اللغوي، بل تمتد لتشمل عناصر أخرى، وفي مقدمة هذه العناصر الإيموجي (Emoji) التي أصبحت وسيلة أساسية في التواصل الرقمي، خاصة في التعليقات القصيرة التي يعتمد

فيها المستخدمون على التعبير البصري. وقد كشفت الدراسات أن الإيموجي لا تؤدي دورًا تكميليًا فحسب، بل كثيرًا ما تحمل المشاعر الرئيسية للرسالة وقد تعكس دلالة النص تمامًا في حالات التهكم والسخرية [10].

كما يستخدم كثير من الجزائريين نظامًا من الرموز الرقمية تحل فيها الأرقام محل الأصوات العربية التي لا مقابل لها في الأبجدية اللاتينية، كتوظيف الرقم 3 للتعبير عن حرف العين، والرقم 7 للحاء، والرقم 9 للقاف. ويُعرف هذا النظام بـ Arabizi وهو شائع بشكل خاص في أوساط الشباب الجزائري [9]. أما التعابير العامية فتُمثّل تحديًا إضافيًا إذ تحمل دلالات ثقافية ومحلية لا تُدرك إلا بمعرفة السياق الاجتماعي الجزائري، وهو ما يجعل فهمها صعبًا بالنسبة للنماذج المدرّبة على العربية الفصحى أو على لهجات عربية أخرى.

4.6.1 نقص الموارد اللغوية والبيانات المصنّفة

يُمثّل شُح الموارد اللغوية المتخصصة باللهجة الجزائرية العائقَ الأكبر أمام بناء أنظمة تحليل مشاعر فعالة. فعلى عكس اللغة الإنجليزية التي تتمتع بمئات قواعد البيانات المصنّفة والمعاجم الحاسوبية والنماذج المدرّبة مسبقًا، تكاد اللهجة الجزائرية تخلو من هذه البنية التحتية اللغوية [8].

ويظهر هذا النقص في عدة جوانب، منها غياب معاجم حاسوبية شاملة تُغطي المفردات العامية الجزائرية، وندرة قواعد البيانات النصية المعنونة (Annotated Corpora) اللازمة لتدريب نماذج التعلم الآلي والتعلم العميق. يُضاف إلى ذلك محدودية أدوات المعالجة المسبقة المتخصصة كمُجزّئات النصوص (Tokenizers) ومحللات الأشكال الصرفية (Morphological Analyzers) المصممة خصيصًا للدارجة الجزائرية [8]. وقد اضطرت هذه الدراسة إلى بناء قاعدة بيانات مصنّفة يدويًا انطلاقًا من الصفر، وهو ما يبرز حجم النقص الموجود في هذا المجال.

5.6.1 مقارنة اللهجة الجزائرية باللغة العربية الفصحى

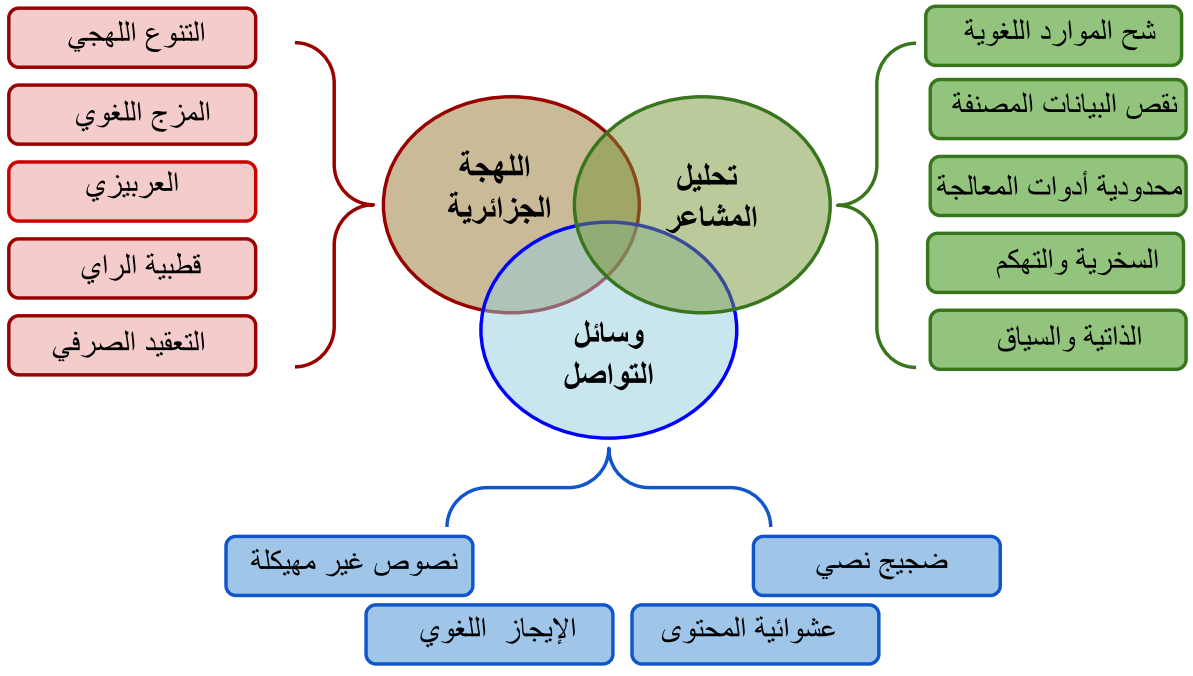
تتباين اللهجة الجزائرية الدارجة عن اللغة العربية الفصحى (Modern Standard Arabic - MSA) في عدة مستويات لغوية، مما يجعل نقل الأدوات والنماذج المطورة للفصحى إلى الدارجة أمرًا يستلزم تكيفًا لا مجرد تعديلات طفيفة. إذ تمتاز اللغة العربية الفصحى بنظام صرفي منضبط ومقنّن، بينما تتميز اللهجة الجزائرية بمرونة أكبر وتأثر واضح بالاقتراض اللغوي والتعابير المحلية [8].

ويمكن إجمال أبرز الفوارق بين الفصحى والدارجة الجزائرية من المنظور الحاسوبي في الجدول التالي:

المعيار	العربية الفصحى	الدارجة الجزائرية
الموارد اللغوية	وفيرة و متنوعة	شحيحة ومحدودة
المزج اللغوي	نادر	واسع الانتشار
النماذج المتاحة	متعددة ومتطورة	قليلة
التحليل الصرفي	أدوات ناضجة	أدوات في طور النشأة
البيانات المصنفة	متوفرة بكثرة	قليلة
المعيار الكتابي	موحد وموثق	غائب وغير منضبط

الجدول 1.1: مقارنة بين العربية الفصحى والدارجة الجزائرية في المعالجة الحاسوبية

ويتضح من هذه المقارنة أن بناء نظام فعال لتحليل مشاعر النصوص الجزائرية لا يمكن أن يرتكز فقط على النماذج المطورة للفصحى، بل يستلزم مقاربةً بحثيةً مستقلة تأخذ في الحسبان الخصوصية اللغوية والثقافية للدارجة الجزائرية في جميع مراحل بناء النظام، من جمع البيانات وصولًا إلى اختيار النماذج و تقييم أدائها.



المخطط 2.1: التحديات المشتركة بين معالجة اللهجة الدارجة الجزائرية وتحليل المشاعر في سياق المحتوى الرقمي

7.1 الخاتمة :

استعرض هذا الفصل الإطار النظري الذي يرتكز عليه البحث، من خلال خمسة نقاط متكاملة سعت في مجملها إلى بناء الخلفية المعرفية اللازمة لفهم الخيارات المنهجية والتقنية التي ستُناقش في الفصول اللاحقة.

انطلق الفصل من السياق التسويقي للدراسة، إذ تبيّن أن التحول الرقمي أعاد رسم ملامح العلاقة بين المؤسسات وعملائها، وجعل من البيانات الرقمية المتولدة على منصات التواصل الاجتماعي مصدرًا استراتيجيًا لا غنى عنه في صياغة القرارات التسويقية. ثم انتقل إلى تقنية كشط الويب بوصفها الأداة العملية لجمع هذه البيانات، مع الوقوف عند حدودها التقنية واعتباراتها القانونية والأخلاقية. وتوسّع بعد ذلك في مجال معالجة اللغات الطبيعية وتحليل المشاعر، مُبيّنًا الأسس النظرية التي تقوم عليها هذه التقنيات وتطبيقاتها في السياق التسويقي. واختتم بتشخيص دقيق للتحديات الخاصة التي تطرحها اللهجة الجزائرية الدارجة على أنظمة المعالجة الآلية، والتي تُمثّل في جوهرها المسوّغ الرئيسي لهذا البحث وإضافاته العلمية.

يبدو جليا من خلال هذا الفصل أن المعالجة الآلية للهجة الجزائرية لا تزال تفتقر إلى البنية التحتية اللغوية الكافية، وأن الأدوات المطورة للغة العربية الفصحى أو اللغات الأكثر توثيقًا لا يمكن نقلها مباشرةً إلى هذا السياق دون تكيّف منهجي معمّق. وهذا بالضبط ما يدفع الدراسة نحو المقاربة التجريبية المقارنة التي ستُعرض في الفصلين التاليين.

وعليه، يُمثّل الفصل الثاني امتدادًا طبيعيًا لهذا الإطار النظري، إذ سينتقل من المفاهيم العامة إلى التعمق في التقنيات والنماذج الخوارزمية المستخدمة في تحليل المشاعر، من نماذج التعلم الآلي التقليدي إلى التعلم العميق وصولًا إلى أحدث نماذج المحولات، تمهيدًا للجانب التطبيقي الذي يُشكّل صلب هذه الدراسة.

الفصل الثاني : تقنيات تمثيل النصوص ونماذج تحليل المشاعر

1.2مقدمة :

يواجه أي نظام يسعى إلى فهم المشاعر في النصوص تحدياً جوهرياً، قبل كل شيء كيف يُحوّل اللغة الإنسانية بكل تعقيدها إلى صورة يمكن للحاسوب معالجتها دون أن يفقد في هذا التحويل جوهر المعنى؟ من هذا التساؤل تنبثق الحاجة إلى استيعاب التطور التقني الذي شكّل أدوات تحليل المشاعر على مدى العقود الأخيرة، وهو ما يتصدى له هذا الفصل.

ينتقل الفصل من الإطار المفاهيمي الذي أرساه الفصل الأول إلى التعمق في الأدوات الخوارزمية والنماذج الحاسوبية التي يرتكز عليها البحث، وقد اقتضت طبيعة الموضوع استعراض ثلاثة أجيال متعاقبة من تقنيات تحليل المشاعر، لا لاستعراضها بحد ذاتها، بل لأن فهم نقاط قوة كل جيل و قصوره هو الذي يُبرر الخيارات التجريبية المعتمدة لاحقاً.

ينطلق الفصل من مسألة تمثيل النصوص بوصفها المدخل الأساسي لأي نظام تحليل نصي، ثم ينتقل إلى نماذج التعلم الآلي التقليدي، فنماذج التعلم العميق التي أحدثت نقلة نوعية في فهم البنية الدلالية للنصوص، وصولاً إلى نماذج المحولات التي باتت تُمثّل المرجعية السائدة في معالجة اللغات الطبيعية حتى وقت إجراء هذا البحث. ويختتم الفصل بمقارنة تحليلية بين هذه المقاربات الثلاث توفر الأساس المنطقي للانتقال إلى الفصل التجريبي.

2.2 تمثيل النصوص

1.2.2 مفهوم تمثيل النص وأهميته

تمثيل النصوص (Text Representation) هو عملية تحويل النص الخام إلى صيغة رقمية يمكن للنماذج الحاسوبية معالجتها. وجودة التمثيل المختار تُحدد في نهاية المطاف السقف الأعلى لأداء أي نموذج تحليل نصي، إذ لا يمكن لأي نموذج أن يتجاوز حدود المعلومات التي يوفرها له هذا التمثيل [6]. وبعبارة أخرى، التمثيل الجيد يحمل المعنى، والتمثيل الرديء يُضيّع المعنى قبل أن يصل النموذج إليه.

وتكمن أهمية التمثيل حول ثلاثة أبعاد متكاملة: أولها البُعد الدلالي، وهو مدى قدرة التمثيل على الإبقاء على المعنى الحقيقي للكلمات والعلاقات بينها. وثانيها البُعد الحسابي، وهو حجم الذاكرة والوقت اللازمَين لمعالجة هذا التمثيل. وثالثها البُعد التعميمي، وهو التعامل مع كلمات جديدة لم تظهر في بيانات التدريب [6].

2.2.2 مفهوم التمثيل التقليدي القائم عن الاحصاء

يُقصد بالتمثيل التقليدي للنصوص (Traditional Text Representation) تحويل النصوص إلى تمثيلات عددية تعتمد أساسًا على التكرار الإحصائي للكلمات داخل الوثائق، دون أخذ المعنى السياقي أو العلاقات الدلالية العميقة بين الكلمات بعين الاعتبار. وتعتمد هذه المقاربات على اعتبار النص كحقيبة من الكلمات (Bag of Words)، حيث يتم تمثيل كل وثيقة على شكل متجه يعكس وجود الكلمات أو تكرارها داخلها.

وتُعد هذه الأساليب من أقدم تقنيات تمثيل النصوص في معالجة اللغات الطبيعية، وقد استُخدمت في العديد من مهام التصنيف النصي واسترجاع المعلومات، نظرًا لبساطتها وسهولة تنفيذها، رغم محدوديتها في تمثيل المعنى والسياق اللغوي [6].

1.2.2.2 نموذج (TF-IDF) و (n-grams):

يُمثّل نموذج TF-IDF (Term Frequency–Inverse Document Frequency) أحد أبرز أساليب التمثيل النصي الإحصائي، إذ يهدف إلى تحديد أهمية كل مصطلح داخل الوثيقة بالنسبة إلى مجموعة الوثائق الكلية، من خلال الجمع بين تكرار المصطلح داخل الوثيقة وندرته عبر مجموعة البيانات. ويقيس TF-IDF هذه الأهمية بضرب تكرار المصطلح في الوثيقة (TF) في معكوس تكراره عبر مجموعة الوثائق (IDF)، مما يمنح المصطلحات النادرة المميزة ذات الدلالة العالية وزنًا أكبر، ويُقلّص وزن المصطلحات الشائعة [6]، وفق المعادلة (01) :

$$TF - IDF\,(t,d) = TF(t,d) \times IDF(t) \quad \text{.....01}$$

و للحفاظ على قدر من السياق اللغوي وترابط الكلمات، اعتُمد هذا الأسلوب في الدراسة الحالية مدعّمًا بتقنية N-grams، مما يُتيح استخراج الميزات على مستوى الكلمات المفردة (Unigrams)، وكذلك التسلسلات النصية المتعددة مثل Bigrams وTrigrams وينتج عن ذلك متجهات متناثرة (Sparse Vectors) تعكس الخصائص اللغوية للنص وتساعد

على التقاط بعض الأنماط المحلية داخل الجمل. غير أن هذا النهج يظل محدودًا في تمثيل العلاقات الدلالية العميقة، كعلاقات الترادف والتضاد وفهم السياق البعيد المدى.

3.2.2 مفهوم التضمين الدلالي

ظهرت التضمينات الدلالية استجابةً للقصور الذي تعانيه أساليب التمثيل التقليدية. وتقوم فكرتها الأساسية على الفرضية التوزيعية (Distributional Hypothesis) التي صاغها Harris، ومفادها أن: "تُعرف الكلمة من خلال مجاورتها"، أي أن الكلمات التي تتشارك في السياقات نفسها غالبًا ما تحمل معانيَ متقاربة. وتُمثَّل كل كلمة وفق هذا المبدأ كمتجه كثيف الأبعاد (Dense Vector) في فضاء متعدد الأبعاد يعكس علاقاتها الدلالية مع بقية الكلمات [11].

1.3.2.2 نموذج FastText

طوّر بوجانوفسكي وزملاؤه [12] من مختبر FAIR في شركة Meta نموذج **FastText**، الذي يُعد تطويرًا لنموذج Word2Vec، وهو نموذج لتمثيل الكلمات (Word Embedding) يحوّل الكلمات إلى متجهات رقمية تعكس العلاقات الدلالية بينها. ويعتمد FastText على بنية Skip-gram، وهي إحدى البنيتين الأساسيتين في Word2Vec، حيث تهدف إلى التنبؤ بالكلمات المحيطة انطلاقًا من الكلمة المستهدفة. وقد أضاف FastText تحسينًا جوهريًا يتمثل في معالجة الكلمات على مستوى المقاطع الحرفية (Character n-grams) بدلًا من الاعتماد على الكلمة الكاملة كوحدة مستقلة. وتبرز أهمية هذا النهج، خاصة عند التعامل مع اللغة العربية، نظرًا لخصائصها الصرفية المعقدة التي تُنتج أشكالًا اشتقاقية متعددة قد يصعب على نموذج Word2Vec تمثيلها بكفاءة، بينما يتمكن FastText من إنتاج تمثيلات عالية الجودة حتى للكلمات النادرة أو غير الموجودة في بيانات التدريب (**Out-of-Vocabulary - OOV**) [13].

4.2.2 مفهوم التمثيلات السياقية باستخدام نماذج المحوّلات

تعتمد التمثيلات السياقية على إنتاج تمثيل مختلف للكلمة الواحدة بحسب السياق الذي ترد فيه داخل الجملة، بخلاف التضمينات الساكنة مثل FastText التي تمنح تمثيلاً ثابتاً لكل كلمة. وقد ساهمت نماذج المحوّلات (Transformers) الحديثة في تطوير هذا النوع من التمثيل، مما أتاح فهماً أدق للعلاقات الدلالية والسياقية داخل النصوص [14][15].

5.2.2 الفرق بين التمثيل التقليدي والدلالي

يلخّص الجدول 2 أوجه المقارنة الجوهرية بين الأساليب الثلاثة

المعيار	TF-IDF	FastText
طبيعة التمثيل	متجه متناثر إحصائي	متجه كثيف ثابت
الفهم الدلالي	معدوم	سياق جزئي
تعدد المعاني Polysemy	لا يدعمه	لا يدعمه
الكلمات النادرة oov	غير مدعوم	يدعم
التكلفة الحسابية	منخفضة جدا	منخفضة إلى متوسطة
متطلبات البيانات	منخفضة	متوسطة
الملائمة للعربية	محدودة	جيد

الجدول 1.2: يبين الجدول مقارنة بين أساليب تمثيل النصوص

وقد أثبتت الدراسات المقارنة أن التمثيلات السياقية كـ MARBERT تحقق أداءً أفضل بفارق ملحوظ في مهام معالجة اللغة العربية. إذ أشارت نتائج [16] إلى تفوق MARBERT على XLM-R Large بفارق ملحوظ على مقياس F1 الماكروي عبر مهام متعددة. ويعود هذا التفوق إلى قدرة التمثيلات السياقية على تحديد معنى الكلمة بحسب سياقها الفعلي،على عكس النماذج TF-IDF و FastText التي تعتمد على تمثيلات ثابتة للكلمة دون مراعاة السياق[12].

3.2 التعلم الآلي التقليدي في تصنيف النصوص

1.3.2 مفهوم التصنيف النصي

التصنيف النصي (Text Classification) هو مهمة تعلم آلي إشرافي تهدف إلى تعيين كل وثيقة نصية إلى فئة أو أكثر من فئات محددة مسبقًا. وفي تحليل المشاعر تكون هذه الفئات عادةً: إيجابي،سلبي ومحايد. وقد صنّف Bishop هذا النوع من المشاكل ضمن مسائل التصنيف الإشرافي (Supervised Classification) ، حيث يتعلم النموذج من أمثلة مُعنونة قاعدةً لتصنيف الأمثلة الجديدة غير المرئية[17].

2.3.2 خوارزمية SVM و مبدأ عملها

تُعدّ آلة المتجهات الداعمة (Support Vector Machine - SVM) من أكثر خوارزميات التعلم الآلي فعاليةً و موثوقيةً في تصنيف النصوص. ويقوم مبدأ عملها على إيجاد المستوى الفاصل الأمثل (Optimal Hyperplane) الذي يُعظّم الهامش (Margin) بين فئتين في الفضاء متعدد الأبعاد [17]. ويُعرَّف الهامش بأنه المسافة بين المستوى الفاصل و أقرب نقاط البيانات من كل فئة، و التي تُعرف بالمتجهات الداعمة (Support Vectors).

4.2 نماذج التعلم العميق للنصوص

1.4.2 مفهوم التعلم العميق وأهميته في NLP

يُعرَّف التعلم العميق (Deep Learning) بأنه فرع من فروع التعلم الآلي يعتمد على شبكات عصبية اصطناعية ذات طبقات متعددة (Deep Neural Networks) قادرة على تعلم تمثيلات هرمية متزايدة التجريد من البيانات الخام. وقد أشار Goodfellow وآخرون [18] إلى أن ما يُميّز التعلم العميق عن التعلم الآلي التقليدي هو قدرته على استخراج الخصائص (Feature Extraction) تلقائياً دون الحاجة إلى هندستها يدوياً .

وفي مجال NLP، ساهم التعلم العميق في تحسين فهم العلاقات اللغوية المعقدة، مثل التبعيات البعيدة بين الكلمات والبُنى التركيبية الدقيقة، وهي أمور كانت صعبة على النماذج التقليدية.

2.4.2 الشبكات العصبية الالتفافية CNN و تطبيقها على النصوص

صُمّمت الشبكات العصبية الالتفافية (Convolutional Neural Networks - CNN) في الأصل لمعالجة الصور والبيانات ثنائية الأبعاد، غير أنها أثبتت كفاءة استثنائية في مهام معالجة اللغات الطبيعية وتصنيف النصوص. وتعمل الـ CNN في هذا السياق عبر معالجة النص كإشارة أحادية البعد (1D Signal)، متمثلة في مصفوفة ميزات متسلسلة يُطبق عليها مجموعة من المرشحات الالتفافية (Convolutional Filters) القادرة على التقاط الأنماط الدلالية المحلية (Local Patterns) والروابط السياقية بين العبارات بصرف النظر عن موقعها الجملي [18].

5.2 نماذج المحولات

1.5.2 مفهوم معمارية Transformer و آلية الانتباه

أحدث نشر ورقة "Attention Is All You Need" عام 2017 من قِبَل Vaswani وآخرين [14]، تحولًا كبيرًا في مجال معالجة اللغات الطبيعية، إذ قدّمت معمارية المحوّل (Transformer) التي تتخلى كلياً عن التكرار والتلافيف لصالح آلية الانتباه الذاتي (Self-Attention).

وتقوم آلية الانتباه الذاتي على مبدأ جوهري يُتيح لكل موقع في التسلسل حساب درجة انتباهه تجاه كل موقع آخر في نفس التسلسل، بغض النظر عن المسافة بينهم. وهذا يُمكّن النموذج من التقاط التبعيات البعيدة المدى مباشرةً دون الحاجة إلى المعالجة المتتابعة كما في RNN وتُحسب درجات الانتباه وفق الصيغة (02):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right) \times V \quad \text{.....02}$$

Q (Query) : الاستعلام (الكلمة الحالية المراد تحليلها)

K(Key) : المفاتيح (جميع الكلمات للمقارنة)

V (Value) : القيم (المعلومات المستخرجة بعد الانتباه)

d_K : بعد المتجهات (للتطبيع)

softmax : تحويل القيم إلى أوزان احتمالية

QK^T: قياس التشابه بين الكلمات

2.5.2 نموذج MARBERT

قدّم عبد المجيد وزملاؤه [16] في ورقتهم البحثية المنشورة ضمن وقائع مؤتمر ACL 2021 نموذجَين متخصصَين للغة العربية ARBERT و MARBERT . ويُعدّ Masked Arabic BERT الأوسع تغطيةً من الناحية اللهجية، إذ دُرِّب على 128 مليار رمز لغوي (token) مستخرجة من مليار تغريدة عربية تشمل العربية الفصحى الحديثة (MSA) والعربية الدارجة بلهجاتها المتعددة. وتتمثل أهم خصائصه في حول:

- التغطية اللهجية الشاملة: تجاوز قيود النماذج السابقة كـ AraBERT المدرَّب أساساً على نصوص الفصحى؛
- التمثيل السياقي الديناميكي: إنتاج تمثيل مختلف للكلمة ذاتها بحسب سياقها، خلافاً للتضمينات الساكنة كـ Word2Vec وFastText؛
- الأداء المتفوق : تحقيق أعلى النتائج على معيار ArBench الذي يضم 41 مجموعة بيانات عبر 5 مهام مختلفة، متجاوزاً XLM-R Large رغم صغر حجمه النسبي[16] .

3.5.2 مفهوم Fine-Tuning و أهميته

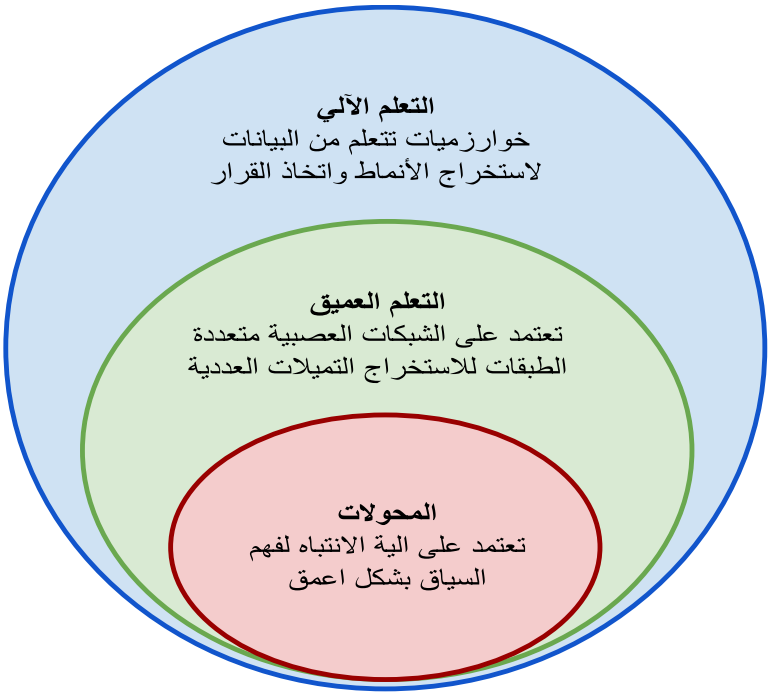
الضبط الدقيق (Fine-Tuning) هو أسلوب يقوم على تكييف نموذج مُدرَّب مسبقاً على بيانات ضخمة (Pre-trained Model) لأداء مهمة محددة أو التعامل مع مجال متخصص، وذلك بتدريبه مجدداً على بيانات أصغر حجماً وأكثر تخصصاً. وقد أشار Devlin وآخرون [15] إلى أن هذا الأسلوب يُتيح تحقيق أداء عالٍ حتى مع مجموعات بيانات صغيرة نسبياً، إذ يُسهم النموذج المُدرَّب مسبقاً بمعرفة لغوية عامة تُغني عن الحاجة إلى بيانات ضخمة في مرحلة التكييف.

وفي سياق هذه الدراسة، يكتسب Fine-Tuning لنموذج MARBERT أهمية مضاعفة. إذ يُتيح تكييف النموذج مع خصوصية اللهجة الجزائرية الدارجة ومجال منتجات العناية بالبشرة ، وهو تحدٍّ مزدوج لا تستطيع النماذج الجاهزة التعامل معه بكفاءة دون هذه المرحلة.

6.2 مقارنة عامة بين المقاربات الثلاث

1.6.2 نقاط القوة والضعف لكل مقاربة

توضح الصورة التالية التطور الهرمي لمقاربات معالجة النصوص من الأساليب التقليدية إلى التعلم العميق فالنماذج الحديثة المعتمدة على المحولات.



المخطط 1.2 : توضح الصورة مستويات نماذج معالجة النصوص من التقليدي إلى المحوّلات

كما يوضح الجدول التالي مقارنة بين مقاربات تمثيل النصوص ونماذج التعلم المختلفة من حيث مجموعة من المعايير الأساسية التي تعكس نقاط القوة والضعف لكل مقاربة.

المعيار	التعلم الآلي	التعلم العميق	المحولات
التمثيل السياقي	غائب	جزئي	كامل
قابلية التفسير	مرتفعة	منخفضة	منخفضة جدا
الأداء العام	متوسط	جيد	أعلى
التعامل مع اللهجات	محدود	متوسط	افضل مع Fine-Tuning
الموارد الحسابية	منخفضة	مرتفعة	مرتفعة جدا
متطلبات البيانات	منخفضة	مرتفعة	متوسطة مع Fine-Tuning
سرعة التدريب	سريعة جدا	متوسطة	بطيئة

الجدول 2.2 : يوضح مقارنة تحليلية بين المقاربات الثلاث المعتمدة

2.6.2 معايير اختيار النموذج المناسب

لا يوجد نموذج أفضل في المطلق، بل يتوقف الاختيار على جملة من المعايير المتشابكة. وقد أشار Goodfellow وآخرون إلى أن المفاضلة بين النماذج ينبغي أن تأخذ بعين الاعتبار حجم البيانات المتاحة والموارد الحسابية والهدف من التطبيق [18]. وفي سياق هذه الدراسة، دفعت ثلاثة اعتبارات نحو اعتماد المقاربة المقارنة الشاملة، أولها شُح البيانات المصنّفة باللهجة الجزائرية الذي يجعل النماذج الخفيفة خياراً منافساً، وثانيها الرغبة في قياس الفجوة الفعلية في الأداء بين الأجيال الثلاثة على بيانات محلية، وثالثها أن الهدف التطبيقي النهائي (منصة الذكاء التسويقي) يستلزم توازناً بين الدقة وسرعة الاستدلال.

7.2 الخاتمة

استعرض هذا الفصل المنظومة التقنية الكاملة التي تقوم عليها هذه الدراسة، منتقلاً من أساليب تمثيل النصوص إلى نماذج التصنيف عبر ثلاثة أجيال تقنية متعاقبة. وقد تبيّن أن كل جيل يُمثّل استجابةً للقصور الذي كشفه الجيل السابق. فالتعلم الآلي التقليدي أتاح التصنيف الأسرع والأكثر تفسيريةً، والتعلم العميق أضاف القدرة على استيعاب البنية التسلسلية والسياق المحلي، فيما جاءت المحولات لتُحقق ما عجز عنه سابقاها من استيعاب كامل للسياق ثنائي الاتجاه.

وتكشف المقارنة بين هذه المقاربات أنه لا يوجد نموذج متفوق في كل الأبعاد، بل ثمة مبادلات حقيقية بين الأداء والتفسيرية والكفاءة الحسابية ومتطلبات البيانات. وهذا بالضبط ما يُسوّغ المنهجية المقارنة التي تتبناها هذه الدراسة، إذ لا يمكن الحكم على أنسب نموذج لتحليل مشاعر التعليقات الجزائرية دون اختبارها ميدانياً على بيانات حقيقية تعكس تعقيداتها اللغوية الخاصة.

وتستند نتائج هذا الفصل مباشرةً إلى الفصل الثالث، الذي يُجسّد هذه المقاربات النظرية في تجارب فعلية على بيانات تعليقات يوتيوب الجزائرية، ويُقيّم أداءها مقارنةً وفق مقاييس موضوعية.

الفصل الثالث: الجانب التطبيقي والتجريبي

1.3 المقدمة

يُمثّل هذا الفصل الجوهر التطبيقي للدراسة، إذ يُجسّد المفاهيم النظرية والتقنية التي أُرسيت في الفصلين السابقين ضمن منهجية تجريبية متكاملة تنتهي بنتائج قابلة للقياس والتحليل. وقد انتظم العمل التطبيقي في سلسلة من المراحل المتتابعة، بدأت بجمع البيانات الخام من منصة يوتيوب، مروراً بمعالجتها وتنظيفها وإعدادها للتدريب، ثم تدريب سبعة نماذج موزعة على ثلاث مقاربات تقنية مختلفة ومقارنة أدائها، وقد سمحت هذه المرحلة باستخلاص مجموعة من الاستنتاجات الأساسية حول أداء النماذج المعتمدة.

وتتميّز هذه الدراسة باعتمادها على بيانات حقيقية مكتوبة باللهجة الجزائرية الدارجة في مجال منتجات العناية بالبشرة، وهو سياق لغوي وتطبيقي لم يحظَ باهتمام بحثي كافٍ، مما يمنح نتائجها قيمةً مضافة على الصعيدين الأكاديمي والتطبيقي معاً.

2.3 وصف البيانات وجمعها

1.2.3 مصدر البيانات و مبرر اختياره

اعتمدت هذه الدراسة على منصة يوتيوب مصدراً أساسياً و حصرياً للبيانات، ويعود هذا الاختيار إلى جملة من المبررات المنهجية المتكاملة. فمن الناحية اللغوية، تحتضن المنصة كماً هائلاً من التعليقات المكتوبة باللهجة الجزائرية الدارجة في سياقات استهلاكية تلقائية وأصيلة، بعيداً عن أي توجيه أو تأثير تسويقي مباشر. ومن الناحية التقنية، توفر المنصة واجهة برمجية رسمية (YouTube Data API v3) تُتيح الوصول المنظم والقانوني إلى بياناتها. أما من الناحية الموضوعية، فإن فيديوهات مراجعة منتجات العناية بالبشرة تحظى بحضور واسع على المنصة وتستقطب تعليقات غنية تعكس تجارب المستهلكين الجزائريين الفعلية مع هذه المنتجات.

2.2.3 العلامات التجارية المستهدفة ومعايير اختيارها

جرى اختيار العلامات التجارية المستهدفة وفق معيار منهجي محدد يقوم على تجانس خط الإنتاج، إذ اقتصر الاختيار على العلامات التي تُنتج نفس فئات منتجات العناية بالبشرة التي تتخصص فيها شركة ماسيليا، وذلك لضمان

قابلية المقارنة بين التعليقات وتجانس السياق التسويقي للبيانات المجموعة. وقد شملت الدراسة مجموعة من العلامات التجارية الناشطة في السوق الجزائرية والمتخصصة في منتجات العناية بالبشرة، من بينها :(Venus، Caledarme، Darmasover، Shaline، Lizas، Biova، Carisons، Nial، Broderma، Zora، Toshi، Selinaz، Swalis، Biolily، Felix Car، Abou Saad)

3.2.3 آلية جمع البيانات عبر YouTube Data API v3

1.3.2.3 وصف الأداة المستخدمة

اعتُمدت YouTube Data API v3 الواجهةُ البرمجية الرسمية التي توفرها Google أداةً وحيدة لجمع البيانات، وقد جاء هذا الاختيار لاعتبارات تقنية وقانونية متعددة. إذ تُقدّم هذه الواجهة ضمانات عالية من حيث الاستقرار والموثوقية مقارنةً بأساليب الكشط المباشر، كما تضمن الامتثال الكامل لشروط استخدام المنصة التي تحظر الوصول غير المُرخَّص إلى بياناتها.

2.3.2.3 منهجية الجمع البرمجي

صُمِّم برنامج جمع البيانات وفق المنطق البرمجي الموضح في الخوارزمية التالية:

تنطلق العملية بتهيئة خدمة يوتيوب باستخدام مفتاح API، ثم تستدعي دالة get_comments لكل فيديو مستهدف. تُرسل هذه الدالة طلبات متتالية إلى نقطة النهاية commentThreads بتحديد معرّف الفيديو (video_Id) وضبط معامل maxResults على 100 تعليق لكل طلب، مع استخراج التعليقات الرئيسية والردود عليها في آنٍ واحد. وتتولى آلية Pagination عبر معامل pageToken ضمان استخراج جميع التعليقات المتاحة دون اقتصار على الصفحة الأولى، وتتوقف الحلقة تلقائياً عند استنفاد جميع الصفحات. و تُستخرج من كل تعليق البيانات التالية:

- نص التعليق (comment text)؛
- معرّف التعليق (comment ID)؛
- اسم صاحب التعليق (author)؛
- تاريخ النشر (published date)؛
- عدد الإعجابات (like count)؛

وفي نهاية العملية، تُجمَّع كل التعليقات في قائمة موحدة تُحوَّل إلى إطار بيانات (DataFrame) وتُحفظ بصيغة CSV لمعالجتها في المراحل اللاحقة. وفي ما يلي الخوارزمية المعتمدة لعملية الكشط (Scraping) الخاصة بجمع البيانات من منصة يوتيوب.

```python
def get_comments(video_id, category, brand):
    comments_list = []
    next_page_token   None

    # استرجاع معلومات الفيديو
    video_response = youtube.videos().list(
        part="snippet", id=video_id).execute()
    if not video_response["items"]:
        return comments_list

    while True:
        # طلب التعليقات مع الردود
        response = youtube.commentThreads().list(
            part="snippet,replies",
            videoId=video_id,
            maxResults=100,
            pageToken=next_page_token,
            textFormat="plainText"
        ).execute()

        for item in response["items"]:
            # استخراج التعليق الرئيسي
            top_comment = item["snippet"]["topLevelComment"]["snippet"]
            comments_list.append({
                "video_id"     : video_id,
                "category"     : category,
                "brand"        : brand,
                "comment"      : top_comment[textDisplay],
                "comment_id"   : item["snippet"]["topLevelComment"]["id"],
                "author"       : top_comment["authorDisplayName"],
                "published_at": top_comment["publishedAt"],
                "like_count"   : top_comment["likeCount"],
                "parent_comment_id": None
            })

            # استخراج الردود إن وُجدت
            if "replies" in item:
                for reply in item["replies"]["comments"]:
                    reply_snippet = reply["snippet"]
                    comments_list.append({
                        "video_id"     : video_id,
                        "category"     : category,
                        "brand"        : brand,
                        "comment"      : reply_snippet["textDisplay"],
                        "comment_id"   : reply["id"],
                        "author"       : reply_snippet["authorDisplayName"],
                        "published_at": reply_snippet["publishedAt"],
                        "like_count"   : reply_snippet["likeCount"],
                        "parent_comment_id": reply_snippet["parentId"]
                    })

        # الانتقال للصفحة التالية
        next_page_token = response.get("nextPageToken")
        if not next_page_token:
            break

    return comments_list
```

الصورة 1.3: خوارزم عملية استخراج البيانات من منصة يوتيوب

3.3.2.3 التحديات التقنية المواجَهة

واجهت عملية الجمع جملةً من التحديات التقنية أبرزها قيود الحصص اليومية (API Quotas) التي تفرضها Google والتي استلزمت تخطيطاً زمنياً دقيقاً لتوزيع طلبات الجمع على فترات متعددة. كما أفرزت التعددية اللغوية في التعليقات ضرورة تطبيق مرشح لغوي لاحق لحصر البيانات في التعليقات العربية الدارجة الجزائرية .

4.2.3 وصف مجموعة البيانات الخام والتصنيف اليدوي

1.4.2.3 حجم البيانات و توزيعها

أسفرت عملية الجمع عن بناء قاعدة بيانات خام تضم 11,112 تعليقاً باللهجة الجزائرية الدارجة في مجال منتجات العناية بالبشرة. وبعد مراحل التنظيف والتصفية، احتُفظ بـ 8,267 تعليقاً صالحاً، خضعت هذه التعليقات لعملية تصنيف يدوي دقيق وُزِّعت فيها على ثلاث فئات شعورية، توضح الصورة التالية مجموعة من التعليقات المستخرجة كما هو موضح في التالي:

هاذي توالم بشرة جافة بإمتياز فقط
صحيح❤❤
شكرا حبيبتي نفس رأيي ❤❤❤❤❤
العفو غاليتي❤❤
السيروم روعة روعة اشتريتوا 40الف مكنتش متوقعة النتيجة الحقيقة منتوج وطني علامة يا بنات رخيص اشروه
اختي مليح للدسة ❤❤❤
@عبدالقادربوبكر-خ1رمليح لانه فيه فيتامين سي لازم برك دوامي عليه وكاين عند العشاب كريمة رخيصة وسمها ليمون در
واش جربتي اختي أنا راني نعاني
سلام. كفاه ديريه هو اول ولا لكخام نتاع شمس من فضلك
@Jdjdjdjd-tj6yxختي ترطبي وجهك ومن بعد تع الشمس
هو مشي بيض مثل حليب
لا شفاف
عندي 3 ايام ملي راني ندير فيه دارلي حساسية نحبسو ولا عادي؟؟
حبسيه
جربتوا مافيه حتى فائده والله ما تغير والا شي
بالعكس يعطي نضاره واشراقة

الصورة 2.3: عينة من التعليقات الخام المجمعة من منصة يوتيوب قبل المعالجة

وتكشف هذه الأرقام عن ظاهرة عدم توازن الفئات (Class Imbalance) التي تُمثّل أحد أبرز التحديات في هذه الدراسة، إذ تهيمن الفئة المحايدة بفارق واسع عن الفئتين الإيجابية والسلبية. وقد استلزم هذا الواقع اعتماد تقنيات خاصة للتعامل معه كما سيُفصَّل في 4.3 .

الفئة	العدد	النسبة التقريبية
محايد	5,643	68.3%
ايجابي	1,784	21.6%
سلبي	840	10.1%

جدول 1.3: يبين توزيع التعليقات على الفئات الثلاث

2.4.2.3 منهجية التصنيف اليدوي

نظراً لغياب أدوات تصنيف آلية موثوقة للهجة الجزائرية الدارجة، اعتُمد التصنيف اليدوي منهجيةً أساسية لبناء مجموعة البيانات المعنوَنة. وقد جرى هذا التصنيف وفق معايير محددة تأخذ في الاعتبار السياق الكامل للتعليق لا الكلمات المفردة فحسب، مع الاستناد إلى المعرفة اللغوية والثقافية الجزائرية للتعامل مع الحالات الملتبسة كالسخرية والتهكم وازدواجية المعنى. وتُعدّ هذه المجموعة المُصنَّفة يدوياً الأساس الذي تقوم عليه جميع مراحل التدريب والتقييم اللاحقة.

3.3 المعالجة المسبقة للنصوص

من أجل تعزيز جودة البيانات وتحسين أداء نماذج تحليل المشاعر، صُمِّم خط معالجة مسبقة مخصص يتكون من ثماني مراحل متتابعة ومتكاملة، تهدف في مجملها إلى تهيئة النصوص الخام بما يتوافق مع متطلبات تحليل اللهجة الجزائرية الدارجة المستخرجة من منصات التواصل الاجتماعي. وقد راعى تصميم هذا الخط خصوصيةَ البيانات المجمعة من حيث تنوع أساليب الكتابة وظاهرة المزج اللغوي وشيوع الرموز التعبيرية، وهي سمات لا تعالجها أدوات المعالجة الجاهزة المصممة للغة العربية الفصحى بشكل كافٍ.

1.3.3 المرحلة الأولى: تصفية البيانات

خضعت البيانات الخام في هذه المرحلة لعملية تصفية أولية هدفها حصر مجموعة البيانات في التعليقات ذات الصلة الفعلية بموضوع التحليل. وقد اشتملت هذه العملية على خطوتين رئيسيتين. تمثّلت الأولى في حذف التعليقات

المحتوية على إشارات المستخدمين (@mentions) والروابط الإلكترونية باعتبارها عناصر لا تحمل قيمة دلالية في سياق تحليل المشاعر. وتمثّلت الثانية في تطبيق مرشح لغوي قائم على التعبيرات النمطية (Regex) لاستبعاد التعليقات المكتوبة أساساً بالحروف اللاتينية، مع الإبقاء على التعليقات العربية والدارجة التي قد تتضمن كلمات أجنبية معزولة، وهو ما يعكس الواقع اللغوي الفعلي للمستخدم الجزائري. وقد أسهمت هذه الإجراءات في تحسين تجانس مجموعة البيانات واتساقها تمهيداً للمراحل اللاحقة.

2.3.3 المرحلة الثانية: تنظيف البيانات

طُبِّقت في هذه المرحلة دالة مخصصة للمعالجة على مستوى المحرف، قائمة على مبدأ الاحتفاظ بالعناصر ذات القيمة الدلالية وإزالة ما سواها. وقد شملت العملية الاحتفاظ بالحروف العربية والأرقام والرموز التعبيرية، مع إزالة جميع الرموز الخاصة وعلامات الترقيم وإشارات المستخدمين والعناصر غير ذات الصلة التي قد تُضيف ضجيجاً نصياً يؤثر سلباً على دقة التحليل. كما جرى توحيد تنسيق المسافات بدمج الفراغات المتعددة المتتالية في فراغ واحد، مما يضمن اتساق الشكل النهائي للنصوص ويُهيّئها للمراحل التالية.

3.3.3 المرحلة الثالثة: معالجة الرموز التعبيرية

تُمثّل الرموز التعبيرية (Emoji) وحدات دلالية ذات حمولة شعورية صريحة في نصوص وسائل التواصل الاجتماعي الجزائرية، مما يجعلها سمة لغوية فعّالة في تحديد التوجه الشعوري للنص لا يمكن إغفالها في سياق تحليل المشاعر. وعليه، جرى الحفاظ عليها خلال مرحلة التنظيف بدلاً من حذفها، واعتُمد في معالجتها قاموس مخصص (Emoji Sentiment Lexicon) مستورد من مجموعة بيانات متخصصة، يربط كل رمز تعبيري بمكافئه النصي باللغة العربية. ثم طُبِّقت عملية استبدال منتظمة تحوّل كل رمز إلى اسمه العربي المقابل مع إحاطته بمسافات فاصلة تكفل عزله عن المحتوى النصي المجاور تفادياً لأي تداخل يؤثر على دقة التحليل. ويوضح المثال التالي هذه العملية: قبل المعالجة " أنا راني نعاني 😢 " بعد المعالجة" *أنا راني نعاني وجه يبكي*".

ويوضح الكود البرمجي التالي آلية تنفيذ هذه العملية :

```
emoji_dict = dict(zip(emoji_df["Emoji"],
emoji_df["Arabic_Name"]))
def replace_emoji_with_arabic(text):
    if pd.isna(text):
        return text
    for emoji, arabic_word in emoji_dict.items():
        if emoji in text:
            text = text.replace(emoji, " "   arabic_word   " ")
    return text
```

الصورة 3.3: خوارزم عملية معالجة الرموز التعبيرية

4.3.3 المرحلة الرابعة: تطبيع النصوص للهجة الجزائرية

تُعدّ هذه المرحلة من أكثر مراحل خط المعالجة تخصصاً وأهميةً، إذ تهدف إلى توحيد أشكال الكلمات وتقليل التشتت اللغوي الناتج عن عفوية الكتابة في منصات التواصل الاجتماعي. وبدلاً من الاعتماد على قواعد التطبيع القياسية للغة العربية الفصحى التي لا تعكس خصوصية اللهجة المحلية، صُمِّمت دالة معالجة متخصصة تراعي الخصائص الصوتية والكتابية للهجة الجزائرية، وشملت أربع عمليات متكاملة:

- **توحيد الأنماط الكتابية:** جرى توحيد مختلف أشكال حرف الألف (أ، إ، آ) في صورة موحدة (ا)، وتحويل التاء المربوطة (ة) إلى (ه)، والألف المقصورة (ى) إلى (ي)، مما أسهم في تقليص حجم المفردات ورفع كفاءة المطابقة بين الكلمات؛
- **المعالجة الصوتية للدارجة:** عولجت الحروف الخاصة الشائعة في الكتابة الجزائرية الرقمية كتحويل (ڤ) إلى (ق) و(ڤ) إلى (ف)، لضمان إرجاع الكلمات إلى أصولها اللغوية الأكثر شيوعاً في نماذج التحليل؛
- **تجريد النصوص من الزخارف :** شملت العملية حذف علامات التشكيل وحروف الإطالة التي تُستخدم للتأكيد البصري دون أن تحمل قيمة دلالية إضافية في سياق تحليل المشاعر؛
- **معالجة الإطالة الحرفية:** عولجت ظاهرة التمديد الحرفي التعبيري الشائعة في الكتابة الرقمية من خلال تقليص الحروف المتكررة المتتالية إلى حرفين كحد اقصى، مما يُعيد الكلمات إلى صيغتها المعجمية المبسطة. ويوضح المثال التالي هذه العملية: قبل المعالجة *"هاايله"* ، *"عياااااااان"* بعد المعالجة *"هايله"* ، *"عيان"*.

```python
def normalization_dz(text):
    if pd.isna(text): return ""
    text = str(text)
    text = re.sub(r'[\u0617-\u061A\u064B-\u0652]', '', text)  # إزالة التشكيل
    text = re.sub(r'[إآا]', 'ا', text)   # توحيد الألفات
    text = re.sub(r'ه' ,'ة', text)        # التاء المربوطة
    text = re.sub(r'ي' ,'ى', text)        # الألف المقصورة
    text = re.sub(r'ق' ,'ڤ', text)        # حروف الدارجة
    text = re.sub(r'ف' ,'ڤ', text)
    text = re.sub(r'-+', '', text)        # حذف التطويل
    text = re.sub(r'(.)\1+', r'\1\1', text) # تقليص التكرار
    return text.strip()
```

الصورة 4.3: خوارزم عملية تطبيع النصوص

5.3.3 المرحلة الخامسة: معالجة التكرار النصي

صُمِّمت دالة مخصصة (handle_redundancy) لمعالجة ظاهرة التكرار النصي التي تشيع في تعليقات وسائل التواصل الاجتماعي، وتعمل هذه الدالة على ثلاثة مستويات متكاملة:

- تكرار علامات الترقيم: تُقلَّص علامات الترقيم المتكررة المتتالية (كـ ؟؟؟ أو !!!) إلى رمز واحد، تفادياً لتضخيم وزن هذه الرموز في التمثيل العددي اللاحق؛
- تكرار العبارات: يُرصد التكرار على مستوى العبارات المتتالية التي تتراوح بين كلمة وأربع كلمات، وتُحذف النسخ الزائدة مع الإبقاء على نسخة واحدة فقط. ويوضح المثال التالي هذه العملية: قبل المعالجة *"روعه حبيت قلب احمر قلب احمر"* بعد المعالجة *"روعه حبيت قلب احمر"*؛
- تقييس المسافات : تُدمج الفراغات المتعددة وتُزال المسافات الطرفية لضمان اتساق الشكل النهائي للنص.

6.3.3 المرحلة السادسة: معالجة النفي

نظراً لتعدد صيغ النفي في الدارجة الجزائرية وأهميتها البالغة في تحديد التوجه الشعوري للنص، إذ كثيراً ما يُحيل النفي المعنى من الإيجابية إلى السلبية أو العكس، صُمِّمت دالة مخصصة (negation_handling) تعالج هذه الظاهرة وفق أربعة إجراءات:

- توحيد النفي المنفصل: يُربط أداة النفي "ما" بالكلمة التالية لها مباشرة لمعالجة حالات الكتابة المنفصلة مثل *"ما عندو" ⟵ "ماعندو"*؛
- توحيد النفي المزدوج: تُدمج صيغة النفي (ما + فعل + ش) في وحدة نصية واحدة مما يُغني عن بناء قوائم يدوية لكل فعل منفي مثل :*"ما فهمت ش" ⟵ "مافهمتش"*؛
- معالجة حالات النفي الشائعة : تُوحَّد التعبيرات النافية المتكررة كـ "ما كاش" و"ما كانش" في صيغة موحدة ثابتة؛
- ربط أدوات النفي بالكلمة التالية: تُوصَّل أدوات النفي الجزائرية الشائعة (كـ "مش"، "موش"، "مانيش"، "ماشي") بالكلمة التالية لها بشرطة سفلية لضمان تماسك الوحدة الدلالية للنفي أثناء التحليل: *"مش ملاح" ⟵ "مش_ملاح"*.

ويوضح الكود البرمجي التالي آلية تنفيذ هذه العملية :

```
def negationhandling(text):
    text = str(text)
    text = re.sub(r'\bما\s+(\w+)\b', r'1\ما', text)
    text = re.sub(r'\bما\s+(\S+?)\s*ش\b', r'ما1\ش', text)
    text = re.sub(r'\bما\s+(كاش|كاش|كانش|كانت)\b', r'1\ما', text)

    prefixes = ['مارانيش' ,'ماهيش' ,'مكيش' ,'مانيش' ,'بلا' ,'موش' ,'مش' ,'مشي' ,'ماشي']

    for pref in prefixes:
        # الكلمة + فراغ + البحث عن الأداة
        text   re.sub(rf'\b{pref}\s+(\S+)', rf'{pref}_\1', text)

    return text
```

الصورة 5.3: خوارزم عملية معالجة النفي

7.3.3 المرحلة السابعة: التقطيع

اعتُمد في هذه المرحلة أسلوب التقطيع القائم على الفراغات (Whitespace-based Tokenization)، إذ يُجزَّأ النص إلى قائمة من الوحدات اللغوية (Tokens) بالاعتماد على المسافات الفاصلة بين الكلمات. وقد جاء هذا الاختيار مبرراً بطبيعة البيانات المعالجة في المراحل السابقة، التي ضمنت توحيد المسافات وعزل كل وحدة نصية بشكل صحيح، مما

جعل التقطيع بالفراغات كافياً وفعالاً دون الحاجة إلى أدوات تقطيع متخصصة قد لا تتلاءم مع خصائص الدارجة الجزائرية.

8.3.3 المرحلة الثامنة : إزالة كلمات الوقف (الحشو)

اعتُمدت في هذه المرحلة قائمة كلمات وقف مخصصة (StopwordsList) أُعِدَّت يدوياً وتضم 3,136 كلمة، صُمِّمت لتتلاءم مع خصائص الدارجة الجزائرية واللغة العربية معاً، بدلاً من الاعتماد على القوائم الجاهزة التي تغلب عليها مفردات الفصحى وقد لا تعكس طبيعة البيانات المستخدمة. وقد خُزِّنت هذه القائمة في هيكل بيانات من نوع (Set) لضمان سرعة البحث والمطابقة أثناء المعالجة. ثم طُبِّقت دالة إزالة على قائمة الوحدات اللغوية الناتجة عن مرحلة التقطيع، بحيث تُحتفظ فقط بالكلمات غير الموجودة في قائمة الوقف، مما يُركّز التحليل على المفردات ذات القيمة الدلالية الفعلية ويُقلّص حجم المدخلات اللاحقة للنماذج.

4.3 إعداد بيانات التدريب

بعد استكمال مراحل المعالجة المسبقة وتهيئة النصوص بالشكل المناسب، انتقلت الدراسة إلى مرحلة إعداد البيانات للتدريب، وهي المرحلة التي تُحوَّل فيها البيانات النصية المنظفة إلى مدخلات قابلة للاستخدام المباشر في تدريب النماذج واختبارها. وقد اشتملت هذه المرحلة على ثلاثة محاور متكاملة: تشخيص مشكلة عدم توازن الفئات ومعالجتها، وتقسيم البيانات بين التدريب والاختبار، وتحويل النصوص إلى تمثيلات عددية ملائمة لكل فئة من النماذج المستهدفة.

1.4.3 توزيع الفئات وتشخيص مشكلة عدم التوازن

كشف تحليل توزيع الفئات في مجموعة البيانات المعنوَنة عن تفاوت واضح بين الفئات الثلاث كما هو موضح في *1.4.2.3* ، إذ تهيمن الفئة المحايدة بنسبة تتجاوز الثلثين مقارنةً بالفئتين الإيجابية والسلبية .

وتُعدّ هذه الظاهرة المعروفة بـ Class Imbalance من أبرز التحديات في مهام تصنيف النصوص، إذ تدفع النماذج نحو التحيز للفئة الأكثر تمثيلاً على حساب الفئات الأقل حضوراً. وفي سياق هذه الدراسة تحديداً، قد يدفع هذا التحيز إلى نماذج تُصنّف غالبية التعليقات كمحايدة بصرف النظر عن محتواها الشعوري الفعلي، مما يُفقدها قدرتها التمييزية بين الإيجابي والسلبي وهو الهدف الجوهري لتحليل المشاعر.

2.4.3 معالجة عدم توازن الفئات

اعتُمدت في هذه الدراسة تقنيتان مختلفتان لمعالجة عدم التوازن، وُزِّعت هاتان التقنيتان على النماذج وفق طبيعة كل نموذج و إمكانيات دمجها داخل بنية التدريب الخاصة به

1.2.4.3 تقنية SMOTE

وُظِّفت تقنية SMOTE (Synthetic Minority Over-sampling Technique) مع نموذج SVM المقترن بتمثيل TF-IDF. و تقوم هذه التقنية على توليد أمثلة اصطناعية للفئات الأقل تمثيلاً عبر الاستيفاء بين الأمثلة الموجودة في الفضاء المتجهي، بدلاً من الاكتفاء بتكرار الأمثلة الحالية. وقد جاء اختيار SMOTE لهذا النموذج مبرراً بطبيعة تمثيل TF-IDF الذي يُنتج متجهات متفرقة (Sparse Vectors) ذات أبعاد محددة تجعل الاستيفاء بين الأمثلة ممكناً وذا معنى.

وتجدر الإشارة إلى أن SMOTE طُبِّقت حصراً على بيانات التدريب دون بيانات الاختبار، وذلك لضمان موضوعية التقييم وتجنب أي تسرب للمعلومات (Data Leakage) قد يُشوّه النتائج.

2.2.4.3 تقنية Class Weights

تعتمد تقنية Class Weights على إعادة موازنة تأثير كل فئة أثناء التدريب عبر تعيين أوزان عكسية لتكرار كل فئة في دالة الخسارة، إذ تحظى الفئات الأقل تمثيلاً بأوزان أعلى مما يجعل أخطاء تصنيفها أكثر تأثيراً في تحديث معاملات النموذج. و بالتالي يساعد على تحسين أدائه تجاه جميع الفئات بشكل متوازن.

وتتميز هذه الطريقة بكونها لا تتطلب تعديل البيانات نفسها، بل تُدمج مباشرة داخل مرحلة التدريب.

3.4.3 تقسيم البيانات

جرى تقسيم مجموعة البيانات المعالجة وفق نسبة للتدريب 80% و 20% للاختبار، و ذلك باستخدام أسلوب التقسيم العشوائي المتقطع (Stratified Split) الذي يضمن الحفاظ على نفس نسب توزيع الفئات في كلا المجموعتين، وهو اعتبار بالغ الأهمية في حالة عدم توازن الفئات التي تتسم بها هذه البيانات.

المجموعة	النسبة	العدد	الغرض
التدريب	80%	6614	تعلم النموذج وضبط معاملاته
الاختبار	20%	1653	التقييم النهائي الموضوعي

جدول 2.3: تقسيم البيانات المعالجة

وقد حُرص على إجراء التقسيم قبل تطبيق أي معالجة لعدم التوازن، بحيث تقتصر هذه المعالجة على بيانات التدريب وحدها دون المساس ببيانات الاختبار، مما يضمن أن يعكس التقييم النهائي الأداء الحقيقي للنماذج على بيانات لم تتعرض لها من قبل.

5.3 تدريب النماذج وتقييم الأداء

يستعرض هذا الجزء المراحل التطبيقية لتدريب النماذج الثلاث المعتمدة في هذه الدراسة وتقييم أدائها، وذلك في إطار منهجية مقارنة شاملة تهدف إلى تحديد النموذج الأنسب لتحليل مشاعر التعليقات الجزائرية الدارجة. وقد اعتُمدت بيئة **Google Colab** منصةً للتطوير والتدريب، نظراً لما توفره من موارد حوسبة مجانية تشمل وحدات معالجة الرسوميات (GPU) الضرورية لتدريب نماذج التعلم العميق والمحولات.

1.5.3 البيئة التقنية والمكتبات المستخدمة

اعتُمد في تنفيذ هذه الدراسة مجموعة من المكتبات البرمجية المتخصصة الموضحة في الجدول التالي:

المكتبة	الاستخدام
scikit-learn	SMOTE.Pipeline,TF-IDF,SVM
TensorFlow /Keras	بناء وتدريب CNN
Gensim	تدريب نموذج FastText
HuggingFace Transformers	Fine-Tuning لنموذج marbert
PyTorch	دعم نماذج المحولات

المكتبة	الاستخدام
Pandas/NumPy	معالجة البيانات وتحليلها
Matplotlib/Seaborn	التمثيل البياني للنماذج

جدول 3.3:- الأدوات والمكتبات البرمجية المستخدمة

2.5.3 النماذج الكلاسيكية

1.2.5.3 نموذج SVM مع TF-IDF

التمثيل (TF-IDF Representation)

يُعتمد تمثيل TF-IDF في هذا العمل كمدخل لنموذج SVM، حيث يتم تحويل النصوص إلى متجهات عددية تعكس أهمية كل كلمة داخل الوثيقة مقارنةً ببقية وثائق مجموعة البيانات. ويتميز هذا التمثيل بكونه مناسبًا لنموذج SVM، نظرًا لاعتمادها على المتجهات العددية وقدرتها العالية على معالجة البيانات المتفرقة (Sparse Representations) بكفاءة.

يتم حساب هذا الوزن عبر عدة خطوات، تبدأ بحساب تردد المصطلح (TF) باستعمال التحجيم اللوغاريتمي للحد من تأثير التكرار المفرط وفق المعادلة (03) :

$$tf_{i,j} = 1 + log(f_{i,j}) \quad .....03$$

حيث: i يمثل الكلمة

j يمثل الوثيقة (النص)

ثم يتم حساب تردد الوثيقة العكسي idf مع التنعيم لتفادي القيم الصفرية وضمان استقرار الأوزان وفق المعادلة (04):

$$idf(t) = 1 + log\left(\frac{1+N}{1+df(t)}\right) \quad .....04$$

حيث: N يمثل عدد الوثائق الكلي

$df(t)$ عدد الوثائق التي تحتوي على المصطلح t.

بعد ذلك يتم حساب الوزن النهائي لكل كلمة وفق العلاقة (05) :

$$w_{i,j} = tf_{i,j} \times idf(t) \quad \text{.....05}$$

ثم يتم تطبيق عملية التقييس (L2 Normalization) لضمان أن طول كل متجه يساوي 1، مما يزيل تأثير طول النص ويجعل المقارنة بين الوثائق أكثر عدلاً.

وقد ضُبِّطت المعاملات على النحو التالي:

المعامل	القيمة	المبرر
max_features	10,000	لاستيعاب مفردات الدارجة المتنوعة
ngram_range	(1,3)	لالتقاط السياق الثلاثي لفهم النفي والتوكيد
sublinear_tf	True	تخفيف تأثير الكلمات المتكررة
min_df	3	تجاهل الأخطاء الإملائية النادرة

جدول 4.3: إعدادات TF-IDF المعتمدة مع نموذج SVM

كما توضح الصورة 6.3 أمثلة عن المفردات المستخرجة باستخدام TF-IDF، حيث يتم التعامل مع الأشكال الاشتقاقية المختلفة للكلمة الواحدة، مثل "بشرتك" و"بشرتها" و"بشرتي"، كمصطلحات مستقلة تمامًا. ويعكس ذلك أحد حدود هذا التمثيل، والمتمثل في عدم قدرته على التقاط التشابه الصرفي أو الدلالي بين الكلمات ذات الجذر الواحد، مما يؤدي إلى فقدان جزء من المعلومات اللغوية ذات الصلة.

```
print(vocab[1000:1030])
```

```
['بزيت الجوجوبا' 'بزيوت' 'بس' 'بسبب' 'بسرعه' 'بسعر' 'بسومه' 'بسيروم'
 'بشامبو' 'بشامبوان' 'بشحال' 'بشر' 'بشرتك' 'بشرتك جافه' 'بشرتك حساسه'
 'بشرتك دهنيه' 'بشرتك مختلطه' 'بشرتنا' 'بشرتها' 'بشرتها جافه'
 'بشرتها دهنيه' 'بشرتها مختلطه' 'بشرتهم' 'بشرتي' 'بشرتي جافه'
 'بشرتي جافه' 'بشرتي حساسه' 'بشرتي دهنيه' 'بشرتي ذهنيه' 'بشرتي زيتيه']
```

صورة 6.3: عينة من القاموس الناتج عن TF-IDF

المعالجة (Balancing: SMOTE / class_weight)

تتضمن مرحلة المعالجة في نموذج SVM مجموعة من الخطوات الهادفة إلى تحسين جودة التعلم ومعالجة عدم توازن البيانات. حيث تم استخدام معامل **class_weight=balanced** داخل نموذج LinearSVC من أجل إعادة موازنة تأثير الفئات المختلفة أثناء التدريب، وذلك عبر إعطاء وزن أكبر للفئات الأقل تمثيلاً، مما يمنع انحياز النموذج نحو الفئات الأكثر تكراراً ويضمن تعلماً أكثر عدلاً بين جميع الفئات.

بالإضافة إلى ذلك، تم تطبيق تقنية **SMOTE (Synthetic Minority Over-sampling Technique)** لتوليد عينات اصطناعية للفئات الأقل تمثيلاً في فضاء الخصائص، بهدف تعزيز توازن مجموعة البيانات وتحسين قدرة النموذج على تعلم خصائص الفئات الضعيفة التمثيل بشكل أفضل.

وبذلك، تساهم هذه المرحلة في رفع قدرة النموذج على التعميم وتقليل مشكلة الانحياز الناتج عن عدم توازن البيانات، مما يؤدي إلى أداء أكثر استقراراً وموثوقية في عملية التصنيف.

ج- التصنيف (LinearSVC Hyperplane)

يتلقى النموذج المتجهات العددية ويبحث عن أفضل مستوٍ فاصل (Hyperplane) يُميّز بين الفئات الثلاث (إيجابي، سلبي، محايد) في الفضاء عالي الأبعاد. والممثل بالمعادلة (06):

$$w \cdot x + b = 0 \quad \text{.....06}$$

ويتم ذلك عبر حل مسألة تحسين تهدف إلى تقليل دالة الهدف التالية(07):

$$min_{w,b} \frac{1}{2}||w||^2 + C \sum_{i=0}^{n} \xi_i \quad .....07$$

حيث : w الأوزان (تحدد شكل الخط الفاصل)؛

b الانحياز (يحرك الخط)؛

$||w||^2$ تعقيد النموذج (كلما صغر كان أفضل)؛

C يوازن بين الدقة وبساطة النموذج؛

ξ_i أخطاء التصنيف لكل نقطة؛

$\sum_{i=0}^{n} \xi_i$ مجموع كل الأخطاء.

حيث يتم تقليل تعقيد النموذج $||w||^2$ الذي يساهم في تعظيم الهامش (Margin)، مما يسمح بفصل أفضل بين الفئات. وفي المقابل، يقوم معامل التنظيم C بالتحكم في مدى التساهل مع الأخطاء، بحيث يوازن بين تقليل عدد التصنيفات الخاطئة والحفاظ على نموذج عام قادر على التعميم.

النتائج

تم تقييم النموذج باستخدام مقاييس Accuracy وF1-Macro إلى جانب F1-score لكل فئة على حدة.

النموذج	المعالجة	Accuracy	F1 Macro	F1 (N)	F1 (NEU)	F1 (P)
SVM + TF-IDF	without SMOTE	0.77	0.68	0.58	0.84	0.62
SVM + TF-IDF	SMOTE	0.69	0.62	0.48	0.77	0.61

جدول 5.3: نتائج تقييم نماذج التعلم الآلي التقليدية

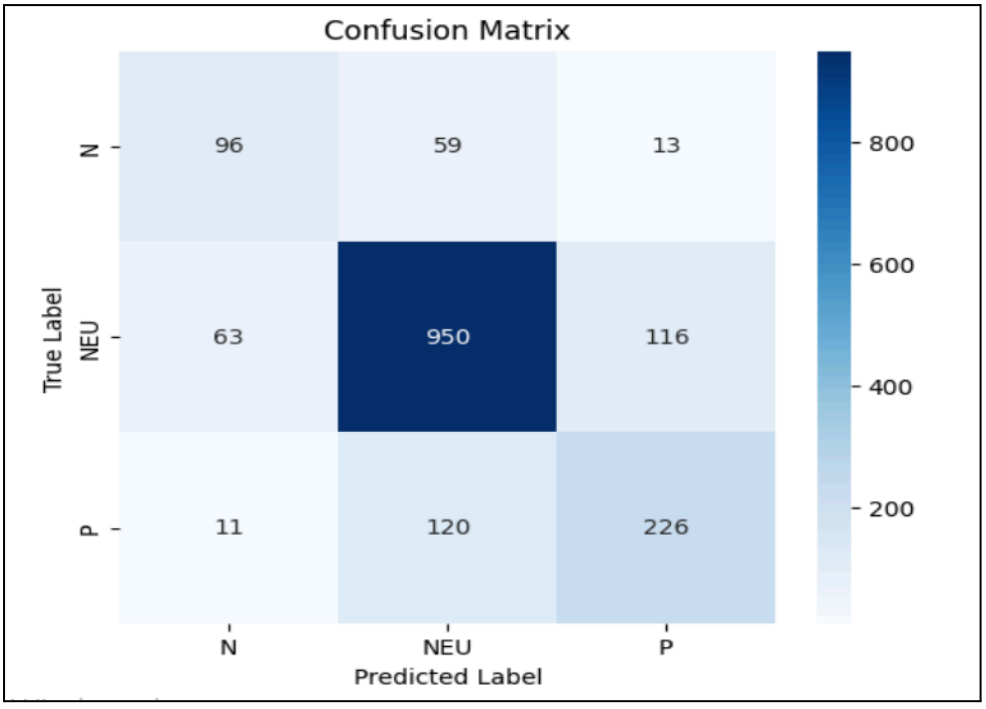
التفسير والتحليل

حقّق النموذج بدون SMOTE دقةً كليةً بلغت 77% وF1-Macro قدره 0.68، مما يؤكد فعالية نموذج SVM مع تمثيل TF-IDF في تصنيف التعليقات المكتوبة بالدارجة الجزائرية. ويمكن تفسير هذه النتائج على ثلاثة مستويات:

- **على مستوى الفئات:** سجّلت الفئة المحايدة أعلى F1 بـ 0.84، وهو ما يعكس هيمنتها العددية في مجموعة البيانات، إذ يميل النموذج طبيعياً نحو الفئة الأكثر تمثيلاً. في المقابل، جاءت الفئة السلبية بأضعف أداء عند **0.58**، مما يدل على صعوبة التمييز بين التعبيرات السلبية والمحايدة في الدارجة الجزائرية، ولا سيما في ظل غياب التوحيد الإملائي وتعدد أشكال التعبير عن الرفض؛
- **على مستوى SMOTE:** أدى إضافة SMOTE إلى انخفاض ملحوظ في الدقة من 77% إلى 69%، ويُعزى ذلك إلى أن توليد العينات الاصطناعية يتم في فضاء TF-IDF المتناثر (Sparse Space)، حيث لا تعكس العينات المولّدة بالاستيفاء الخطي أنماطاً لغوية حقيقية، مما يُدخل ضوضاءً تؤثر سلباً على التدريب؛
- **على مستوى الخيار النهائي:** بناءً على هذه النتائج، اعتُمدت النسخة **بدون SMOTE** كنموذج مرجعي نهائي لـ SVM، كونها تحقق توازناً أفضل بين الدقة الكلية وأداء الفئات الثلاث.

مصفوفة الارتباك

تكشف مصفوفة الارتباك الموضحة عن أنماط الأخطاء التي يرتكبها النموذج بدقة أكبر مما توفره مقاييس الأداء الكلية وحدها.



صورة 7.3: مصفوفة الارتباك لنموذج ال SVM

الملاحظات الرئيسية:

- **الفئة المحايدة (NEU)** حقّقت أفضل أداء بـ 950 تصنيفاً صحيحاً من أصل 1129، غير أن النموذج صنّف 63 تعليقاً محايداً خطأً على أنها سلبية، و116 على أنها إيجابية، مما يعكس التداخل الدلالي الطبيعي بين الفئات في الدارجة الجزائرية؛
- **الفئة السلبية (N)** هي الأكثر إشكالية، إذ صُنِّف 59 تعليقاً سلبياً على أنها محايدة وهو أكبر مصدر للخطأ في النموذج. ويُفسَّر ذلك بأن التعبير عن الرفض والنقد في الدارجة كثيراً ما يأتي بصيغ خفية أو غير مباشرة لا يلتقطها TF-IDF الذي يعتمد على التكرارات الإحصائية للكلمات بمعزل عن سياقها؛
- **الفئة الإيجابية (P)** سجّلت 226 تصنيفاً صحيحاً من أصل 357، لكن 120 تعليقاً إيجابياً صُنِّفت خطأً كمحايدة. ويعود ذلك جزئياً إلى أن كثيراً من التعليقات الإيجابية في الدارجة تُصاغ بعبارات قصيرة أو مقتضبة تفتقر إلى الكلمات الدالة الصريحة التي يعتمد عليها TF-IDF.

الخلاصة

يتضح أن النموذج يُجيد التمييز حين تكون الفئة واضحة المفردات، لكنه يُخفق حين يتطلب الأمر فهم السياق الكامل للجملة هو أحد القيود الأساسية للنماذج الإحصائية التقليدية .

3.5.3 نماذج التعلم العميق (Word Embeddings)

1.3.5.3 نموذج CNN مع FastText

التمثيل (FastText Embeddings)

اعتمدت هذه المقاربة على نموذج FastText لتعلم التمثيلات الشعاعية للكلمات انطلاقًا من مجموعة التعليقات المدروسة. وقد تم تدريب النموذج وفق معمارية Skip-gram باستعمال بعد تضمين مقداره d=100 ونافذة سياقية حجمها 5 كلمات.

يعتمد FastText على الوحدات الحرفية الفرعية (Subword n-grams)، حيث تُفكك كل كلمة w إلى مجموعة من المقاطع الحرفية نرمز لها بـ:

$$G\ (w)\ =\ \{g1, g2, \ldots, gn\}$$

يحتوي النموذج على مصفوفتين قابلتين للتعلم:

$$W\in R^{|G|\times d}$$

$$W'\in R^{|G|\times d}$$

حيث تمثل Wمصفوفة الإدخال التي تخزن متجهات الوحدات الحرفية الفرعية، بينما تمثلW' مصفوفة السياق التي تحتوي متجهات الكلمات المستعملة أثناء التنبؤ بالسياق.

يُنشأ تمثيل الكلمة المركزية من خلال جمع متجهات جميع الوحدات الحرفية المكوِّنة لها(08) :

$$h\ =\ \sum_{g\in G(w)} z_g \quad \ldots\ldots\ldots 08$$

حيث تمثل z_g متجه الوحدة الحرفية g ، ويمثل h المتجه الخفي Hidden Vector النهائي للكلمة المركزية.

بعد ذلك يُستعمل هذا المتجه للتنبؤ بالكلمات الموجودة داخل النافذة السياقية. فإذا كانت c كلمة سياق حقيقية، يُحسب الجداء السلمي بين المتجه الخفي و متجه كلمة السياق في مصفوفة السياق(09):

$$s_c\ =\ v'_c{}^{\top} h \quad \ldots\ldots\ldots 09$$

ثم تُحوَّل هذه القيمة إلى احتمال باستعمال دالة سيغمويد(10):

$$\sigma(s_c)\ =\ \frac{1}{1+e^{-s_c}} \quad \ldots\ldots\ldots 10$$

ولتقليل التكلفة الحسابية الناتجة عن التعامل مع جميع كلمات القاموس، يعتمد FastText على تقنية العينات السالبة (Negative Sampling)، حيث تُعامل الكلمات السياقية الحقيقية كعينات موجبة، وتُختار مجموعة من الكلمات العشوائية كعينات سالبة. وعليه تصبح دالة الخسارة(11)(Loss Function) :

$$L = log\ \sigma({v'}_c^{\top} h) + \sum_{i=1}^{k} log\ \sigma({v'}_{n_i}^{\top} h) \quad .....11$$

حيث تمثل n_i الكلمات السالبة المختارة عشوائياً، ويمثل k عدد العينات السالبة.

أثناء عملية التدريب تُحدَّث أوزان مصفوفة السياق W' وفق Gradient Descent:

$${v'}_{w_j} \leftarrow {v'}_{w_j} - \eta(\sigma({v'}_{w_j}^{\top} h) - t_j)\, h \quad .....12$$

ثم يُحسب متجه الخطأ المتراكم e الناتج عن جميع العينات، ويُوزَّع على متجهات الوحدات الحرفية في مصفوفة الإدخال W:

$$z_g \leftarrow z_g - \frac{\eta}{|G(w)|} e \quad \forall\, g \in G(w) \quad .....13$$

ومع تكرار هذه العملية على كامل بيانات التدريب، تتعلم المتجهات العلاقات الدلالية والسياقية بين الكلمات.

بعد اكتمال التدريب، يُعاد بناء تمثيل كل كلمة انطلاقًا من الوحدات الحرفية المكونة لها وفق نفس العلاقة (14):

$$e_w = \sum_{g \in G(w)} z_g \quad .....14$$

وتُجمع هذه التمثيلات لبناء مصفوفة التضمين:

$$E = [e_{w_1}\ e_{w_2} \ldots\ldots\ldots\ldots\ e_{w_v}]^{\top} \in R^{V \times 100} \quad .....15$$

وأخيراً تُحمَّل مصفوفة التضمين E كأوزان ابتدائية لطبقة التضمين داخل شبكة CNN، بحيث يُحوَّل كل تعليق نصي إلى سلسلة من المتجهات العددية قبل تمريرها إلى طبقات الالتفاف لاستخراج الميزات اللازمة لعملية التصنيف.

توضح الصورة 8.3 نموذجاً من المتجهات العددية الكثيفة التي يُنتجها FastText، إذ تُمثَّل كل كلمة بمصفوفة من القيم الحقيقية عبر أبعاد متعددة تعكس موقعها في الفضاء الدلالي .

	Word	Dim_1	Dim_2	Dim_3	Dim_4	Dim_5	Dim_6	Dim_7	Dim_8	Dim_9	...
0	قلب	-0.670655	1.015498	-0.011661	0.496401	-0.108822	-0.218997	0.159867	0.294008	0.106520	...
1	وجه	-1.703619	2.580137	-0.060250	1.245625	-0.278565	-0.574453	0.397617	0.739463	0.278681	...
2	احمر	-0.629135	0.966305	-0.023370	0.461313	-0.100981	-0.214737	0.150968	0.279400	0.103387	...
3	ولا	-1.200160	1.844447	-0.037490	0.881306	-0.203124	-0.407939	0.280043	0.532196	0.195446	...
4	وجهي	-1.232196	1.876149	-0.042830	0.902875	-0.201484	-0.416168	0.284218	0.535086	0.202106	...

صورة 8.3: متجهات تضمين الكلمات بواسطة FastText

وفي الصورة 9.3، يُظهر النموذج قدرته على إيجاد الكلمات الأقرب دلالياً للفظة "وجوههكم"، وهي كلمة غير موجودة بشكل مباشر ضمن بيانات التدريب. و يعكس ذلك قدرة FastText على معالجة مثل هذه الحالات بفضل اعتماده على التمثيل المعتمد على الأجزاء الحرفية (subword units)، بخلاف نماذج التي تعتمد على تمثيل سطحي للكلمات و تواجه صعوبة مع الكلمات غير الموجودة في معجم التدريب.

```
نجح! لقّى جيران للكلمة 'وجوههكم' رغم أنها غريبة FastText:
[('0.9998847246170044', 'وجهكم'), ('0.999877393245697', 'وجهي'), ('0.9998710751533508', 'بالوجه'),

('0.9998703598976135', 'والوجه'), ('0.9998695254325867', 'وجوهكم'),
```

صورة 9.3: نتائج التشابه الدلالي باستخدام FastText

مرحلة التصنيف CNN

تُطبق معمارية CNN على التعليقات النصية بعد تحويلها إلى تمثيلات عددية، حيث يتم تمرير كل تعليق على شكل تسلسل من الكلمات بطول L=100 إلى طبقة Embedding المبنية على FastText، مما ينتج مصفوفة تمثيل لكل تعليق.

بعد ذلك، يتم إدخال هذه التمثيلات مباشرة إلى بقية طبقات النموذج بشكل متتابع. تعمل طبقة SpatialDropout1D على تعطيل بعض قنوات التضمين بشكل عشوائي أثناء التدريب، مما يساعد النموذج على التعميم على بيانات جديدة وعدم الاعتماد على كلمات محددة.

ثم تقوم طبقة Conv1D بتمرير مرشحات تلافيفية عبر كامل التسلسل النصي لاستخراج الأنماط المحلية المرتبطة بسياق الكلمات داخل التعليقات. بعد ذلك يتم تلخيص هذه المخرجات عبر طبقة GlobalMaxPooling1D لاختيار أهم الإشارات الدلالية في كل تعليق بغض النظر عن موقعها.

تُمرَّر بعدها التمثيلات المختزلة إلى طبقة Dense مع Dropout من أجل دمج السمات المستخرجة واتخاذ القرار النهائي. في المرحلة الأخيرة، تقوم طبقة Softmax بإنتاج احتمالات الانتماء إلى كل فئة، ويتم تصنيف كل تعليق حسب الفئة ذات الاحتمال الأعلى.

يوضح الجدول 6.3 معمارية النموذج بشكل مفصل.

الطبقة	الدور
Embedding	تحويل الكلمات إلى متجهات FastText
SpatialDropout1D(0.3)	حماية من الإفراط في التعلم على مستوى الميزات
Conv1D(256, 5)	استخلاص الأنماط المحلية من نوافذ سياقية بطول 5 كلمات
BatchNormalization	تثبيت التدريب وتسريع التقارب
GlobalMaxPooling1D	الاحتفاظ بأبرز ميزة من كل مرشح
Dense(128) + Dropout(0.5)	التصنيف مع الحد من الإفراط في التعلم
Dense(3, softmax)	إخراج احتمالات الفئات الثلاث

جدول 6.3: معمارية نموذج CNN المعتمد في الدراسة

كما استُخدم EarlyStopping وReduceLROnPlateau لضمان أفضل نقطة تقارب وتجنب الإفراط في التدريب، مع تفعيل أوزان الفئات (class_weight) لمعالجة التفاوت في توزيع البيانات.

النتائج

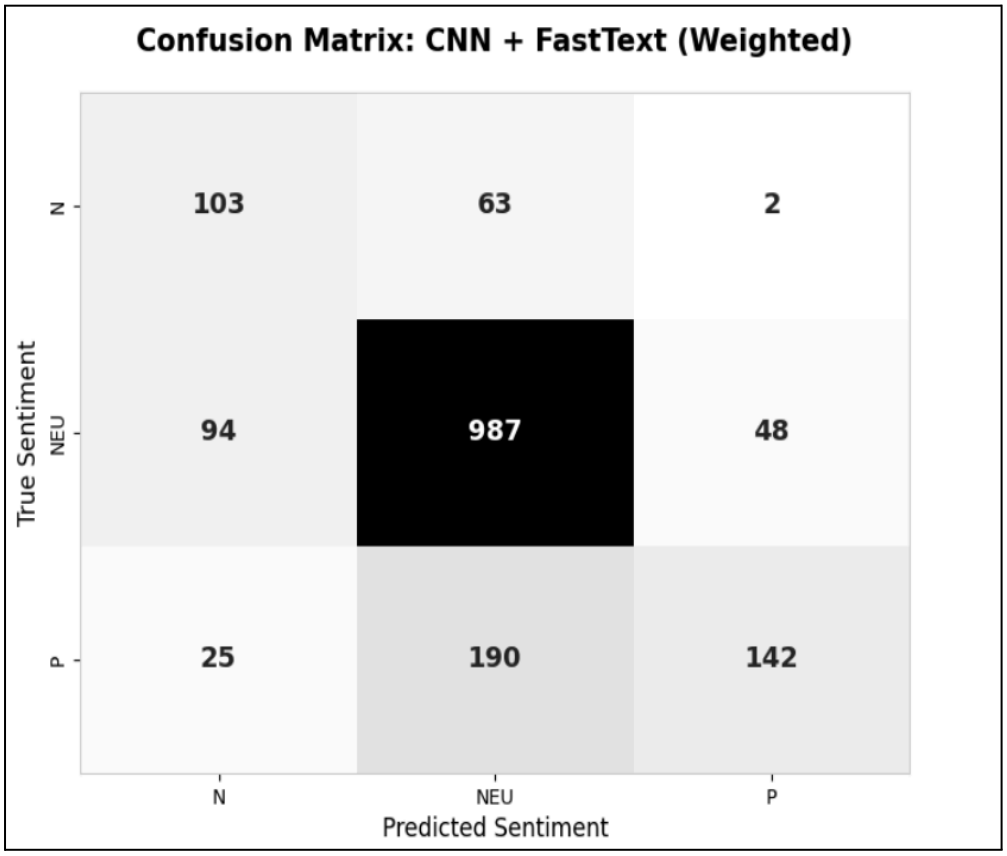
تم تقييم النموذج باستخدام مقاييس Accuracy وF1-Macro إلى جانب F1-score لكل فئة على حدة.

النموذج	المعالجة	Accuracy	F1 Macro	F1 (N)	F1 (NEU)	F1 (P)
CNN + FastText	Class Weights	0.74	0.63	0.53	0.83	0.52

جدول 7.3 : نتائج تقييم نماذج التعلم العميق

تحليل مصفوفة الارتباك والتفسير

تكشف مصفوفة الارتباك التالية عن ثلاث ملاحظات جوهرية:



صورة 10.3 : مصفوفة الارتباك لنموذج ال CNN

- **الفئة المحايدة (NEU)** حقّقت **987** تصنيفاً صحيحاً من أصل 1129، وهو أداء متين يعكس قدرة CNN على التقاط الأنماط المحلية للتعليقات المحايدة التي تغلب عليها الصياغة الوصفية المباشرة؛

- **الفئة السلبية (N)** سجّلت 103 تصنيفاً صحيحاً من أصل 168، غير أن 63 تعليقاً سلبياً صُنِّفت خطأً كمحايدة وهو نفس نمط الخطأ الملاحَظ في SVM، مما يدل على أن هذا الالتباس يعكس خصوصية لغوية حقيقية في الدارجة لا قصوراً في النموذج وحده. والملفت هنا أن CNN صنّف فقط تعليقَيْن سلبيَّيْن على أنهما إيجابيان، وهو تحسّن ملحوظ مقارنة بـ SVM كان ب 13 تعليقاً؛
- **الفئة الإيجابية (P)** هي الأكثر إشكالية في هذا النموذج، إذ صُنِّف **190** تعليقاً إيجابياً على أنها محايدة من أصل 357، مما أفضى إلى أضعف F1 في هذه الفئة بـ **0.52**. ويُفسَّر ذلك بأن CNN يستخلص أنماطاً محلية من نوافذ محدودة (5 كلمات)، وهو ما لا يكفي لفهم المشاعر الإيجابية التي كثيراً ما تُعبَّر عنها في الدارجة بسياقات أطول أو بتعابير ضمنية تتجاوز النافذة المحلية.

الخلاصة

مقارنةً بـ SVM حقّق CNN + FastText دقةً أقل بـ 3 نقاط (74% مقابل 77%)، ويُعزى ذلك بالدرجة الأولى إلى محدودية حجم بيانات التدريب التي لا تتيح لنماذج التعلم العميق تحقيق كامل إمكاناتها ، إذ تحتاج هذه النماذج عادةً إلى كميات أكبر من البيانات لتفوق على النماذج الكلاسيكية.

4.5.3 نماذج المحولات (Transformers)

1.4.5.3 نموذج MARBERT (Fine-Tuning)

تمثيل الكلمات

يتكون النموذج من ثلاث طبقات رئيسية تعمل بشكل متكامل لمعالجة النصوص وفهمها. تتمثل الطبقة الأولى في **Embedding Layer**، التي تقوم بتحويل الكلمات أو الرموز (Tokens) إلى تمثيلات عددية كثيفة يمكن للنموذج معالجتها. تليها **Encoder Layer**، وهي الجزء الأساسي المسؤول عن تعلم العلاقات السياقية بين الكلمات من خلال آليات الانتباه والشبكات العصبية العميقة. أما الطبقة الأخيرة فهي **Classification Layer**، التي تعتمد على التمثيلات المستخرجة من المشفّر لإنتاج التنبؤ النهائي وتحديد الفئة المناسبة للنص. يمكن تلخيص **مراحل عمل النموذج** في الخطوات التالية:

أولاً: الترميز والتقطيع (Tokenization & Padding):

يبدأ النموذج بتحويل النص الخام إلى وحدات لغوية (Tokens) باستخدام المُجزِّئ المخصص لـ MARBERT. وفق إعداد max_length=256، يُحدَّد طول كل جملة بـ 256 موضعاً، بحيث يوضع رمز [CLS] اول الجملة السطر الاول و الرمز [SEP] اخر الجملة و تُحشى المواقع المتبقية برمز الحشو [PAD]، وتُقطَّع الجمل الأطول عند الحد الأقصى.

ثانياً : استرجاع التمثيلات (Embeddings Lookup):

يمتلك النموذج ثلاث مكونات أساسية لتمثيل المدخلات: تضمين المفردات (Token Embeddings)، وتضمين المواقع (Position Embeddings)، وتضمين التجزئة (Segment Embeddings). تُستخلص تمثيلات الـ 256 موضعًا الأولى من مصفوفة المواقع الأصلية ذات الأبعاد (512×768)، ثم تُجمع مع تمثيلات المفردات وتمثيلات التجزئة لتكوين التمثيل النهائي للمدخلات وفق المعادلة:

Input Representation = Token Embedding + Position Embedding + Segment Embedding 16

وينتج عن ذلك مصفوفة إدخال نهائية:

$$X \in R^{256\times768}$$

حيث يمثل 256 الطول الأقصى للتسلسل، و768 بُعد التمثيل المستخدم في نموذج MARBERT

ثالثاً : آلية الانتباه المتعدد الرؤوس (Multi-Head Attention):

تدخل مصفوفة المدخلات X إلى طبقات Transformer الاثنتي عشرة. وداخل كل طبقة تعمل 12 رأسًا للانتباه بالتوازي، حيث يُحوَّل تمثيل كل كلمة إلى ثلاثة متجهات تمثل الاستعلامات (Queries) والمفاتيح (Keys) والقيم (Values)، وذلك من خلال ضرب مصفوفة المدخلات في مصفوفات أوزان قابلة للتعلم:

$$Q = XW_Q \quad .....17$$

$$K = XW_K \quad .....18$$

$$V = XW_V \quad \text{.....19}$$

حيث تمثل W_K وW_Q وW_V مصفوفات الأوزان القابلة للتعلم. وفي كل رأس تنتج مصفوفات K وV و Q بأبعاد (64×256). وتُستخدم هذه المصفوفات لحساب درجة ارتباط كل كلمة ببقية الكلمات داخل التسلسل، مما يسمح للنموذج بالتركيز على الكلمات الأكثر أهمية في فهم المعنى والسياق. ثم يُحسب الانتباه وفق المعادلة :

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right) \times V \quad \text{.....02}$$

حيث يمثل $d_K = 64$ بُعد مصفوفات المفاتيح، ويُستخدم القسمة على $\sqrt{d_K}$ لضمان الاستقرار العددي وتجنب تشبع دالة Softmax. كما تُطبَّق آلية الحجب (Masking) لإهمال مواضع الحشو [PAD] وتركيز الأوزان على الكلمات الفعلية فقط.

بعد ذلك تُدمج مخرجات الرؤوس الاثني عشر لإنتاج تمثيل موحد، ثم تُطبّق عملية(Add(Residual Connection بإضافة دخل الطبقة إلى خرجها، يليها Norm(Layer Normalization)بهدف تحسين استقرار التمثيلات وتسريع عملية التدريب:

$$\text{Output} = \text{LayerNorm} + (X+\text{Sublayer}(X)) \quad \text{.....20}$$

ثم تمر التمثيلات الناتجة إلى الشبكة الأمامية (Feed Forward Network)، وبعدها تُطبّق مرة أخرى نفس عملية Residual + LayerNorm قبل تمريرها إلى الطبقة التالية من النموذج.

رابعاً : الدمج والشبكة الأمامية (Concatenation & Feed-Forward):

بعد حساب مخرجات رؤوس الانتباه المتعددة (Multi-Head Attention)، يتم دمج هذه المخرجات (Concatenation) ثم إسقاطها خطيًا للحصول على تمثيل موحد. بعد ذلك، تُمرَّر التمثيلات الناتجة إلى الشبكة الأمامية (Feed-Forward Network - FFN)، والتي تُطبَّق بشكل مستقل على كل عنصر من عناصر التسلسل.

تتكون الشبكة الأمامية (FFN) من طبقتين خطيتين تفصل بينهما دالة التنشيط GELU. في الطبقة الأولى، يتم توسيع البعد من 768 إلى 3072 لزيادة القدرة التعبيرية للنموذج، ثم تُطبَّق دالة GELU لإدخال اللاخطية وتحسين جودة

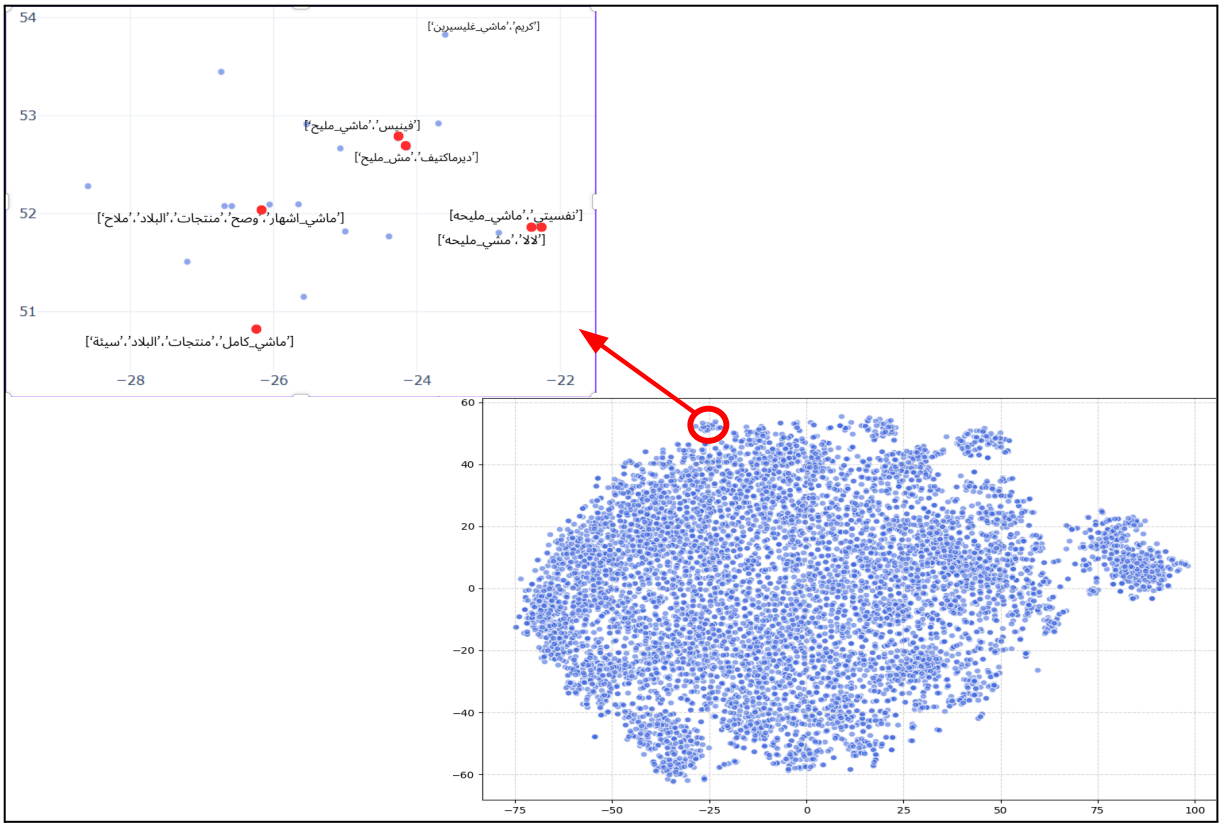
التمثيلات المتعلمة. بعد ذلك، تُستخدم الطبقة الخطية الثانية لإعادة إسقاط التمثيل إلى البعد الأصلي 768، مما يضمن توافق أبعاد المخرجات مع بقية طبقات النموذج.

ويمكن تمثيلها رياضيًا كما يلي، حيث تمثل $(W1)$ و$(W2)$ مصفوفتَي الأوزان القابلتين للتعلم، بينما يمثل $(b1)$ و$(b2)$ متجهي الانحياز :

$$FFN(x) = max(0, xW1 + b1)W2 + b2 \quad \text{.....21}$$

تتكون بنية المشفّر من 12 كتل متماثلة (Encoder Blocks)، حيث تضم كل كتلة طبقة Multi-Head Attention، تليها Feed-Forward Network (FFN)، مع تطبيق Residual Connections و Layer Normalization في كل مرحلة لضمان استقرار عملية التدريب وتحسين جودة التمثيلات المتعلمة.

وتظهر الصورة 11.3مخطط لتضمينات MARBERT حيث يتجلى التوزيع الكثيف للنقاط مع ظهور عنقود منفصل على يمين المخطط، مما يُشير إلى قدرة النموذج على التمييز بين مجموعات دلالية متباينة داخل البيانات .



الصورة 11.3: مخطط التوزيع الفضائي لتضمينات نموذج MARBERT

آلية عمل النموذج على البيانات

يختلف نموذج MARBERT عن النماذج التقليدية، إذ لا يتم تدريبه من الصفر، بل يعتمد على تمثيلات لغوية مسبقة التعلم تم اكتسابها من كميات ضخمة من النصوص العربية والعامية، ثم يتم تكييفه لاحقًا من خلال عملية الضبط الدقيق (Fine-Tuning) لمهمة تصنيف المشاعر في هذه الدراسة.

المرحلة الأولى: تقسيم البيانات

على عكس بعض النماذج التي تعتمد تقسيمًا ثنائيًا، تم اعتماد تقسيم ثلاثي للبيانات: 80% للتدريب، 10% للتحقق (Validation)، و10% للاختبار (Test)، مما يسمح بمراقبة أداء النموذج أثناء التدريب واختيار أفضل إعدادات ممكنة.

المرحلة الثانية: الترميز (MARBERT Tokenizer)

يتم ترميز التعليقات باستخدام Tokenizer الخاص بـ MARBERT مع حد أقصى يبلغ 256 رمزًا، مع إضافة الرموز الخاصة مثل [CLS] و [SEP]. يعتمد هذا الترميز على أسلوب WordPiece الذي يقوم بتقسيم الكلمات غير المعروفة إلى وحدات فرعية، مما يحسن قدرة النموذج على التعامل مع الدارجة وأخطاء الكتابة.

المرحلة الثالثة: الضبط الدقيق (Fine-Tuning)

تم ضبط النموذج باستخدام مجموعة من المعاملات الموضحة في الجدول 8.3، حيث تشمل عدد العصور التدريبية (num_train_epochs)، معدل التعلم (learning_rate)، نسبة warmup، معامل تنظيم الأوزان (weight_decay)، وآلية التوقف المبكر (EarlyStopping)، وذلك لضمان أفضل أداء ممكن وتجنب الإفراط في التعلم.

المعامل	القيمة	المبرر
num_train_epochs	3	كافٍ للضبط الدقيق دون إفراط في التعلم
learning_rate	2e-5	معدل تعلم منخفض للحفاظ على الأوزان المسبقة
warmup_ratio	0.1	تثبيت بداية التدريب تدريجياً
weight_decay	0.01	تنظيم لمنع الإفراط في التعلم
EarlyStopping	patience=2	إيقاف التدريب عند توقف التحسن

جدول 8.3: جدول إعدادات الضبط الدقيق (Hyperparameters) لنموذج MARBERT

طبقة التصنيف (Classification Layer)

تعتمد مرحلة التصنيف في نموذج MARBERT على التمثيل الناتج عن رمز [CLS] الذي يمثل ملخصًا سياقيًا شاملًا للنص. يتم تمرير هذا التمثيل إلى طبقة Dense مكونة من 128 وحدة عصبية لدمج المعلومات المستخرجة من طبقات Encoder وتحويلها إلى تمثيل مناسب لاتخاذ القرار.

بعد ذلك، يتم تطبيق Dropout بنسبة 0.5 أثناء التدريب للحد من الإفراط في التعلم وتحسين قدرة النموذج على التعميم. في المرحلة النهائية، يتم استخدام طبقة Dense بثلاث وحدات مع دالة Softmax لإنتاج توزيع احتمالي عبر الفئات الثلاث (إيجابي، سلبي، محايد)، ويتم اختيار الفئة ذات الاحتمال الأعلى كتصنيف نهائي للنص.

النتائج

تم تقييم النموذج باستخدام مقاييس Accuracy وF1-Macro إلى جانب F1-score لكل فئة على حدة

النموذج	المعالجة	Accuracy	F1 Macro	F1 (N)	F1 (NEU)	F1 (P)
MARBERT	Class Weights	0.79	0.72	0.64	0.85	0.68

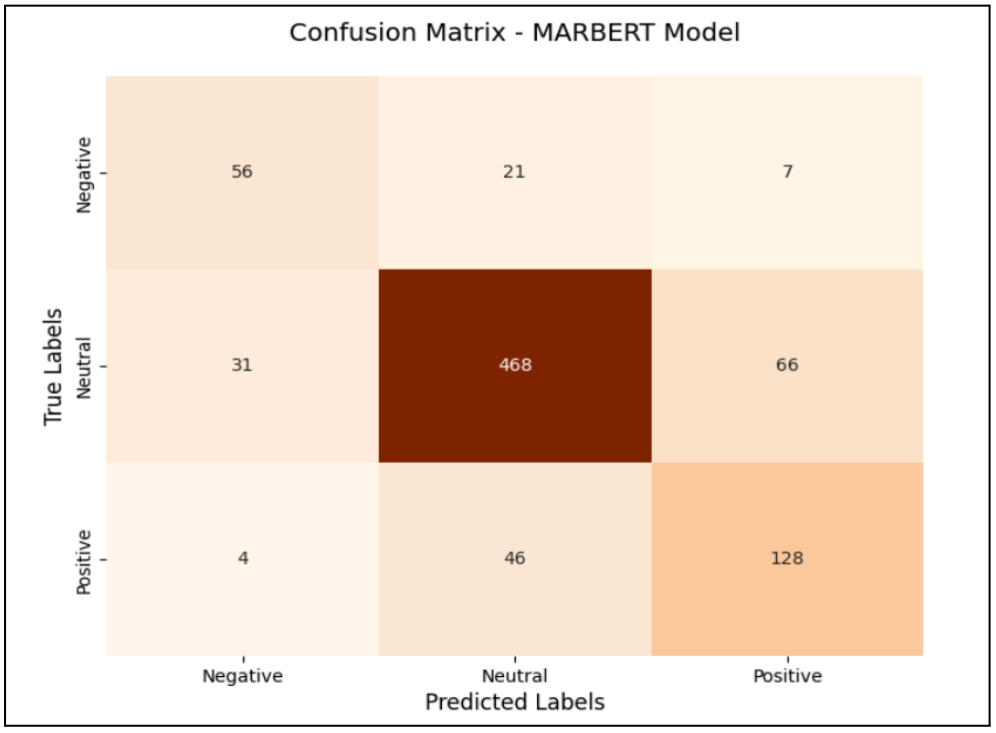
جدول 9.3: نتائج نموذج الMARBERT

هـ- تحليل مصفوفة الارتباك والتفسير

تكشف مصفوفة الارتباك التالية عن ثلاث ملاحظات جوهرية:

- **الفئة المحايدة (NEU)** حقّقت 468 تصنيفاً صحيحاً من أصل 565، مع 31 خطأً نحو السلبي و66 نحو الإيجابي. والملاحَظ أن هذا التوزيع أكثر تناسقاً مقارنة بالنموذجين السابقين، مما يدل على أن MARBERT يفهم حدود الفئة المحايدة بشكل أدق؛
- **الفئة السلبية (N)** سجّلت 56 تصنيفاً صحيحاً من أصل 84، مع 21 خطأً نحو المحايد و7 نحو الإيجابي. والأهم هنا أن نسبة التصنيف الصحيح بلغت 67%، وهو تحسّن جوهري مقارنة بـ57% = SVM و61 =CNN، ويعكس قدرة MARBERT على فهم أساليب النفي والتعبير السلبي الضمني في الدارجة؛

- **الفئة الإيجابية (P)** حقّقت 128 تصنيفاً صحيحاً من أصل 178، مع 46 خطأً نحو المحايد و4 فقط نحو السلبي. ويُلاحَظ الانخفاض الكبير في الخطأ السلبي/إيجابي (4 مقابل 7 في SVM و2 في CNN)، مما يعني أن النموذج نادراً ما يقع في الخطأ الجسيم المتمثل في الخلط بين القطبين المتعاكسين.



صورة 12.3: مصفوفة الارتباك لنموذج MARBERT

الخلاصة التفسيرية:

يتفوق MARBERT على النموذجين الآخرين لسببين جوهريين:

أولاً، آلية الانتباه (Attention Mechanism) التي تمكّنه من ربط الكلمات ببعضها بغض النظر عن مواقعها في الجملة، فيفهم السياق الكامل لا الأنماط المحلية فقط. وثانياً، التدريب المسبق على العربية العامية الذي يمنحه معرفة لغوية بالدارجة لم تتوفر للنموذجين السابقين اللذين بنيا تمثيلاتهما من الصفر انطلاقاً من بياناتنا المحدودة.

6.3 الاستنتاجات التحليلية للنماذج

يتيح الجمع بين نتائج النماذج الثلاثة استخلاص جملة من الاستنتاجات التحليلية:

النموذج	المعالجة	Accuracy	F1 Macro	F1 (N)	F1 (NEU)	F1 (P)
MARBERT	Class Weights	0.79	0.72	0.64	0.85	0.68
SVM + TF-IDF	Class Weight	0.77	0.68	0.57	0.84	0.63
CNN + FastText	Class Weights	0.74	0.63	0.53	0.83	0.52

جدول 10.3:- نتائج تقييم النماذج

- **أولاً تفوق MARBERT في الفئات الأصعب:** الفارق الجوهري بين MARBERT والنموذجين الآخرين لا يتجلى في الدقة الكلية (فرق بسيط لا يتجاوز 5 نقاط) بل في أدائه على الفئات الأقل تمثيلاً. فقد حقق F1=0.64 للفئة السلبية و0.68 للإيجابية، مقابل 0.58 و0.62 لـ SVM، وهو ما يُثبت قدرته على التعامل مع التفاوت في توزيع البيانات بكفاءة أعلى؛

- **ثانياً تفوق SVM على CNN في غياب بيانات كافية:** النتيجة اللافتة هي أن SVM+TF-IDF تفوّق على CNN+FastText رغم بساطته النسبية. ويُعزى ذلك إلى أن نماذج التعلم العميق تحتاج إلى كميات أكبر من البيانات لتحقيق كل امكانياتها، في حين يؤدي SVM أداءً متيناً حتى مع البيانات المحدودة؛

- **ثالثاً التحدي المشترك للفئة السلبية:** أظهرت النماذج الثلاثة صعوبة مشتركة في تصنيف التعليقات السلبية، وهو ما يعكس خصوصية لغوية حقيقية في الدارجة الجزائرية، إذ كثيراً ما يُعبَّر عن الرفض والنقد بأساليب ضمنية أو ساخرة تفتقر إلى المؤشرات اللغوية الصريحة التي تعتمد عليها النماذج في التصنيف؛

- **رابعاً الاختيار النهائي:** بناءً على مجمل هذه النتائج، يُعتمَد **MARBERT** نموذجاً نهائياً للمنصة المقترحة، كونه يحقق أفضل توازن بين الدقة الكلية والأداء على الفئات الثلاث، ولا سيما في السياق اللغوي الجزائري الذي يستفيد بشكل مباشر من التدريب المسبق على العربية العامية.

7.3 الخاتمة

وبناءً على ما تم عرضه في هذا الفصل، تم بناء إطار تطبيقي شامل مكّن من اختبار مختلف النماذج على بيانات واقعية تنتمي إلى سياق الدارجة الجزائرية في مجال منتجات العناية بالبشرة. وقد سمح هذا العمل بتقييم فعالية المقاربات المختلفة واستخلاص الفروق الجوهرية بينها من حيث القدرة على فهم النصوص وتصنيفها.

كما أبرزت النتائج أهمية اختيار التمثيل المناسب للنصوص ومعالجة عدم توازن الفئات في تحسين أداء النماذج، خاصة في سياق لغوي غير رسمي يتميز بالتنوع والاختلاف في الصياغة.

وتمثل هذه النتائج مرحلة انتقالية أساسية نحو الفصل الموالي، الذي يهدف إلى توظيف مخرجات النماذج ضمن منظومة تطبيقية متكاملة، تمهيدا لتصميم منصة ذكاء تسويقي تُحوّل مخرجات هذه النماذج إلى توصيات استراتيجية قابلة للاستثمار المباشر.

الفصل الرابع: تجسيد وبناء النظام المقترح

1.4 مقدمة

نستعرض في هذا الفصل الحل المقترح وتجسيده الفعلي في شكل نظام برمجيات متكامل (ويب). نقوم بوصف الأجهزة والعتاد، الأدوات البرمجية، وبيئات التطوير المستخدمة في البناء، إلى جانب استعراض هندسة النظام (Architecture) والواجهات الرسومية المنشأة التي توضح الخصائص والميزات الأساسية للتطبيق.

2.4 بيئة التطوير

1.2.4 العتاد والأجهزة (Hardware)

تم الاعتماد على الأجهزة التالية لتطوير واختبار النظام ومحاكاة عمليات معالجة البيانات النصية:

اسم الجهاز	الميزات
حاسوب Dell	RAM: 16GB
	CPU:11th Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz
	Hard Disk: 512 SSD
	Operating System: Windows 11 Pro
حاسوب Lenovo	RAM: 8 GB
	CPU:Intel(R) Core(TM) i7-7500U @ 2.70GHzO
	Hard Disk: 1 TB HDD
	Operating System: Windows 10 Pro

جدول1.4 : العتاد المستخدم في التطوير

2.2.4 الأدوات البرمجية

1.2.2.4 React

هي مكتبة جافا سكريبت (JavaScript Library) مفتوحة المصدر تديرها شركة ميتا (فيسبوك سابقاً). تُستخدم على نطاق واسع لإنشاء واجهات مستخدم تفاعلية وسريعة الاستجابة (Reactive User Interfaces). وتعتمد على تقسيم الواجهة إلى مكونات مستقلة وقابلة لإعادة الاستخدام (Reusable Components)، مما يسهل بناء وصيانة تطبيقات الويب المعقدة وضمان كفاءة الأداء [19].

2.2.2.4 Vite

هو أداة بناء (Build Tool) حديثة وبيئة تطوير برمجية تتميز بالسرعة الفائقة والخفة. يعتمد Vite على تقنية وحدات الجافا سكريبت الأصلية في المتصفح (Native ES Modules) لتوفير خادم تطوير سريع جداً ولحظي، بالإضافة إلى ميزة التحديث الفوري للأكواد (Hot Module Replacement) دون الحاجة لإعادة تحميل الصفحة بالكامل. يدعم Vite بيئات عمل متعددة مثل React و Vue.js، وقد تم اختياره في هذا النظام لتسريع وتيرة التطوير وضمان الأداء العالي أثناء معالجة وعرض البيانات المستمرة [19].

3.2.2.4 FastAPI

هو إطار عمل (Framework) حديث، عالي الأداء، ومفتوح المصدر مخصص لبناء واجهات برمجة التطبيقات (APIs) باستخدام لغة Python، ويعتمد بشكل أساسي على تلميحات الأنواع القياسية (Python Type Hints). تكمن القوة التقنية لهذا الإطار في الميزات الأساسية التالية [20]:

- **الأداء الفائق وسرعة المعالجة:** يوفر كفاءة تشغيلية عالية جداً تقارن بأداء بيئات تشغيل مثل Node.js و Go، وذلك بفضل اعتماده على مكتبات حديثة مثل Starlette لعمليات الويب و Pydantic للتحقق من البيانات؛
- **كفاءة وتطوير متسارع:** يساهم في تسريع عملية كتابة البرمجيات وبناء الميزات بنسب تتراوح بين 200% إلى 300%، مع تقليل الأخطاء البشرية البرمجية بنسبة تصل إلى 40% نتيجة التحقق اللحظي من المدخلات؛
- **بيئة برمجية بديهية وسهلة التعلم:** يوفر دعماً استثنائياً لمحاميل الأكواد (إكمال تلقائي وتصحيح لحظي)، مما يقلل الوقت المستغرق في مراجعة التوثيقات وتتبع الأخطاء البرمجية؛

- **المتانة والتوثيق التلقائي التفاعلي (Robustness):** ينتج أكواداً برمجية جاهزة للبيئات الإنتاجية الحية، ويقوم بتوليد واجهات توثيق تفاعلية بشكل آلي للمطورين بالاعتماد على معايير OpenAPI (Swagger) و JSON Schema..

Node.js 4.2.2.4

بيئة تشغيل مجانية ومفتوحة المصدر ومتعددة المنصات لتشغيل جافا سكريبت تم استخدامها لإنشاء خادم محلي خفيف وسريع (JSON Server). يهدف هذا الخادم إلى محاكاة عمليات قاعدة البيانات وإدارة صلاحيات المستخدمين والتحقق من الهوية بشكل مرن وسلس في مراحل التطوير الحالي[21].

Firebase 5.2.2.4

منصة سحابية متكاملة تقدمها شركة Google. تم استخدامها لإدارة وتخزين البيانات السحابية الحية؛حيث تُستغل قاعدة البيانات الآنية (Realtime Database) لتحديث التقارير فوراً، بينما تُستخدم (Firestore) كقاعدة بيانات وثائقية مرنة لحفظ السجلات والمستخدمين[22].

Visual Studio Code 6.2.2.4

بيئة تحرير وتطوير برمجية متكاملة ومفتوحة المصدر (مستندة إلى رخصة MIT الحرة)، تم تطويرها من قِبل شركة Microsoft لتدعم طيفاً واسعاً من لغات البرمجة وعمليات التطوير عبر منظومة إضافات (Extensions) مرنة. يتميز المحرر بدعمه البرمجي المتقدم لتقنيات الإكمال التلقائي الذكي، تمييز النحو البرمجي (Syntax Highlighting)، أدوات تصحيح الأخطاء (Debugging)، والتحكم في النسخ عبر نظام Git مدمج[23].

3.4 التقنيات المدمجة في النظام (Implemented Technologies)

التقنيات البرمجية التي تم دمجها معاً لتعمل كتلة واحدة داخل النظام هي :

- **الواجهة الأمامية:** React | Vite | JavaScript
- **الواجهة الخلفية والذكاء الاصطناعي:** FastAPI | Node.js | Python
- **قواعد البيانات والسحابة:** Firebase Firestore | Realtime DB | JSON

- **النماذج اللغوية:** MARBERT | Gemini API | Sentiment Model .pkl

4.4 مكونات النظام الأساسية (Application Components)

ينقسم النظام برمجياً ووظيفياً إلى ثلاثة مستويات مترابطة:

1. مستوى العرض (The View): ويمثله تطبيق React المسؤول عن عرض واجهات المستخدم ولوحات التحكم وتوفير آليات التفاعل لكلٍّ من المشرف والعميل.
2. مستوى الخدمات والتحكم (The Controllers / Backend Services): ويتكون من خادمين رئيسيين:
   - خادم التطبيق (Node.js): يتولى إدارة منطق الأعمال العام، بما في ذلك عمليات التسجيل، وتسجيل الدخول، والتحقق من الهوية (Authentication).
   - خادم الذكاء الاصطناعي (FastAPI): يستقبل طلبات تحليل المشاعر، ويستخدم مكتبة youtube_comment_download لاستخراج التعليقات من مقاطع YouTube، ثم يمرر التعليقات المستخرجة إلى نماذج التحليل لإجراء عملية التنبؤ وإرجاع النتائج.
3. مستوى النماذج والبيانات (The Models and Data Layer): ويشمل نماذج الذكاء الاصطناعي المستخدمة في التصنيف، سواء المحلية أو المستضافة (MARBERT وGPT-4o-mini)، بالإضافة إلى مكونات تخزين البيانات المتمثلة في قاعدة البيانات السحابية Firebase والملف المحلي db.json.

5.4 واجهات التطبيق (Application Interfaces)

1.5.4 الواجهة الرئيسية للمنصة (Home Page Interface)

تُمثل الصورة 1.4 الواجهة الرئيسية، والتي تُعد البوابة الأساسية ونقطة الانطلاق الأولى للمستخدمين داخل المنصة. وقد تم تصميم هذه الواجهة وتطويرها باستخدام مكتبة React وبيئة البناء Vite، مع التركيز على توفير تجربة مستخدم تفاعلية وواضحة ومبسطة تدعم اللغة العربية. وقد قُسِّمت الصفحة الرئيسية من الناحيتين البرمجية والوظيفية إلى ثلاثة أقسام رئيسية متتالية، بهدف تعريف المستخدم بآلية عمل المنصة والخدمات التي تقدمها قبل الانتقال إلى مرحلة تسجيل الدخول أو إنشاء حساب جديد.

الصورة 1.4 : الواجهة الرئيسية للمنصة و قسم التعريف بخدمة تحليل المشاعر.

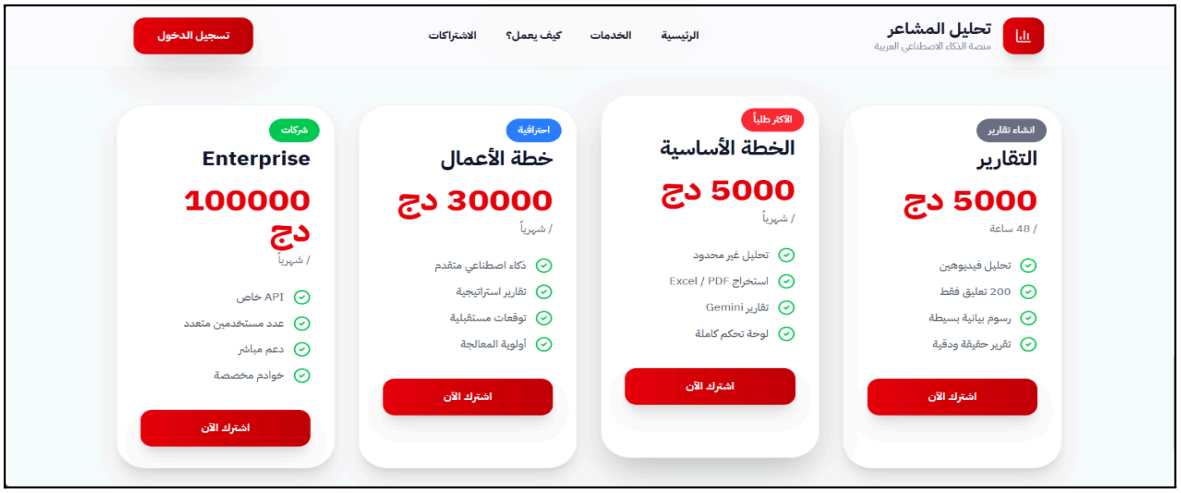
تمثل الصورة 2.4 المميزات والخدمات التي توفرها المنصة للمستخدمين، والتي تشمل استخراج التعليقات من مقاطع يوتيوب وتصنيفها وفقاً للمشاعر المختلفة، بالإضافة إلى إنشاء رسوم بيانية ومؤشرات إحصائية تساعد على فهم نتائج التحليل بصورة أوضح. كما تتيح المنصة إنشاء تقارير ذكية مدعومة بتقنيات الذكاء الاصطناعي، مع إمكانية تصدير هذه التقارير وطباعتها، فضلاً عن إنشاء ملف بصيغة Excel يتضمن جميع التعليقات المستخرجة ونتائج تصنيفها، مما يسهّل عملية حفظ البيانات وإعادة استخدامها في التحليلات اللاحقة.

الصورة 2.4 : قسم المميزات والخدمات المتاحة في المنصة

تمثل الصورة 3.4 آلية عمل النظام المقترح، حيث تبدأ العملية بإدخال رابط الفيديو المراد تحليله، ثم استخراج التعليقات المرتبطة به وإجراء عمليات المعالجة المسبقة اللازمة لتنظيف البيانات وتجهيزها. بعد ذلك، تُمرَّر التعليقات إلى نموذج تحليل المشاعر المدرب بالاعتماد على MARBERT لإجراء عملية التصنيف والتنبؤ بالفئات الشعورية المناسبة. وفي المرحلة الأخيرة، تُعرَض النتائج للمستخدم في شكل تقارير ورسوم بيانية و إحصائيات تسهّل عملية تفسير المخرجات والاستفادة منها.

صورة 3.4 : قسم كيف يعمل ؟ آلية عمل نظام تحليل المشاعر المقترح.

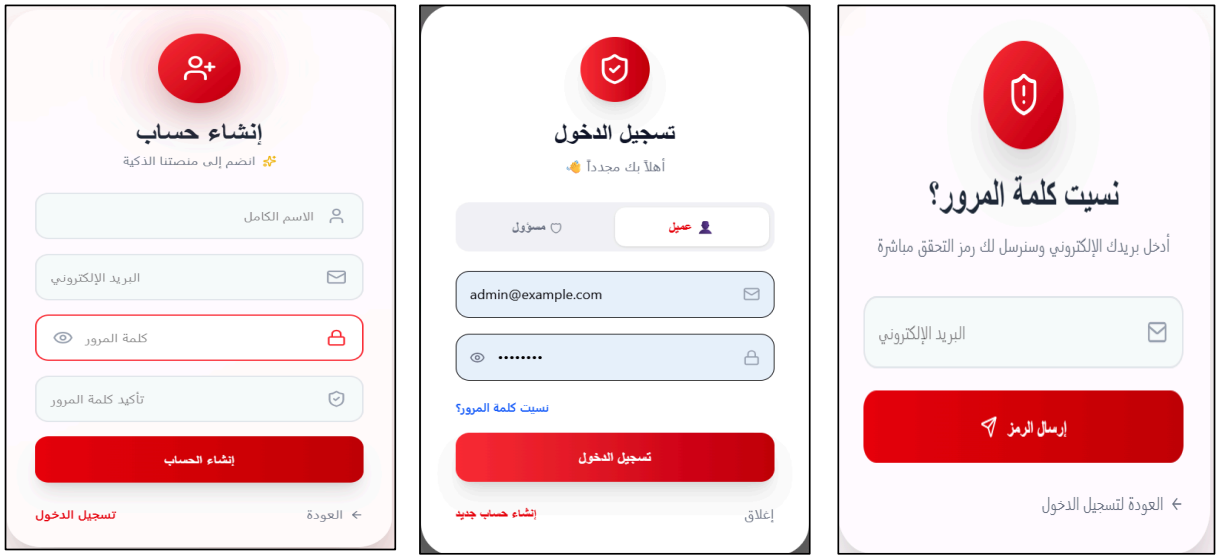
تمثل الصورة 4.4 خطط الاشتراك المتاحة في المنصة، حيث توفر المنصة عدة باقات تختلف فيما بينها من حيث السعر والخصائص والخدمات المقدمة. وتتيح هذه الخطط للمستخدمين اختيار الباقة التي تتناسب مع احتياجاتهم ومتطلبات استخدامهم للنظام.



صورة 4.4: خطط الاشتراك المتاحة في المنصة ومزاياها.

2.5.4 واجهات المصادقة وإدارة الحسابات

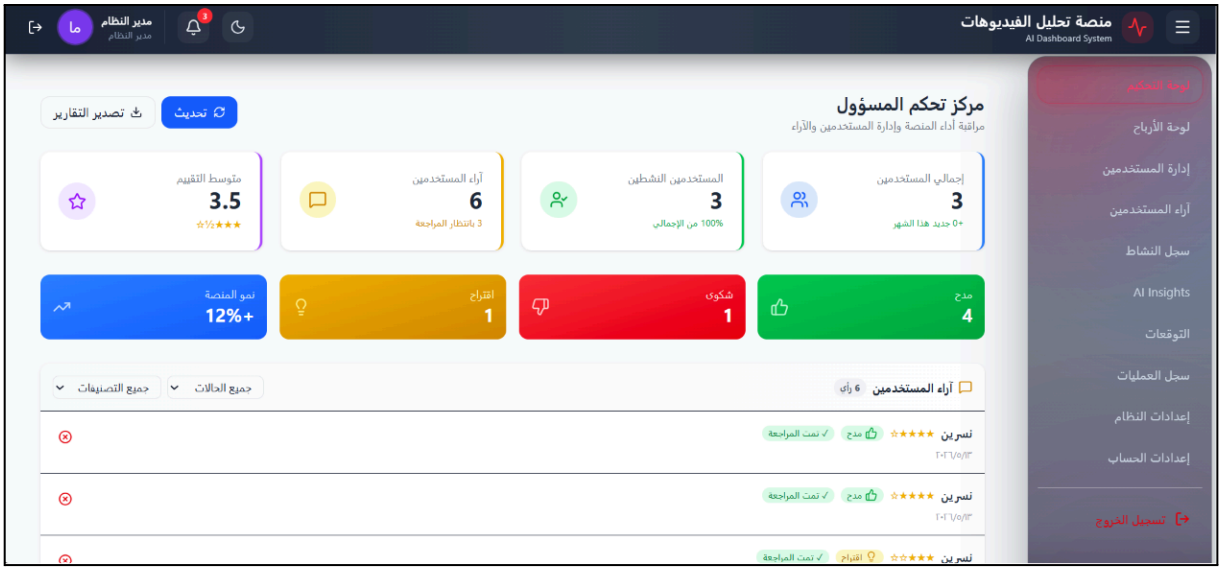
تمثل الصورة 5.4 واجهات مخصصة للمستخدم تتيح إنشاء حساب جديد، تسجيل الدخول، أو طلب رمز مؤقت لإعادة تعيين كلمة المرور في حال نسيانها. وتعتمد هذه الواجهة في الخلفية على بروتوكول JWT لتأمين الجلسات وإدارة عمليات المصادقة، بالإضافة إلى استخدام مكتبة bcryptjs لتشفير كلمات المرور وحمايتها من أي وصول غير مصرح به.



صورة 5.4: واجهة تسجيل الدخول وإنشاء الحساب واستعادة كلمة المرور

3.5.4 إدارة المستخدمين الخاصة بالمشرف

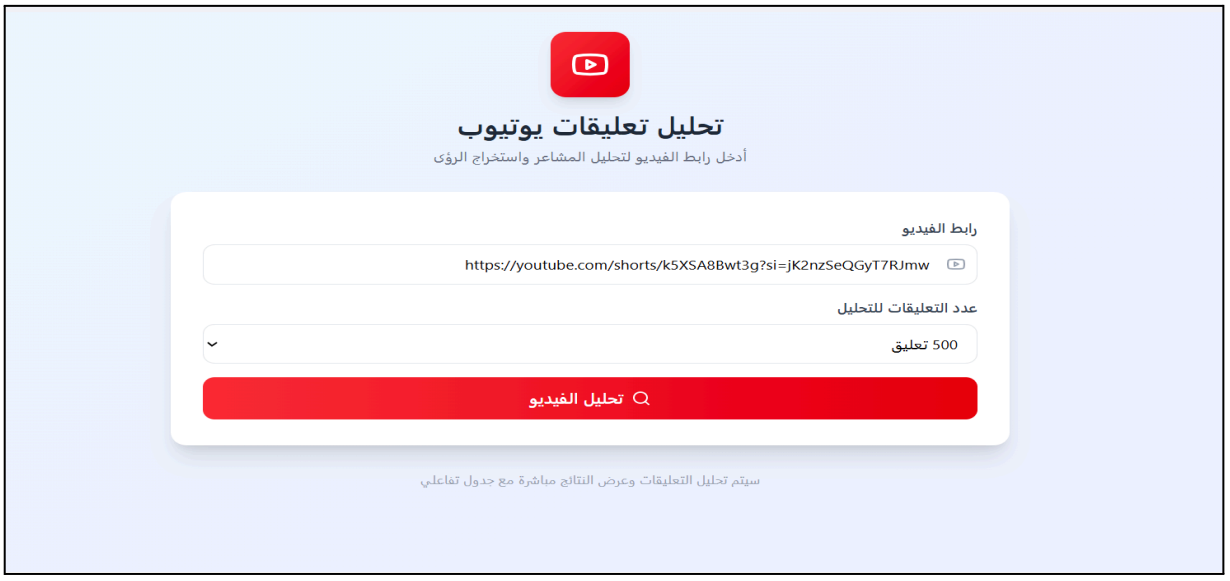
تعرض الصورة 6.4 واجهة الإحصائيات العامة الخاصة بالنظام، والتي توفر نظرة شاملة حول أداء المنصة، مثل عدد المستخدمين النشطين، وسجلات الأنشطة الحالية (Activity Logs)، إضافة إلى مراقبة أداء النظام والإعدادات العامة. وتهدف هذه الواجهة إلى تمكين المشرف من متابعة حالة النظام بشكل لحظي واتخاذ القرارات المناسبة عند الحاجة.



صورة 6.4: إدارة المستخدمين الخاصة بالمشرف

4.5.4 واجهات العميل (Client & Sentiment Analysis)

تعرض الصورة 7.4 واجهة مخصصة للمستخدم (العميل) لإدخال رابط فيديو يوتيوب المراد دراسته. يقوم النظام مباشرة باستخراج التعليقات المتعلقة بالفيديو، ثم معالجتها باستخدام نموذج MARBERT ونماذج تحليل المشاعر، قبل عرض النتائج بشكل نسبي (إيجابي، سلبي، محايد) على شكل رسوم بيانية تفاعلية، مع دعم التحديث الآني للبيانات عبر Firebase.

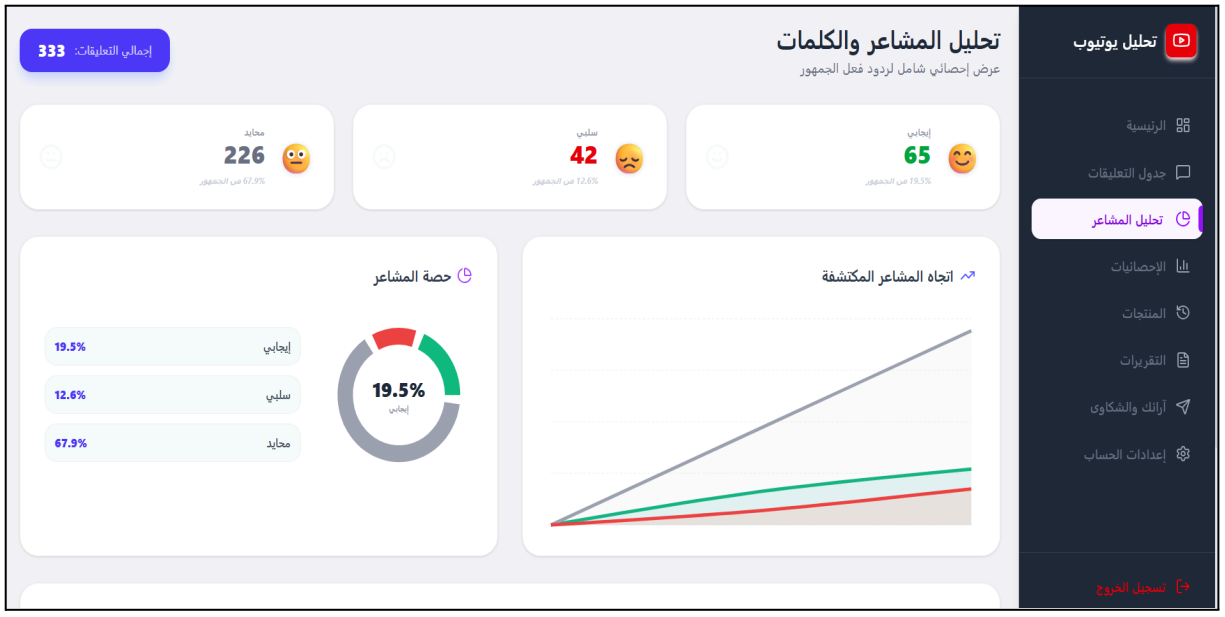


الصورة 7.4: واجهة العميل استخراج التعليقات وتحليلها

توضح الصورة 8.4 قسم إدارة وتصنيف التعليقات المستخرجة من منصة "يوتيوب". تظهر الواجهة تقسيماً منظماً يحتوي على قائمة جانبية للتنقل بين مختلف وظائف النظام، بالإضافة إلى جدول مركزي يعرض بيانات التعليقات: اسم الحساب، نص التعليق بالعامية، ومستوى التفاعل (عدد الإعجابات). كما يبرز الجدول نتائج معالجة اللغة الطبيعية والتحليل الآلي للمشاعر، حيث يتم تصنيف كل تعليق تلقائياً إلى (إيجابي، سلبي، أو محايد) مع توفير أدوات لتصفية البيانات، البحث، وتصدير التقارير بصيغة Excel.

صورة 8.4: جدول التعليقات والتحليل الآلي

توضح صورة 9.4عرضاً إحصائياً شاملاً لردود فعل الجمهور. تحتوي الواجهة في أعلاها على بطاقات رقمية تعرض العدد الإجمالي للتعليقات المستخرجة (333 تعليقاً)، مصنفة إلى: تعليقات إيجابية (65)، سلبية (42)، ومحايدة (226) مع نسبها المئوية. كما تتضمن الواجهة مخططيْن بيانيين الأيمن يمثل **"اتجاه المشاعر المكتشفة"** عبر منحنيات بيانية لمراقبة التطور الزمني والمقارنة بين الفئات، والأيسر يمثل **"حصة المشاعر"** على شكل مخطط دائري مجوف (Donut Chart) يبرز الهيمنة النسبية لكل فئة من المشاعر لتسهيل القراءة السريعة للبيانات المستهدفة.



صورة 9.4: واجهة لوحة التحكم الإحصائية لتحليل اتجاهات

توضح الصورة 10.4 خلاصة تحليل المحتوى الذكي (بواسطة Gemini AI). تبرز الواجهة في جزئها العلوي إحصائيات نوعية سريعة للتعليقات، بينما ينقسم الجزء السفلي إلى قسمين تحليليين رئيسيين يعتمدان على المعالجة الدلالية للنصوص: القسم الأيمن مخصص لاستخراج **"نقاط القوة"** والميزات الإيجابية للمنتجات المذكورة في التعليقات (مثل الكفاءة، الترطيب، والتوصيات الطبية)، والقسم الأيسر مخصص لرصد **"نقاط الضعف"** والشكاوى المستخلصة (مثل الأسعار، المشاكل الجلدية المحتملة، أو دعوات المقاطعة). وتتيح الواجهة أيضاً خيار "طباعة التقرير" لاستخدامه في اتخاذ القرارات التسويقية.

صورة 10.4: واجهة التقرير الذكي لملخص نقاط القوة والضعف المستخلصة آلياً من التعليقات.

6.4 خلاصة

قدم هذا الفصل نظرة شاملة وعميقة حول البنية التقنية والتطبيقية لنظام تحليل المشاعر المطوّر. من خلال استعراض أدوات التطوير، لغات البرمجية المتنوعة، وهندسة الخدمات المصغرة الهجينة (FastAPI & Node.js)، بالإضافة إلى قواعد البيانات السحابية والمحلية، أظهرنا كيف تتكامل هذه المكونات برمجياً لتشكل نظاماً فعالاً، آمناً، وسهل الاستخدام لمعالجة وفهم البيانات النصية.

الفصل الخامس : الجانب الاقتصادي

1.5 المقدمة

يُعدّ هذا الفصل الإطار التطبيقي للمشروع، إذ ينتقل من البعد البحثي والأكاديمي إلى البعد الريادي والعملي. فإذا كانت الفصول السابقة قد أرست الأسس النظرية وقدّمت النماذج التقنية لتحليل مشاعر الدارجة الجزائرية، فإن هذا الفصل يُجسّد الرؤية التجارية لتحويل هذه النتائج إلى منتج رقمي قابل للتسويق والاستدامة.

يُقدّم الفصل المشروع بوصفه شركة ناشئة (Startup) في مجال الذكاء الاصطناعي وتحليل البيانات، موجّهة للسوق الجزائري، وذلك من خلال ستة نقاط متكاملة تشمل: تقديم الفكرة وقيمها المضافة، الجوانب الابتكارية، التحليل الاستراتيجي للسوق، خطة الإنتاج والتنظيم، التوقعات المالية، إضافة إلى النموذج الأولي التجريبي وخارطة الطريق.

ويهدف هذا الفصل في مجمله إلى إبراز أن المشروع لا يقتصر على كونه إنجازاً أكاديمياً، بل يمثّل فرصة ريادية ذات جدوى اقتصادية ومجتمعية، في ظل الحاجة المتزايدة إلى حلول ذكية لتحليل البيانات اللغوية المحلية، خاصة الدارجة الجزائرية.

2.5 المحور الأول : تقديم المشروع

1.2.5 فكرة المشروع (الأبعاد الاستراتيجية والحل المقترح)

يرتبط نجاح أي مؤسسة اقتصادية ذات نشاط خدمي بمدى قدرتها على التفاعل مع زبائنها وفهم متطلباتهم، وفي ظل التحول الرقمي المتسارع، أصبحت منصات التواصل الاجتماعي في السنوات الأخيرة فضاءً رئيسياً لتبادل الآراء والتجارب الاستهلاكية، الأمر الذي جعلها مصدراً مهماً للمعلومات التي تعتمد عليها المؤسسات في فهم سلوك الزبائن وتقييم مدى تقبلهم للمنتجات والخدمات المقدمة. وفي هذا السياق، يشهد المجتمع الجزائري توسعاً ملحوظاً في استخدام التقنيات الرقمية. حيث بلغ عدد مستخدمي الإنترنت في الجزائر 36.2 مليون مستخدم مع بداية سنة 2025 ، أي ما يعادل 76.9% من إجمالي السكان. كما قُدِّر عدد مستخدمي منصات التواصل الاجتماعي بحوالي 25.6 مليون مستخدم، وهو ما يمثل 54.2% من مجموع السكان [24].

وقد أدى هذا التوسع إلى توليد كميات ضخمة من البيانات النصية يومياً، خاصة على شكل تعليقات ومنشورات مكتوبة باللهجة الجزائرية بمختلف تنوعاتها. غير أن أغلب أدوات تحليل المشاعر المتاحة تجارياً أو بحثياً صُممت أساساً للتعامل مع اللغة الإنجليزية أو العربية الفصحى وبعض اللهجات العربية الأخرى، مما يجعل قدرتها على فهم الخصائص اللغوية والدلالية للهجة الجزائرية محدودة، ويؤثر سلباً على دقة النتائج المستخلصة منها.

وتُعد منصة يوتيوب من أبرز المنصات الرقمية المؤثرة في السلوك الاستهلاكي للمستخدم الجزائري؛ إذ تشير بيانات Google الإعلانية إلى أن عدد مستخدمي يوتيوب في الجزائر بلغ حوالي 21.1 مليون مستخدم خلال سنة 2025 [24]، وهو ما يعكس الأهمية المتزايدة للمحتوى المرئي في تشكيل الانطباعات واتخاذ القرارات الشرائية. وتستغل العديد من المؤسسات والعلامات التجارية هذه المنصة للترويج لمنتجاتها وخدماتها، مما يؤدي إلى تراكم أعداد كبيرة من التعليقات التي تتضمن آراء المستهلكين ومواقفهم الحقيقية تجاه تلك المنتجات.

استجابةً لهذه التحديات، يقترح هذا المشروع تطوير منصة رقمية تفاعلية تعتمد على تقنيات الذكاء الاصطناعي ومعالجة اللغات الطبيعية، بهدف أتمتة عملية تحليل مشاعر المستهلك الجزائري انطلاقاً من التعليقات المنشورة على منصات التواصل الاجتماعي، مع اعتماد منصة يوتيوب كنموذج أولي للتطبيق.

تتيح المنصة للمستخدم إدراج رابط الفيديو المراد تحليله، لتتكفل الواجهة الخلفية بجمع التعليقات المرتبطة به ومعالجتها آلياً. وبعد تنفيذ مراحل التنظيف والمعالجة المسبقة، تُمرَّر البيانات إلى نماذج تصنيف المشاعر المدربة مسبقاً لتحديد طبيعة كل تعليق ضمن ثلاث فئات رئيسية: **إيجابي، سلبي، أو محايد**.

كما توفر المنصة لوحة قيادة تفاعلية (Dashboard) تعرض نتائج التحليل في شكل مؤشرات إحصائية ورسوم بيانية تسهّل على متخذي القرار فهم توجهات الجمهور وتقييم مدى تقبل الحملات التسويقية أو المنتجات المطروحة. ويسهم ذلك في تحويل البيانات النصية غير المنظمة إلى معلومات قابلة للاستثمار في تحسين الاستراتيجيات التسويقية وتعزيز جودة القرارات المبنية على البيانات.

2.2.5 القيم المقترحة

تمثل القيمة المقترحة جوهر الخدمات التي تقدمها المنصة، وتعكس المنافع العملية التي يحصل عليها العملاء عند استخدام الحل المقترح. وفي إطار هذا المشروع، تم تحديد مجموعة من القيم المضافة التي تستجيب لاحتياجات المؤسسات الاقتصادية والشركات الناشئة الساعية إلى الاستفادة من البيانات الرقمية في تحسين قراراتها التسويقية وتعزيز قدرتها التنافسية. ويبين الجدول التالي إسقاط هذه القيم على الهيكل الوظيفي للمشروع.

المؤشر	القيمة المقترحة	تجسيدها التطبيقي والاقتصادي في المنصة
الحداثة	معالجة موجهة للدارجة الجزائرية	يندرج المشروع ضمن المبادرات المحلية التي تسعى إلى توظيف تقنيات الذكاء الاصطناعي في تحليل النصوص المكتوبة بالدارجة الجزائرية، من خلال الاستفادة من نموذج MARBERT بعد إعادة تدريبه (Fine-Tuning) ليتلاءم مع طبيعة البيانات المستهدفة، بالإضافة إلى اعتماد قوائم كلمات مستبعدة مخصصة تدعم خصوصية السياق اللغوي المحلي.
الأداء والكفاءة	تحليل آلي سريع للبيانات	تتيح المنصة أتمتة عمليات جمع ومعالجة وتصنيف أعداد كبيرة من التعليقات المنشورة على منصة يوتيوب في زمن قصير، مما يساهم في تقليص الجهد البشري وتسريع الحصول على النتائج، مع الحد من التحيزات والأخطاء المرتبطة بالتقييم اليدوي.
إنجاز المهمة	دعم اتخاذ القرار التسويقي	تساعد المنصة المؤسسات على دراسة ردود فعل الجمهور تجاه منتجاتها أو حملاتها الإعلانية من خلال تصنيف المشاعر وتحويلها إلى مؤشرات قابلة للتفسير، بما يسهم في تقييم الأداء التسويقي واتخاذ قرارات أكثر استناداً إلى البيانات الفعلية.
خفض التكاليف	تقليل النفقات التشغيلية	يساهم الحل المقترح في تقليص الاعتماد على المتابعة اليدوية واسعة النطاق للتعليقات، مما يقلل من الوقت والموارد البشرية اللازمة لتحليل البيانات، ويحد من المخاطر المرتبطة بالاستمرار في حملات تسويقية لا تتوافق مع توقعات الجمهور المستهدف.
سهولة الوصول	واجهة تفاعلية مبسطة	صُممت المنصة وفق مبدأ البساطة وسهولة الاستخدام؛ حيث يمكن للمستخدم إدراج رابط المحتوى المراد تحليله للحصول على نتائج مرئية وإحصائية بصورة مباشرة، مع إمكانية الوصول إلى الخدمة عبر الويب دون الحاجة إلى تجهيزات متخصصة.

سهولة الاستخدام والملائمة	تكييف المعالجة حسب طبيعة البيانات	تتيح المنصة إمكانية الاستفادة من قوائم كلمات مستبعدة مخصصة (Custom Stop-Words) يمكن تكييفها بما يتناسب مع طبيعة البيانات أو القطاع المستهدف، بما يدعم جودة المعالجة ويزيد من ملاءمة النتائج لاحتياجات المستخدم النهائي.

جدول 1.5: القيم المقترحة للمشروع

3.2.5 فريق العمل

يتكوّن فريق المشروع من طالبتين في إطار مشروع تخرج لنيل شهادة الماستر في الإعلام الآلي، حيث تم توزيع المهام وفق المهارات التقنية لكل عضو.

العضو	الدور	المهارات
لقمة حفيظة	مسؤولة معالجة اللغة الطبيعية والنمذجة	معالجة اللغة الطبيعية، تدريب النماذج (SVM, CNN, MARBERT)
جعيل نسرين	مطورة المنصة	تطوير منصة الويب،قواعد البيانات،ربط النماذج عبر الAPI

جدول 2.5: فريق العمل وأدوارهم

تم توزيع المهام و المسؤوليات كتالي :

- جمع البيانات وتوسيمها: العضو الاول؛
- بناء النماذج و تقييم الأداء: العضو الأول؛
- تطوير المنصة و الواجهة الأمامية: العضو الثاني؛
- دمج النموذج داخل المنصة وربط الAPI: العضو الثاني؛
- الاختبار النهائي و التسليم: مشترك بين العضوين.

اعتمد الفريق على التنسيق المستمر والاجتماعات الدورية لمتابعة تقدم المشروع، مع استخدام أدوات رقمية لتبادل الملفات و تنظيم المهام و ضمان تكامل العمل بين الجانب البرمجي والجانب الذكي للمشروع.

4.2.5 أهداف المشروع

يُعد تحديد الأهداف خطوة محورية لرسم خارطة الطريق وضمان كفاءة التنفيذ. ولإخراج هذه الأهداف من سياقها النظري التخميني إلى إطار عملي قابل للتطبيق. يهدف هذا المشروع، باعتباره منصة رقمية تعتمد على نموذج البرمجيات كخدمة (SaaS)، إلى تحقيق استدامة اقتصادية تدريجية من خلال تطوير نموذج أعمال قابل للتوسع، يرتكز على تحويل التحليل التقني إلى قيمة تجارية موجهة للمؤسسات. ويتم تحديد الأهداف الاقتصادية وفق مراحل زمنية متدرجة كما يلي:

1. **على المدى القريب (6 أشهر) :**

- تحقيق أول تجربة تجارية فعلية عبر تقديم خدمة تحليل بيانات موجهة لعميل أو مؤسسة محلية؛
- اختبار قابلية تحويل نتائج النموذج إلى منتج قابل للبيع ؛
- تقييم الطلب الأولي في السوق الجزائري على خدمات تحليل المشاعر الرقمية.

2. **على المدى المتوسط (سنة إلى سنتين):**

- توسيع المنصة لتشمل منصات تواصل اجتماعي أخرى مثل Facebook وInstagram؛
- العمل على بناء قاعدة عملاء تضم وكالات التسويق الرقمي والمؤسسات التجارية، والعمل على تعزيز حضور المشروع داخل سوق أدوات التسويق الرقمي في الجزائر؛
- اعتماد نموذج اشتراك (SaaS Subscription Model) كآلية رئيسية لتحقيق الإيرادات؛
- تحقيق استقرار مالي أولي يسمح بتغطية تكاليف التشغيل والتطوير؛
- تحسين القيمة التجارية للمنصة من خلال تطوير الأداء ودقة التحليل.

3. **على المدى البعيد (أكثر من سنتين):**

- توسيع قاعدة العملاء لتشمل أسواقاً إقليمية ذات خصائص لغوية مشابهة؛

- تعزيز مكانة المنصة كحل رقمي في مجال تحليل البيانات التسويقية؛
- تطوير نموذج أعمال مستدام يعتمد على التوسع التدريجي وتحسين الخدمات؛
- رفع القيمة السوقية للمشروع تدريجياً من خلال النمو في عدد المستخدمين والعملاء.

5.2.5 الجدول الزمني لتحقيق المشروع

يوضح الجدول 3.5المراحل الزمنية الأساسية (**بالاشهر**)لإنجاز المشروع، بدايةً من جمع البيانات ومعالجتها إلى تطوير المنصة واختبارها ثم إعداد التقرير النهائي.

المرحلة	المهام	1	2	3	4	5	6
1	الدراسة الأولية واختيار الفكرة	✓	✓				
2	جمع بيانات Dataset (YouTube)		✓	✓			
3	معالجة النصوص (Preprocessing + Stop Words)			✓	✓		
4	تدريب نموذج MARBERT+ Fine-Tuning				✓	✓	
5	تطوير Backend				✓	✓	
6	تصميم Dashboard (واجهة المستخدم)					✓	✓
7	اختبار النظام وتحسين الأداء						✓
8	إطلاق النسخة التجريبية (MVP)						✓

جدول 3.5: الجدول الزمني لتحقيق المشروع

3.5 المحور الثاني :الجوانب الابتكارية

1.3.5 طبيعة الابتكارات

بالاستناد إلى مصفوفة الابتكار المعتمدة في المشروع، وفي سياق التحول الرقمي القائم على استغلال البيانات الضخمة في المجال التسويقي، يمكن تصنيف مشروع المنصة ضمن الابتكار التكنولوجي (Technological Innovation) المدمج مع ابتكار السوق (Market Innovation).

لا يقوم المشروع على ابتكار جذري عالمي، وإنما على تكييف تقنيات الذكاء الاصطناعي، خاصة التعلم العميق (Deep Learning) ومعالجة اللغة الطبيعية (NLP)، مع خصوصية السوق المحلي الجزائري. ويتمثل جوهر الابتكار في معالجة التحدي المرتبط بـ"عدم التجانس اللغوي" في البيانات النصية، حيث يتم تحويل التعليقات غير المنظمة على منصات التواصل الاجتماعي، والمكتوبة أساساً بالدارجة الجزائرية، إلى مؤشرات تحليلية دقيقة تدعم القرار التسويقي.

وعليه، تكمن القيمة الابتكارية للمشروع في تطويع تقنيات تحليل النصوص الحديثة لفهم السياق المحلي غير المدعوم بشكل كافٍ في الحلول العالمية الحالية، مما يتيح استخراج رؤى استراتيجية دقيقة حول سلوك المستهلك الجزائري.

2.3.5 مجالات الابتكارات

يتجسد الابتكار في المنصة عبر ثلاثة مجالات رئيسية:

1.2.3.5 ابتكار العمليات (Process Innovation)

يتمثل في إعادة هندسة طريقة تحليل السوق الرقمي، من أسلوب تقليدي يعتمد على المعالجة اليدوية للتعليقات وما يرافقه من بطء وأخطاء بشرية، إلى نظام آلي يعتمد على المعالجة الفورية (Real-time Processing)، يسمح بسحب وتصنيف آلاف التعليقات بشكل آني ودقيق، مما يرفع كفاءة تحليل البيانات بشكل كبير

2.2.3.5 ابتكار الخدمات والخصائص (Service & Feature Innovation)

تقدم المنصة خدمة تحليلية رقمية من نوع SaaS موجهة لفئة الأعمال (B2B)، تعتمد على لوحات تحكم تفاعلية (Interactive Dashboards) وتحليل المشاعر. وتميزها الأساسي يكمن في قدرتها على فهم وتحليل الدارجة الجزائرية، على

عكس الأدوات العالمية التي غالباً ما تفشل في تفسير السياق المحلي أو تصنيفه بشكل غير دقيق، مما يمنح المنصة قيمة مضافة في السوق الجزائري.

3.2.3.5 ابتكار نموذج العمل (Business Model Innovation)

يعتمد المشروع نموذجاً اقتصادياً قائماً على الاشتراك (Subscription-Based Model)، يسمح بتحقيق إيرادات دورية قابلة للتوسع. كما يتيح هذا النموذج مرونة في تلبية احتياجات مختلف الفئات من المؤسسات، من الشركات الناشئة إلى المؤسسات المتوسطة والكبيرة، عبر باقات خدمات تختلف حسب حجم البيانات ودرجة التحليل المطلوبة.

4.5 المحور الثالث : التحليل الاستراتيجي للسوق

1.4.5 عرض القطاع السوقي

1.1.4.5 السوق المحتمل

تنشط شركتنا الناشئة في سوق تكنولوجيا التسويق الرقمي والذكاء الاصطناعي، وهو سوق يتميز بالمرونة العالية والقدرة السريعة على التوسع البرمجي . على الرغم من أن انطلاقتنا التجارية ستكون مركزة، إلا أن سوقنا المحتمل يمتد أفقياً وعمودياً ليشمل:

- **استهداف قطاعات اقتصادية إضافية :** المنصة مصممة برمجياً لخدمة أي قطاع يعتمد على رضا المستهلك والحملات الإعلانية الموجهة للجمهور الجزائري، مثل: شركات التجارة الإلكترونية الشاملة (E-commerce)، قطاع الصناعات الغذائية والمشروبات، شركات الاتصالات، والمؤسسات الخدمية والمصرفية؛
- **التوسع نحو منصات رقمية متعددة (Multi-Platform Integration):** لن يقتصر النموذج التقني لشركتنا على سحب وتحليل تعليقات منصة "يوتيوب" فقط، بل نمتلك خارطة طريق برمجية لربط الواجهات (APIs) الحية لسحب وتنظيف البيانات اللحظية بالدارجة عبر المنصات الاخرى ؛
- **محددات وحجم السوق المحتمل:** المستهدفون هم الشركات والمؤسسات الجزائرية التي تمتلك حضوراً رقمياً وتضخ ميزانيات إعلانية . الدافع الأساسي (المحفز) لشراء خدمتنا البرمجية هو *"عقلنة الإنفاق الإعلاني وحماية السمعة الرقمية"* عبر التواجد في جميع الفضاءات الرقمية التي يتواجد فيها المستهلك الجزائري.

2.1.4.5 السوق المستهدف

في مرحلة الدخول الأولي إلى السوق ، حددت شركتنا الناشئة قطاع مستحضرات التجميل والعناية الشخصية (Cosmetics & Beauty) في الجزائر كسوق مستهدف متاح وقابل للاستهداف الفعلي. ويستند هذا التحديد إلى مجموعة من المؤشرات الاقتصادية والهيكلية والسلوكية التي تعكس جاذبية هذا القطاع وإمكاناته النموية.

أولاً: الحجم المالي للقطاعة: يُعد سوق مستحضرات التجميل والعناية الشخصية في الجزائر من بين القطاعات الاستهلاكية الأسرع نموًا، وهو ما يعكس حجم طلب مرتفع واستمرارية في الاستهلاك، مما يجعله سوقًا واعدًا للاستثمار والتوسع.

ثانيًا: البنية الهيكلية والتجارية للسوق : يتميز هذا القطاع بتنوع الفاعلين الاقتصاديين داخله، إذ يضم مصانع محلية ناشئة، وشركات استيراد وتوزيع كبرى، إضافة إلى علامات تجميل محلية صاعدة ومتاجر إلكترونية تعتمد بشكل كامل على التجارة الرقمية. هذا التنوع يعكس نضجًا تدريجيًا في هيكلة السوق وارتفاع مستوى المنافسة داخله.

ثالثًا: الديناميكية التسويقية وكثافة المحتوى الرقمي : يُعتبر قطاع مستحضرات التجميل من أكثر القطاعات اعتمادًا على التسويق الرقمي ووسائل التواصل الاجتماعي، حيث تعتمد الشركات بشكل كبير على التسويق عبر المؤثرين (Influencer Marketing) وصناعة المحتوى التفاعلي للترويج لمنتجاتها. ويؤدي هذا النشاط الرقمي المكثف إلى توليد كميات كبيرة من البيانات النصية والتعليقات اليومية، مما يجعل القطاع بيئة مناسبة للتحليل البرمجي ونماذج الذكاء الاصطناعي.

3.1.4.5 مبررات اختيار السوق المستهدف

تم بناء مبررات اختيار السوق المستهدف على أساس تحليل ديموغرافي وسلوكي لخصائص المستهلك الجزائري داخل قطاع مستحضرات التجميل والعناية الشخصية، مع التركيز على الفئة الأكثر تأثيرًا في قرارات الشراء والسلوك الرقمي.

- **أولاً: المرأة كمحرك رئيسي لقرارات الشراء في قطاع التجميل :** تشير المعطيات الحديثة إلى أن النساء يشكلن حوالي 42% من مستخدمي مواقع التواصل الاجتماعي في الجزائر، مقابل 58% للرجال [24]. كما تمثل هذه الفئة

نسبة كبيرة من التفاعل الرقمي داخل منصات التواصل، خاصة في المحتوى المرتبط بالبشرة والشعر ومستحضرات التجميل.

ويعكس هذا التواجد الرقمي دورًا مهمًا للمرأة في التأثير على قرارات الشراء داخل هذا القطاع، سواء بشكل مباشر عبر الاستهلاك أو بشكل غير مباشر عبر التوصيات والمراجعات.

- **ثانيًا: غزارة البيانات السلوكية باللغة الدارجة الجزائرية :** تتميز التعليقات والمراجعات الخاصة بالمستهلك الجزائري، وخاصة فئة النساء، بكونها غنية بالتجارب الشخصية التفصيلية حول المنتجات، سواء من حيث الرضا أو عدم الرضا. وغالبًا ما تُكتب هذه التعليقات باللغة الدارجة الجزائرية، بما تحمله من تعبيرات ثقافية وسياقية دقيقة.

هذا النوع من المحتوى يُصنّف كبيانات نصية غير مهيكلة (Unstructured Data)، إلا أن معالجته عبر النماذج التقليدية أو العالمية يبقى محدودًا بسبب الطابع اللغوي المحلي، مما يبرز الحاجة إلى حلول ذكية قادرة على فهم الدارجة الجزائرية وتحليلها لاستخراج الرأي الحقيقي للمستهلك.

- **ثالثًا: إمكانية إبرام عقود B2B وتطوير قنوات مبيعات مستدامة:** تعتمد الشركة على نموذج الأعمال (B2B SaaS)، مع استراتيجية لاكتساب العملاء عبر مسارين رئيسيين:

  - **عقود تجريبية مدفوعة (Pilot Contracts):** استهداف 3 إلى 5 علامات تجارية جزائرية ناشئة في قطاع التجميل، من خلال تقديم خدمات تحليل بيانات تسويقية مخصصة مقابل أسعار تحفيزية، بهدف بناء دراسات حالة (Case Studies) تثبت فعالية الحل.

  - **اتفاقيات شراكة إطارية (Framework Agreements):** التعاون مع وكالات التسويق الرقمي (Digital Marketing Agencies) التي تدير عدة علامات تجارية في نفس الوقت، مما يسمح بتوسيع قاعدة العملاء بشكل غير مباشر وضمان تدفقات إيرادات متعددة ومستقرة.

2.4.5 قياس شدة المنافسة

يتميز سوق تحليل البيانات والتسويق الرقمي في الجزائر ببنية تنافسية مزدوجة، حيث تتداخل المنافسة بين حلول برمجية عالمية من جهة، وأساليب تقليدية تعتمد على العمل البشري من جهة أخرى. وتكمن الميزة التنافسية للشركة الناشئة في قدرتها على سد الفجوة اللغوية والثقافية المرتبطة بالدارجة الجزائرية في تحليل البيانات النصية.

1.2.4.5 المنافسون المباشرون

تتمثل هذه الفئة في فرق إدارة المحتوى (*Community Managers*) و وكالات الدراسات التسويقية المحلية التي تعتمد على التحليل اليدوي التعليقات والآراء.

- **نقاط القوة:** قدرة بشرية طبيعية على فهم السياق المحلي والدارجة الجزائرية بدقة وسياق التهكم (Sarcasm).
- **نقاط الضعف:** بطء كبير في المعالجة، تكلفة تشغيل مرتفعة، وعدم القدرة على التعامل مع كميات ضخمة من البيانات في الزمن الحقيقي. فعلى سبيل المثال، تحليل 5,000 تعليق يدويًا قد يتطلب ما يقارب 40 ساعة عمل بشرية، مع احتمالية حدوث أخطاء ناتجة عن الإجهاد تتجاوز 20% إلى 25%، مما يحد من إمكانية اتخاذ قرارات تسويقية فورية.

2.2.4.5 المنافسون غير المباشرين

تشمل هذه الفئة الشركات العالمية المتخصصة في أدوات تحليل ومراقبة وسائل التواصل الاجتماعي (*Social Listening Tools*) مثل: *Brandwatch*، *Hootsuite* و *Mention*. تستحوذ هذه الأدوات على نسبة من استخدام الشركات الجزائرية، خصوصًا الكبرى منها، نظرًا لغياب بدائل محلية متخصصة.

- **نقاط القوة:** بنية تحتية تقنية متقدمة، قدرات تحليل واسعة، ودعم متعدد اللغات.
- **نقاط الضعف:** ضعف واضح في التعامل مع اللغة المحلية (الدارجة الجزائرية)؛ حيث يتم غالبًا تصنيف النصوص الدارجة على أنها غير دقيقة دلاليًا أو معالجتها ضمن نماذج اللغة العربية الفصحى، مما يؤدي إلى نتائج تحليل غير دقيقة للسياق المحلي ومضللة لأصحاب القرار.

3.4.5 الاستراتيجية التسويقية (Go-to-Market Strategy)

1.3.4.5 الاستراتيجية التنافسية والتسويقية

استناداً إلى نموذج الاستراتيجيات التنافسية العامة لـ Michael Porter، تعتمد المؤسسة الناشئة استراتيجية **القيادة بالتكلفة (Cost Leadership Strategy)**، والتي تهدف إلى تقديم خدمة تحليل المشاعر والآراء الرقمية للشركات بأقل تكلفة ممكنة مع الحفاظ على مستوى مقبول من الجودة والكفاءة.

ترتكز هذه الاستراتيجية على الاستفادة من تقنيات الحوسبة السحابية مفتوحة المصدر، وأتمتة عمليات جمع البيانات وتحليلها، والاعتماد على نماذج ذكاء اصطناعي مدربة مسبقاً، مما يسمح بتقليص تكاليف التشغيل وتقديم أسعار تنافسية مقارنة بالدراسات التسويقية التقليدية أو الحلول الأجنبية مرتفعة التكلفة.

2.3.4.5 المزيج التسويقي للخدمات (Marketing Mix – 7Ps)

المنتج (Product)

تتمثل الخدمة في منصة سحابية (SaaS) متخصصة في تحليل تعليقات وآراء المستهلكين المتعلقة بمنتجات التجميل والعناية الشخصية، حيث توفر تقارير تحليلية ومؤشرات أداء تساعد المؤسسات على فهم احتياجات العملاء واتجاهات السوق بسرعة وبتكلفة منخفضة مقارنة بوسائل الدراسات التسويقية التقليدية.

السعر (Price)

تعتمد المؤسسة سياسة تسعير اقتصادية وتنافسية تستهدف المؤسسات الصغيرة والمتوسطة، من خلال نظام اشتراك دوري منخفض التكلفة يسمح للعملاء بالاستفادة من خدمات التحليل دون الحاجة إلى استثمارات كبيرة في البنية التحتية أو فرق تحليل البيانات المتخصصة. كما يتم تقديم عروض تجريبية في مرحلة الإطلاق لتشجيع تبني الخدمة وزيادة الحصة السوقية.

التوزيع (Place)

يتم تقديم الخدمة بالكامل عبر الإنترنت من خلال منصة سحابية متاحة على مدار الساعة، مما يلغي تكاليف التوزيع التقليدية ويضمن وصول الخدمة إلى مختلف العملاء داخل الجزائر وخارجها دون قيود جغرافية.

الترويج (Promotion)

تعتمد المؤسسة على وسائل ترويج منخفضة التكلفة ومرتفعة الفعالية، تشمل التسويق الرقمي عبر شبكات التواصل الاجتماعي، والتسويق بالمحتوى من خلال نشر تقارير ودراسات دورية حول اتجاهات المستهلكين، إضافة إلى التواصل المباشر مع المؤسسات المستهدفة وعرض النسخ التجريبية للمنصة بهدف اكتساب العملاء بأقل تكلفة تسويقية ممكنة.

الأفراد (People)

يشمل هذا العنصر الفريق المسؤول عن تطوير المنصة وتقديم الدعم الفني والاستشارات للعملاء، حيث يساهم مستوى كفاءة العاملين وسرعة الاستجابة لاحتياجات العملاء في تعزيز رضاهم وبناء الثقة في الخدمة.

العمليات (Process)

تعتمد المنصة على عمليات رقمية مؤتمتة تشمل جمع البيانات وتحليلها وإعداد التقارير بشكل سريع ومنظم، مما يضمن تقديم الخدمة بكفاءة وجودة عالية وتقليل زمن الحصول على النتائج.

الدليل المادي (Physical Evidence)

رغم أن الخدمة رقمية، إلا أن العميل يلمس جودتها من خلال واجهة المنصة، التقارير التحليلية، لوحات المتابعة والإحصائيات المعروضة، إضافة إلى الهوية البصرية والموقع الإلكتروني للمؤسسة.

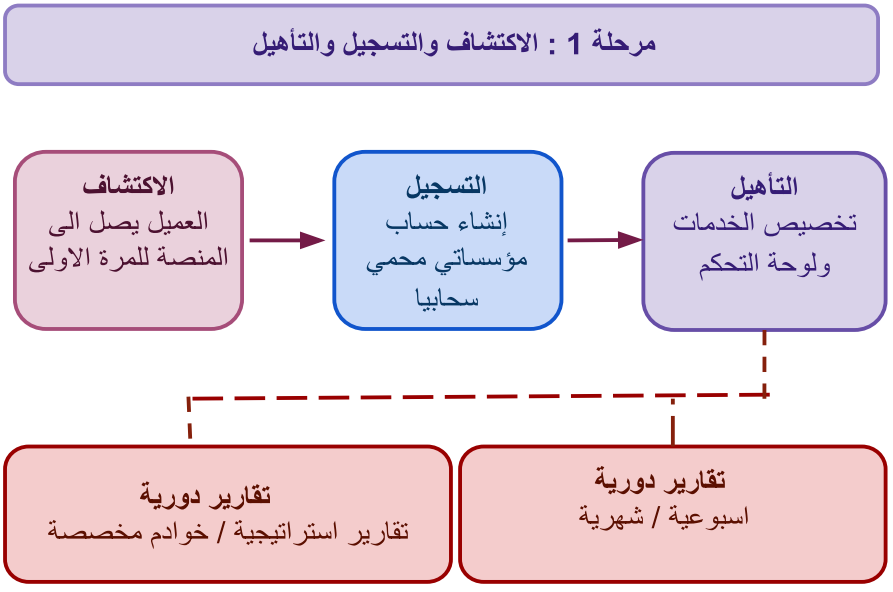
5.5 المحور الرابع : خطة الإنتاج والتنظيم

1.5.5 عملية الإنتاج (تطوير الخدمة)

بما أن المشروع يندرج ضمن قطاع المنصات الرقمية كمؤسسة ناشئة تعتمد على نموذج البرمجيات كخدمة (SaaS)، فإن مفهوم "خط الإنتاج" هنا يتحول من شكل التصنيع المادي التقليدي إلى "سلسلة تدفق برمجية ورحلة للمستخدم (User Journey & Service Pipeline)". تهدف هذه السلسلة إلى تحويل الطلبات الرقمية والبيانات اللغوية الخام إلى تقارير تحليلية ذات قيمة مضافة قابلة للاستخدام، وذلك عبر 6 مراحل أساسية متكاملة:

1.1.5.5 الاكتشاف والتسجيل والتأهيل

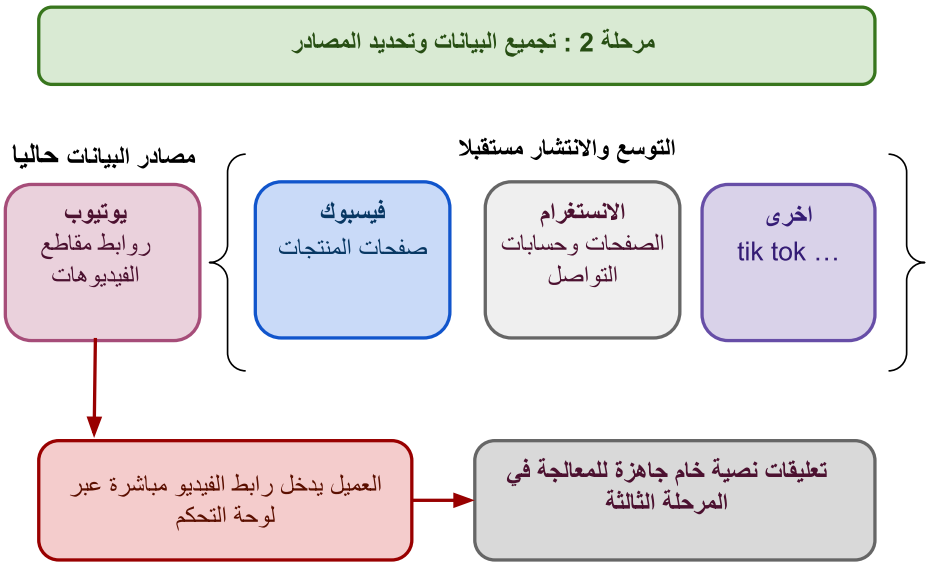
تبدأ دورة تقديم الخدمة عندما يقوم المستخدم (مثل إدارة التسويق لعلامة مستحضرات تجميل) بالولوج إلى المنصة لإنشاء حساب مؤسساتي محمي سحابياً. في هذه المرحلة، يتم تحديد وتخصيص نوع واجهة التحكم (Dashboard) والخدمات المطلوبة بدقة (مثل: مراقبة لحظية للسمعة الرقمية، أو استخراج تقارير تحليلية دورية).



مخطط 1.5 : مرحلة الاكتشاف والتسجيل والتأهيل

2.1.5.5 تجميع البيانات وتحديد مصادر الطلب

يقوم العميل بتحديد قنوات البيانات المراد تحليلها؛ حيث يتم ربط المنصة بصفحات ومقاطع الفيديو المستهدفة (مثل روابط قنوات يوتيوب، أو حسابات التواصل الاجتماعي). يتم هذا النشاط إما يدوياً عبر واجهة المستخدم، أو برمجياً بشكل آلي ومستمر عبر بروتوكولات الربط (APIs) لسحب التعليقات بشكل لحظي.

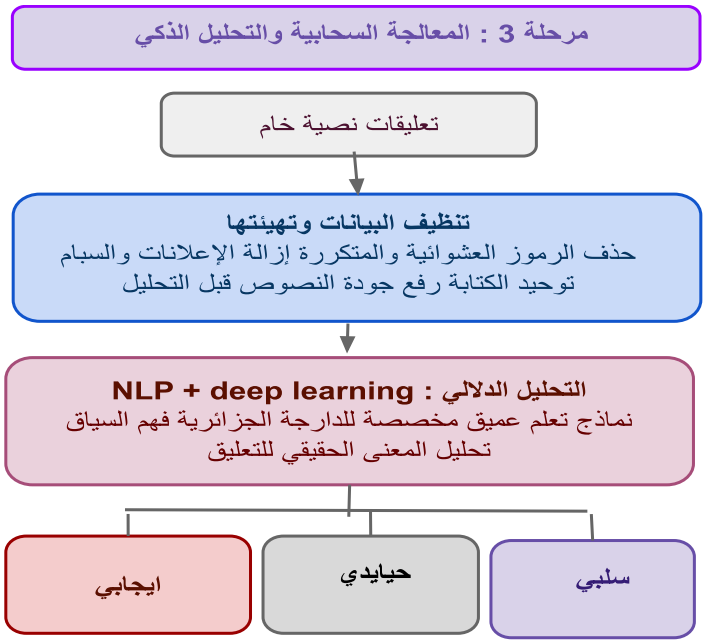


مخطط 2.5 : مرحلة تجميع البيانات وتحديد مصادر الطلب

3.1.5.5 المعالجة السحابية والتحليل الذكي

تنتقل البيانات النصية المسحوبة إلى خوارزميات المعالجة السحابية التابعة للمنصة، حيث تخضع لعمليتين أساسيتين:

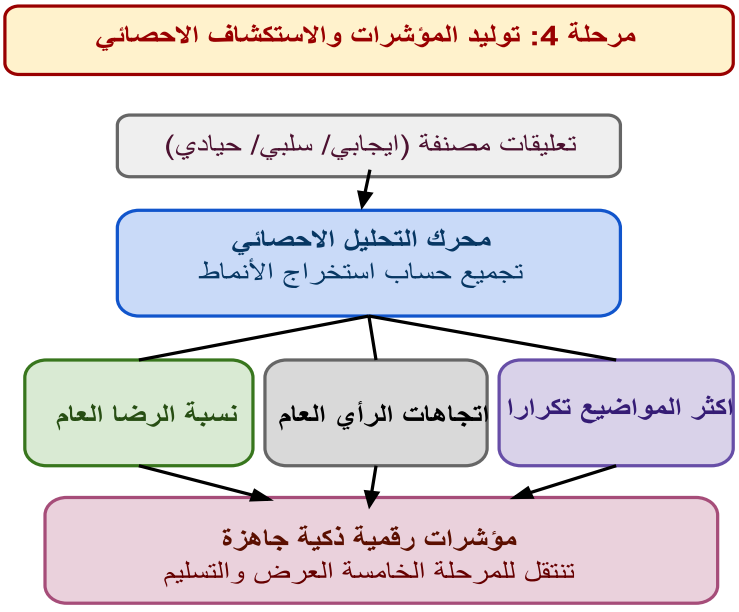
- التنظيف والتهيئة (Data Cleaning): تصفية النصوص من الرموز العشوائية المكررة والإعلانات لرفع جودة البيانات؛
- التحليل الدلالي (NLP Modeling): تشغيل نماذج تعلم عميق (Deep Learning) مخصصة لفهم الدارجة الجزائرية وتصنيف التعليقات والآراء بدقة سياقية إلى (إيجابي، سلبي، محايد).



مخطط 3.5 : مرحلة المعالجة السحابية والتحليل الذكي

4.1.5.5 توليد المؤشرات والاستكشاف الإحصائي

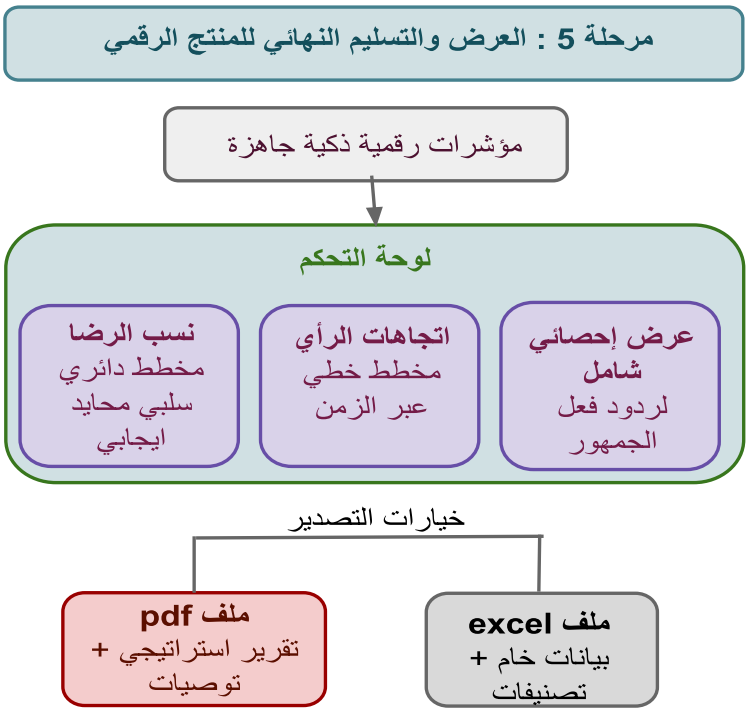
بعد انتهاء المعالجة الخوارزمية، يتم إنتاج القيمة التجارية عبر تحويل النصوص المصنفة إلى مخرجات رقمية ذكية تشمل: حساب نسب الرضا العام للزبائن، استخراج اتجاهات الرأي العام (Trends) حول المنتجات، وتحديد الكلمات المفتاحية والشكاوى الأكثر تكراراً بشكل إحصائي دقيق.



مخطط 4.5 : مرحلة توليد المؤشرات والاستكشاف الاحصائي

5.1.5.5 العرض والتسليم النهائي للمنتج الرقمي

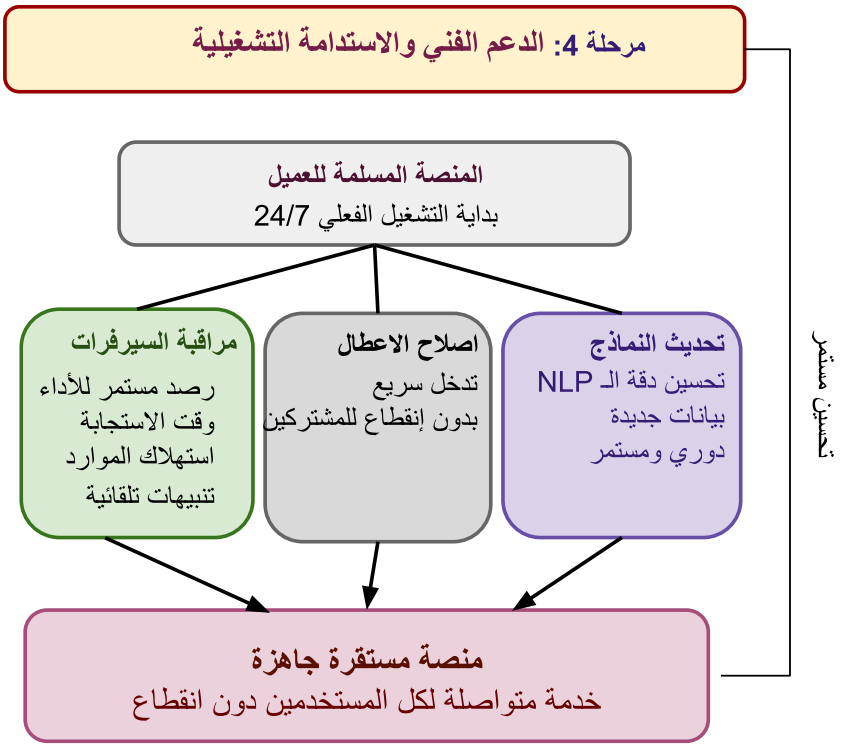
تُعرض المؤشرات المستخلصة داخل لوحة تحكم تفاعلية وحية (Interactive Dashboard) صُممت خصيصاً لتمنح العميل رؤية بصرية واضحة وسريعة حول سمعة علامته الرقمية. تتيح هذه المرحلة للعميل القدرة الفورية على تحميل وتصدير تقارير استراتيجية جاهزة بصيغ متعددة (PDF / Excel) تتضمن توصيات تسويقية مباشرة.



مخطط 5.5 : مرحلة العرض والتسليم النهائي للمنتج الرقمي

6.1.5.5 الدعم الفني والاستدامة التشغيلية :

لا تنتهي عملية تقديم الخدمة بالتسليم، بل تستمر لضمان جاهزية المنصة واستقرارها 24/7 عبر شبكة سحابية مرنة. يتضمن ذلك المراقبة المستمرة لأداء السيرفرات، التدخل السريع لإصلاح الأعطال البرمجية دون انقطاع الخدمة عن بقية المشتركين، والتحديث الدوري لنماذج معالجة اللغة الطبيعية (Model Update) لضمان تحسين دقة التصنيف بشكل مستمر.



مخطط 6.5 : مرحلة الدعم الفني والاستدامة التشغيلية

2.5.5 اليد العاملة

يتكون فريق عمل المشروع من 4 أفراد يتوزعون على المهام الإدارية، القانونية، والتقنية كالتالي:

المُسيّر

- **العدد:** 1
- **المهام والمسؤوليات:**
  - الممثل القانوني والشرعي للمشروع أمام جميع الهيئات الرسمية والشركاء؛
  - إدارة الشؤون المالية، الحساب البنكي، والالتزامات الضريبية وشؤون التأمينات؛
  - توقيع العقود والاتفاقيات الرسمية ومتابعة الأمور الإدارية اليومية للمؤسسة.

المُشرف

- **العدد: 1**
- **المهام والمسؤوليات:**
  - التخطيط التنفيذي للمشروع، ومتابعة سير العمل اليومي لفريق البرمجة؛
  - التنسيق بين متطلبات الإدارة والحلول التقنية لضمان تحقيق أهداف المشروع؛
  - مراقبة جودة المنتج النهائي، وضع الخطط التطويرية، وحل المشكلات التشغيلية.

مطور واجهات المستخدم و البرمجة الخلفية

- **العدد:** 1
- **المهام والمسؤوليات:**
  - برمجة وتطوير الواجهات الخارجية للموقع أو التطبيق التي يتفاعل معها المستخدم مباشرة؛
  - تحويل التصاميم إلى صفحات تفاعلية سريعة، مرنة، ومتوافقة مع جميع الشاشات والهواتف؛
  - ربط الواجهة الأمامية بالنظام الخلفي وقواعد البيانات عبر الـ (APIs) وإدارة الخوادم (Servers).
  - بناء الهيكل البرمجي الأساسي للمشروع، وحماية النظام من الاختراقات والثغرات الأمنية؛
  - تصميم وتسيير قواعد البيانات لضمان معالجة البيانات وحفظها بسرعة وأمان؛

مهندس متخصص في الذكاء الاصطناعي ومعالجة اللغات الطبيعية

- **العدد:** 1
- **المهام والمسؤوليات:**
  - تطوير وتحسين نماذج الذكاء الاصطناعي ومعالجة اللغات الطبيعية المستخدمة في تحليل التعليقات والنصوص.
  - الإشراف على تدريب النماذج وتقييم أدائها لضمان دقة وجودة نتائج التحليل.

3.5.5 الشراكات الرئيسية

يعتمد المشروع على مجموعة من الشركاء لدعم التطوير والتشغيل:

- حاضنة الأعمال الجامعية المسيلة : الإشراف والدعم التقني؛
- الصندوق الوطني لدعم المؤسسات الناشئة؛
- وكالات التسويق الرقمي: اختبار المنصة (Beta Testing)؛
- مزودو الخدمات السحابية: الاستضافة والحوسبة؛
- مجتمعات Open Source: دعم تقني ومكتبات جاهزة؛
- مؤسسات دعم ريادة الأعمال: التمويل والمرافقة.

6.5 المحور الخامس: الخطة المالية

1.6.5 التكاليف الاستثمارية:

تمثل التكاليف الاستثمارية الأموال الموجهة لاقتناء الموارد والتجهيزات اللازمة لإنشاء المشروع وتجهيزه قبل بدء مرحلة التشغيل الفعلي.

التكاليف الاستثمارية العينية المتمثلة في الأصول المادية والتجهيزات الضرورية لإنجاز المشروع وضمان حسن سيره، و التكاليف الاستثمارية المعنوية في النفقات المرتبطة باقتناء أو تطوير الأصول غير المادية اللازمة لتشغيل المشروع.. ويوضح الجدول التالي تفاصيل التكاليف:

نوع الاستثمار	اسم الاستثمار	العدد	تكلفة الوحدة	التكلفة الاجمالية
معنوي	منصة	1	200 000,00	200 000,00
المجموع 01		1	200 000,00	200 000,00
عيني	الحواسيب ولوازمها	4	100 000,00	400 000,00
	تجهيزات المكتبية	4	60 000,00	240 000,00
	الطابعة	**1**	30 000,00	30 000,00
المجموع 02	**المجموع**	9	190 000,00	670 000,00
المجموع 01+02		**10**	**390 000,00**	**870 000,00**

جدول 4.5 : تكاليف استثمارية

2.6.5 التكاليف التشغيلية

تشمل التكاليف التشغيلية جميع النفقات الدورية اللازمة لتشغيل المنصة واستمرار نشاطها، حيث تُنفق بشكل منتظم لضمان تقديم الخدمات بكفاءة واستمرارية. وتُعرض تفاصيل هذه التكاليف في الجداول التالية.

أ- المشتريات

تمثل المشتريات مختلف الموارد والخدمات الضرورية لضمان سير المشروع و تشغيله بكفاءة. وتشمل هذه المشتريات خدمات الاستضافة السحابية، وخدمات الذكاء الاصطناعي، والبرمجيات المكتبية، بالإضافة إلى اللوازم المكتبية اللازمة لمتابعة الأنشطة الإدارية والتشغيلية للمشروع. ويوضح الجدول التالي تفاصيل هذه المشتريات وتكاليفها.

التعيين	العدد/الكمية	تكلفة الوحدة	التكلفة الاجمالية
Cloud Hosting & VPS	12	8 900,00	106 800,00
Gemini API Services	12	2 600,00	31 200,00
Microsoft 365 Original	4	17 000,00	68 000,00
لوازم المكتبية			20 000,00
المجموع		**28 500,00**	**226 000,00**

جدول 5.5 : جدول المشتريات

ب- أجور العمال

تُعد أجور العمال من أهم التكاليف التشغيلية للمشروع، حيث تمثل المقابل المالي للموارد البشرية المكلفة بتسيير المنصة وتطويرها وصيانتها. وقد تم تحديد الأجور وفقًا لطبيعة المهام والمسؤوليات الموكلة لكل منصب، بما يضمن حسن سير المشروع واستمرارية نشاطه. ويوضح الجدول التالي توزيع الأجور الشهرية والسنوية للعاملين في المشروع.

الوظيفة	العدد	الراتب الشهري	إجمالي الرواتب الشهرية	إجمالي الرواتب السنوية
المسيير المؤسسة	1	60 000,00	60 000,00	720 000,00
مشرف المنصة	1	60 000,00	60 000,00	720 000,00
مبرمج الواجهة الخلفية والامامية	1	55 000,00	55 000,00	660 000,00
مهندس متخصص في الذكاء الاصطناعي ومعالجة اللغات الطبيعية	1	55 000,00	55 000,00	660 000,00
المجموع	**4**	**230 000,00**	**230 000,00**	**2 760 000,00**

جدول 6.5 : جدول أجور العمال

ج- أعباء خارجية

تتمثل الأعباء الخارجية في مختلف النفقات التي يدفعها المشروع مقابل خدمات وموارد يتم الحصول عليها من جهات خارجية كما هو موضح في الجدول التالي.

التعيين	الشهرية	السنوية
الإيجار	10 000,00	120 000,00
كهرباء/غاز/ماء	8 000,00	96 000,00
مصاريف التسويق	5 000,00	60 000,00
مصاريف البريد والاتصالات	2 600,00	31 200,00
مصاريف الصيانة	1 000,00	12 000,00
اشتراكات	5 000,00	60 000,00
أجور الوسطاء والأتعاب	8 000,00	96 000,00
المجموع		**475 200,00**

جدول 7.5 : الأعباء الخارجية

3.6.5.تقدير رأس المال و الإيرادات

أ- تقدير رأس المال

يوضح الجدول 22 تقدير رأس المال اللازم لإطلاق المشروع بإجمالي 2,226,106.67 دج. يغطي هذا المبلغ التكاليف الاستثمارية لشراء العتاد والمكاتب (870,000.00 دج)، ورأس المال العامل لتأمين المصاريف التشغيلية لمدة 4 أشهر الأولى (1,153,733.33 دج)، بالإضافة إلى هامش أمان بنسبة 10% لحماية الخزينة وضمان استقرار انطلاقة الشركة الناشئة.

التكاليف الاستثمارية	870 000,00		
التكاليف التشغيلية السنوية	3 461 200,00		
التكاليف التشغيلية الشهرية	288 433,33		
المدة بالأشهر	4	رأس المال العامل	1 153 733,33
نسبة هامش الأمان	10%	هامش أمان	202 373,33
		رأس المال المطلوب	2 226 106,67

جدول 8.5 : تقدير رأس المال

ب- تقدير الايرادات :

تتمثل الإيرادات المداخيل المتوقعة التي يحققها المشروع من خلال الخدمات التي تقدمها المنصة، والمتمثلة في بيع التقارير التحليلية والاشتراكات بمختلف أنواعها. وقد تم تقدير الإيرادات لثلاث سنوات اعتمادًا على عدد الخدمات المتوقع بيعها وأسعارها، بهدف تقييم القدرة المالية للمشروع ومدى تحقيقه للعائدات المستقبلية.

N+4	N+3	N+2	N+1	N	
249	207	173	144	120	**التقارير**
60000,00	60000,00	60000,00	60000,00	60000,00	**Prix HT produit A**
14929920,00	12441600,00	10368000,00	8640000,00	300000,00	**CHIFFRE D'AFFAIRES GLOBAL**
104	86	72	60	50	**الخطة الرئيسية**
120000,00	120000,00	120000,00	120000,00	120000,00	**Prix HT produit B**

CHIFFRE D'AFFAIRES GLOBAL	6000000,00	7200000,00	8640000,00	10368000,00	12441600,00
خطة الاعمال	10	12	14	17	21
Prix HT produit C	30000,00	30000,00	30000,00	30000,00	30000,00
CHIFFRE D'AFFAIRES GLOBAL	300000,00	360000,00	432000,00	518400,00	622080,00
خطة الاحترافية والشركات	16	19	23	28	33
Prix HT produit D	100000,00	100000,00	100000,00	100000,00	100000,00
CHIFFRE D'AFFAIRES GLOBAL	1600000,00	1920000,00	2304000,00	2764800,00	3317760,00
CHIFFRE D'AFFAIRES GLOBAL	8200000,00	18120000,00	21744000,00	26092800,00	31311360,00

جدول 9.5 : جدول الإيرادات

7.5 المحور السادس : النموذج الأولي التجريبي

يمثل النموذج الأولي التجريبي المرحلة العملية الأولى من هذا المشروع، حيث تم تطوير نماذج تحليل مشاعر وظيفية تعتمد على تقنيات النماذج العميقة (MARBERT) لتصنيف التعليقات المكتوبة بالدارجة الجزائرية إلى ثلاث فئات: إيجابي، سلبي، ومحايد.

وقد أظهرت النماذج المطورة أداءً تجريبياً يبلغ تقريباً 79% من الدقة على البيانات المستخدمة في التدريب والتقييم، مما يعكس قابلية الحل للتطبيق في بيئة واقعية.

يمكن اختبار النموذج بشكل مباشر من خلال المنصة المطوّرة، حيث يتم إدخال رابط فيديو من منصة YouTube، ثم استخراج التعليقات المرتبطة به تلقائيًا وتحليلها باستخدام نموذج تحليل المشاعر، مع عرض النتائج بشكل فوري داخل النظام.

كما تم إعداد تصور أولي لواجهة المنصة (Prototype)، مع دمج نموذج تحليل المشاعر داخل تطبيق ويب كامل، حيث أصبح النظام يعمل بشكل متكامل ويعرض نتائج التحليل مباشرة للمستخدم النهائي.

يمثل هذا النموذج مرحلة انتقالية بين البحث التطبيقي والتطبيق العملي، ويهدف إلى التحقق من قابلية الحل قبل إطلاق النسخة النهائية.

8.5 الخاتمة

يتضح من خلال التحليل الشامل المقدَّم في هذا الفصل أن المشروع يرتكز على ثلاث ركائز أساسية تمنحه قابلية للتطبيق العملي.

- أولاً، على الصعيد التقني، تم بناء نماذج تحليل مشاعر مخصصة للدارجة الجزائرية بلغت دقتها حوالي 79%، مما يعكس نموذجاً أولياً قابلاً للتطوير والتحسين؛
- ثانياً، على الصعيد السوقي، كشف التحليل الاستراتيجي عن وجود فجوة واضحة في السوق الجزائري، حيث تعجز الحلول العالمية الحالية عن التعامل مع المحتوى غير الرسمي المكتوب بالدارجة، مما يمنح المشروع ميزة تنافسية يصعب تكرارها على المدى القريب؛
- ثالثاً، على الصعيد المالي، تشير التوقعات المبنية على السيناريو الواقعي إلى إمكانية بلوغ نقطة التعادل خلال السنة الأولى، وتحقيق ربحية ابتداءً من السنة الثانية، مع حاجة تمويلية أولية يمكن تغطيتها عبر برامج دعم الشركات الناشئة الوطنية.

وفي المحصلة، لا يمثّل هذا المشروع مجرد مذكرة تخرج، بل يشكّل نواة مشروع ريادي واعد يجمع بين الجانب البحثي في معالجة اللغة الطبيعية والرؤية التجارية المبنية على حاجة سوقية حقيقية غير مُلبّاة.

الخاتمة العامة

خلصت هذه الدراسة إلى تقديم إجابة عملية وموثّقة عن إشكاليتها المحورية، المتعلقة بمدى قدرة نماذج التعلم الآلي والتعلم العميق ونماذج المحوّلات الحديثة على تحليل مشاعر النصوص المكتوبة باللهجة الجزائرية الدارجة، وكيفية توظيف هذه النتائج في بناء منصة ذكاء تسويقي.

وقد أثبتت التجارب المُجراة على بيانات حقيقية مُستخرَجة من منصة يوتيوب أن التفاوت في الأداء بين المقاربات الثلاث المختبَرة كان واضحاً ودالاً. فقد سجّلت نماذج التعلم الآلي الكلاسيكية، ولا سيما SVM مقترناً بـTF-IDF، أداءً مرجعياً متيناً بدقة بلغت 77%، متفوقاً في جوانب عدة على نماذج التعلم العميق المعتمدة على FastText. في المقابل، جاء نموذج MARBERT في مقدمة جميع النماذج المختبَرة بدقة 79% وF1-Macro قدره 0.72، مؤكداً القيمة المضافة التي تمنحها النماذج المدرَّبة مسبقاً على اللغة العربية في سياقات اللهجات المحلية، خاصةً في تصنيف الفئات الأقل تمثيلاً كالتعليقات السلبية والإيجابية.

كما كشفت الدراسة أن التحديات اللغوية الخاصة بالدارجة الجزائرية من غياب التوحيد الإملائي، وظاهرة المزج اللغوي، وشُح البيانات المصنّفة تمثّل عائقاً حقيقياً يستلزم معالجات مسبقة دقيقة ومخصصة، لا يمكن الاستعاضة عنها بالأدوات العامة المصممة للغة العربية الفصحى. وقد تجلّى هذا في الأثر الإيجابي لخطوات التطبيع النصي ومعالجة النفي والرموز التعبيرية على جودة التصنيف.

وعلى الصعيد التطبيقي، قدّمت الدراسة تصوراً أولياً لمنصة ذكاء تسويقي موجّهة للمؤسسات الجزائرية، تستثمر النتائج التقنية المحققة في تقديم خدمة رصد وتحليل آلي لآراء المستهلكين على وسائل التواصل الاجتماعي. وقد أثبت التحليل الاستراتيجي والمالي المرفق أن هذه المنصة تحمل جدوى اقتصادية حقيقية، في ظل فراغ تنافسي واضح في السوق الجزائري، إذ تعجز الحلول العالمية المتاحة عن تلبية الحاجة المحلية الخاصة بمعالجة المحتوى غير الرسمي المكتوب بالدارجة.

وفي المحصلة، يكشف هذا العمل أن قيمته تتجاوز حدود المقارنة التقنية بين النماذج، لتمتد نحو طرح نهج منهجي متكامل يجمع بين صرامة البحث العلمي وواقعية التطبيق الميداني. ولعل أبرز ما خلصت إليه الدراسة هو أن بناء حلول فعّالة لتحليل اللغة العربية العامية لا يستقيم دون الاستثمار الجاد في بناء موارد لغوية محلية متخصصة، وهو ما يفتح آفاقاً واسعة أمام أبحاث مستقبلية تتمحور حول توسيع مجموعات البيانات الموسومة، وتطوير نماذج ضبط دقيق مخصصة للدارجة الجزائرية، واختبار أداء نماذج الجيل الجديد من المحوّلات متعددة اللغات في هذا السياق اللغوي الفريد.

المراجع

[1] D. Chaffey and F. Ellis-Chadwick, *Digital Marketing: Strategy, Implementation and Practice*, 7th ed. Harlow, UK: Pearson Education Limited, 2019.

[2] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[3] R. Mitchell, *Web Scraping with Python: Collecting More Data from the Modern Web*, 2nd ed. Sebastopol, CA: O'Reilly Media, 2018.

[4] Google Developers, "YouTube Data API v3 Overview," Google LLC. [Online]. Available: https://developers.google.com/youtube/v3. [Accessed: 2024].

[5] A. Markham and E. Buchanan, *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee*, Version 2.0. Association of Internet Researchers, 2012.

[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. (draft). Stanford University, 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[7] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA: Morgan & Claypool Publishers, 2012.

[8] N. Habash, *Introduction to Arabic Natural Language Processing*. San Rafael, CA: Morgan & Claypool Publishers, 2010.

[9] H. Samih and W. Maier, "Detecting Code-Switching in Social Media," in *Proc. of the SocialNLP Workshop at EMNLP*, 2016, pp. 9–15.

[10] F. Barbieri, F. Ronzano, and H. Saggion, "What Does This Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis," in *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, 2016, pp. 526–532.

[11] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2–3, pp. 146–162, 1954.

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.

[13] M. Abdul-Mageed et al., "A comparative analysis of word embedding and deep learning for Arabic sentiment classification," *Electronics*, vol. 12, no. 6, p. 1425, 2023.

[14] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, vol. 30.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[16] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguist.*, Online, Aug. 2021, pp. 7088–7105.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[19] Doccaio, "React + Vite: Why use it?," *DEV Community*, Dec. 15, 2023. [Online]. Available: https://dev.to/doccaio/react-vite-why-use-it-cg2. [Accessed: May 1, 2026].

[20] S. Ramírez, "FastAPI," *FastAPI Documentation*, 2018. [Online]. Available: https://fastapi.tiangolo.com/. [Accessed: May 1, 2026].

[21] OpenJS Foundation, "Node.js," *Node.js Official Website*. [Online]. Available: https://nodejs.org/. [Accessed: May 1, 2026].

[22] Google, "Firebase Documentation," *Google Firebase Official Platform*. [Online]. Available: https://firebase.google.com/docs. [Accessed: May 1, 2026].

[23] Microsoft, "Visual Studio Code Documentation," *Code.visualstudio.com*. [Online]. Available: https://code.visualstudio.com/docs. [Accessed: May 1, 2026].

[24] DataReportal, "Digital 2025: Algeria," Feb. 2025. [Online]. Available: https://datareportal.com/reports/digital-2025-algeria. [Accessed: Jun. 15, 2026].

الملحق الأول : الميزانية

N+4	N+3	N+2	N+1	N		
		ACTIF				
0,00	40000,00	80000,00	120000,00	160000,00	التثبيتات المعنوية / الأصول غير الملموسة	Immobilisation Incorporelles
						Immobilisation Corporelles
					تثبيتات عينية أخرى	Terrain
						Bâtiment
0,00	134000,00	268000,00	402000,00	536000,00		Autres Immobilisations Corporelles
						Immobilisations en concession
						Immobilisation en cours
						Immobilisations Financières
						Titres mis en équivalence
						Autres participations et créances rattachées
						Autres Titres immobilisés
						Prêts et autres titres financiers non courants
						Impôts différés actif
0,00	**174000,00**	**348000,00**	**522000,00**	**696000,00**		**ACTIF NON COURANT**
					المخزونات والمنتجات قيد التنفيذ	Stocks et encours
						Créances et emplois assimilés
						Clients

						Autres débiteurs
						Impôts et assimilés
						Autres créances et emplois assimilés
						Disponibilités et assimilés
						Placements et autres actifs financiers courants
88966465,07	**61358938,67**	**38891866,67**	**20708506,67**	**6094906,67**	**الخزينة**	**Trésorerie**
88966465,07	**61358938,67**	**38891866,67**	**20708506,67**	**6094906,67**		**ACTIF COURANT**
88966465,07	**61532938,67**	**39239866,67**	**21230506,67**	**6790906,67**	**مجموع الأصول**	**TOTAL ACTIF**

		PASSIF				
N+4	N+3	N+2	N+1	N		
						CAPITAUX PROPRES
2226106,67	2226106,67	2226106,67	2226106,67	2226106,67	رأس مال تم اصداره	Capital émis
						Capital non appelé
						Ecart de réévaluation
						Primes et réserves- Réserves Consolidées
27433526,40	22293072,00	18009360,00	14439600,00	4564800,00	النتيجة الصافية - حصة المجموعة	Résultat net- RN part du groupe
59306832,00	37013760,00	19004400,00	4564800,00		أموال خاصة أخرى - ترحيل من جديد	Autres capitaux propres- report à nouveau
						Part de la société consolidante (1)
88966465,07	**61532938,67**	**39239866,67**	**21230506,67**	**6790906,67**		**CAPITAUX PROPRES**

						PASSIFS NON-COURANTS
						Emprunts et dettes financières
						Impôt différé passif
						Autres dettes non courantes
						Provisions et produits constatés d'avance
0,00	**0,00**	**0,00**	**0,00**	**0,00**		**PASSIFS NON-COURANTS**
						PASSIFS COURNATS
						Fournisseurs et comptes rattachés
						Impôts
						Autres dettes
						Trésorerie passif
0,00	**0,00**	**0,00**	**0,00**	**0,00**		**PASSIFS COURANTS**
88966465,07	**61532938,67**	**39239866,67**	**21230506,67**	**6790906,67**	مجموع الخصوم	TOTAL PASSIF
0,00	**0,00**	**0,00**	**0,00**	**0,00**	التحقق من توازن الأصول / الخصوم	Vérification de l'équilibre Actif/Passif

الملحق الثاني: حساب النتائج

N+4	N+3	N+2	N+1	N		
31311360,00	2609280,00	217440,00	1812000,00	820000,00	رقم الأعمال	Vente et produits annexes
						Variation des stocks produits finis et en cours
						Production immobilisée
						Subvention d'exploitation
31311360,00	**2609280,00**	**217440,00**	**1812000,00**	**820000,00**	إنتاج السنة المالية	**Production de l'exercice**
46863,60	39052,80	32544,00	27120,00	22600,00	المشتريات المستهلكة	Achats consommés
47520,00	47520,00	47520,00	47520,00	47520,00	الخدمات الخارجية والاستهلاكات الأخرى	Services Extérieurs et autres consommations
94383,60	**86572,80**	**80064,00**	**74640,00**	**70120,00**	استهلاك السنة المالية	**Consommation de l'exercice**
30367526,40	**2522707,00**	**209433,60**	**1737360,00**	**749880,00**	القيمة المضافة للاستغلال	**Valeur ajoutée d'exploitation**
2760000,00	276000,00	276000,00	276000,00	276000,00	أعباء المستخدمين	Charges de personnel
						Impôts et taxes et versement assimilés
27607526,40	**2246707,00**	**181833,60**	**1461360,00**	**473880,00**	**الفائض الإجمالي عن الاستغلال**	**Excédent Brut d'Exploitation**
						Autres produits opérationnels
						Autres charges opérationnelles
174000,00	17400,00	17400,00	17400,00	17400,00	المخصصات للاهتلاكات والمؤونات	Dotations aux amortissements, Provisions
						Reprise sur pertes de valeurs et provisions
27433526,40	**2229307,00**	**180093,60**	**1443960,00**	**456480,00**	النتيجة العملياتية	**Résultat opérationnel**

						Produits Financiers
						Charges financières
						Résultat financier
0,00	**0,00**	**0,00**	**0,00**	**0,00**		**Résultat Ordinaire avant impôt**
						Impôt exigible sur résultat ordinaire
						Impôt différé sur résultat ordinaire
						Total des produits des activités ordinaires
						Total des charges des activités ordinaires
27433526,40	**22293072,00**	**18009360,00**	**14439600,00**	**4564800,00**		Résultat net des activités ordinaires
						Eléments extraordinaire (produits)
						Eléments extraordinaire (produits)
						Eléments extraordinaire (charges)
						Résultat extraordinaire
27433526,40	**22293072,00**	**18009360,00**	**14439600,00**	**4564800,00**	**النتيجة الصافية للسنة**	RESULTAT NET DE L'EXCERCICE

الملحق الثالث : نموذج العمل التجاري

نموذج العمل التجاري لمؤسسة
Technical Marketing Leader

الشركات الاساسية

تشمل الشراكات الرئيسية للمشروع مجموعة من الهيئات والمؤسسات التي تساهم في دعم وتطوير المنصة، والمتمثلة في:

- جامعة محمد بوضياف بالمسيلة.
- حاضنة الأعمال بجامعة المسيلة.
- الصندوق الوطني لدعم المؤسسات الناشئة.
- مزودو خدمات الحوسبة السحابية.
- وكالات التسويق الرقمي.
- مجتمعات البرمجيات مفتوحة المصدر.
- مؤسسات دعم ريادة الأعمال والابتكار.

الانشطة الرئيسية

تعمل المنصة من خلال:

- جمع البيانات من منصات التواصل الاجتماعي.
- معالجة وتنظيف البيانات النصية.
- تدريب وتحسين نماذج الذكاء الاصطناعي.
- تطوير المنصة الإلكترونية ولوحات التحكم.
- إعداد التقارير والتحليلات الإحصائية.
- صيانة المنصة وتحديثها دورياً.
- إطلاق المنصة وتسويق خدماتها وتعزيز حضورها لدى المؤسسات المستهدفة، بما يساهم في استقطاب العملاء وإبرام العقود والشراكات الاستراتيجية.

القنوات

يعتمد المشروع على عدة قنوات من أجل الوصول إلى العملاء والتعريف بالخدمات المقدمة، والمتمثلة في:

- منصة ويب إلكترونية تسمح للمؤسسات باستخدام خدمات التحليل بسهولة.
- منصة LinkedIn للتواصل مع الشركات والمؤسسات الاقتصادية.
- مواقع التواصل الاجتماعي للتسويق الرقمي ونشر نتائج التحليل والإحصائيات.
- المؤتمرات والفعاليات المتعلقة بالتكنولوجيا والتحول الرقمي.
- الشراكات الجامعية والاقتصادية من أجل توسيع نطاق المشروع.

القيم المقترحة

اختصار الوقت والجهد في اتخاذ القرارات التسويقية من خلال :

- تحليل مشاعر المستهلك الجزائري بدقة.
- دعم اتخاذ القرارات التسويقية المبنية على البيانات.
- تخفيض تكاليف الدراسات التسويقية التقليدية.
- توفير واجهة سهلة وسريعة الاستخدام.
- تقديم تقارير تشمل نقاط قوة وضعف واستراتيجيات تسويقية ومؤشرات إحصائية تساعد على فهم السوق.

العلاقات مع العملاء

يسعى المشروع إلى بناء علاقة قوية ومستدامة مع العملاء من خلال

- خلق علاقة مع العملاء من خلال الحملات التحسيسية عبر مواقع التواصل الاجتماعي، والتعريف بالمنصة بالتعاون مع الشركاء، وتقديم خدمات تجريبية مجانية.
- تطوير العلاقة مع العملاء من خلال تقديم خدمات إضافية مجانية، وخصومات، ومزايا خاصة للمشتركين.
- استدامة العلاقة مع العملاء من خلال بناء الثقة، وتوفير خدمات حسب الطلب، وتجديد الخدمات بصورة مستمرة، والمحافظة على التواصل الدائم معهم.

الموارد الرئيسية

تشمل الموارد الرئيسية التي يعتمد عليها المشروع مجموعة من الوسائل التقنية والبشرية الضرورية لضمان نجاح المنصة، والمتمثلة في:

الموارد المادية:

- الحواسيب، الخوادم، وسائل الاتصال.

الموارد المالية:

- رأس المال الأولي ومصادر التمويل.

الموارد البشرية:

- مطورو الويب.
- مختصو الذكاء الاصطناعي ومعالجة اللغة الطبيعية.
- مسير الشركة .
- المشرف التقني.

الموارد الفكرية:

- منصة تحليل المشاعر.
- نموذج MARBERT المدرب.
- قاعدة البيانات الخاصة بالدارجة الجزائرية.
- واجهات API والهوية الرقمية للمشروع.

شرائح العملاء.

يستهدف المشروع عدة فئات من المؤسسات والعملاء المهتمين بتحليل سلوك المستهلك، والمتمثلة في:

العملاء المستهدفون:

- شركات مستحضرات التجميل والعناية الشخصية.
- شركات التجارة الإلكترونية.
- وكالات التسويق الرقمي.
- المؤسسات الناشئة.
- شركات المنتجات الاستهلاكية.

العملاء المحتملون:

- شركات الصناعات الغذائية والمشروبات.
- شركات الاتصالات.
- المؤسسات الخدمية والمصرفية.
- الباحثون والمهتمون بتحليل البيانات.
- القنوات والوسائط الإعلامية التي تسعى إلى معرفة آراء زبائنها وعملائها ومتابعيها وتقييماتهم.

هيكل التكاليف

تشمل التكاليف الأساسية المرتبطة بالمشروع مجموعة من المصاريف الضرورية لضمان استمرارية العمل وتطوير المنصة، والمتمثلة في:

- تكاليف جمع ومعالجة البيانات.
- تكاليف تدريب النماذج وتطويرها.
- تكاليف تطوير وصيانة المنصة.
- تكاليف الحوسبة السحابية والاستضافة.
- تكاليف واجهات البرمجة والأدوات التقنية.
- تكاليف التسويق واستقطاب العملاء.
- تكاليف الدعم الفني والتطوير المستمر.

مصادر الإيرادات

- يعتمد المشروع على عدة مصادر لتحقيق الإيرادات وضمان الاستمرارية المالية، والمتمثلة في:
- إيرادات الاشتراكات الشهرية (الخطة الاساسية، الخطة الاعمال , خطة الاحترافية والشركات).
- الإيرادات الناتجة عن بيع خدمات تحليل البيانات.
- الإيرادات الناتجة عن التقارير التحليلية حسب الطلب.
- الإيرادات الناتجة عن الدراسات السوقية ومخططات البيانية .
- الإيرادات الناتجة عن الخدمات الاستشارية في التسويق وتحليل البيانات وتقديم روى استراتيجية .