RAPPORT DE PROJET - Analyse de Sentiments

Auteur : Seynabou Julienne Venance
Date : Décembre 2025

1. INTRODUCTION ET OBJECTIFS

Ce projet vise à développer un système complet d'analyse de sentiments basé sur l'apprentissage profond, capable de classifier des avis clients selon 5 niveaux de sentiment (très négatif à très positif).

Objectifs principaux :

- Construire un modèle de classification performant
- Développer une API REST pour l'inférence
- Créer une application web interactive
- Déployer la solution en production

2. DATASET ET EXPLORATION

2.1 Description des données

Sources : Fichiers Parquet optimisés

- Dataset d'entraînement : 650 000 avis clients
- Dataset de test : 50 000 avis clients
- Structure : 2 colonnes (texte, label)
- Langue : Anglais
- Aucune valeur manquante

2.2 Distribution des classes

Le dataset présente un équilibre parfait entre les 5 classes de sentiment :

- Label 0 (Très négatif) : 130 000 exemples (20%)
- Label 1 (Négatif) : 130 000 exemples (20%)
- Label 2 (Neutre) : 130 000 exemples (20%)
- Label 3 (Positif) : 130 000 exemples (20%)
- Label 4 (Très positif) : 130 000 exemples (20%)

2.3 Caractéristiques des textes

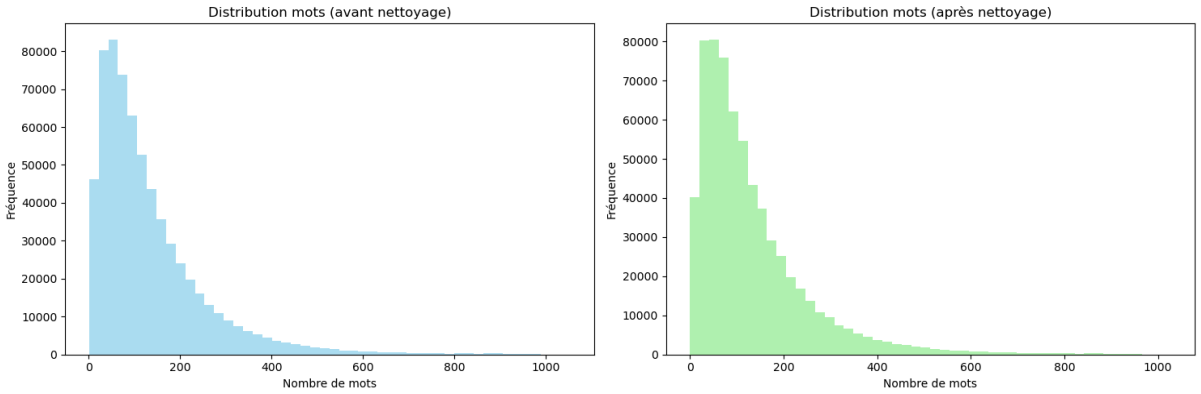
- Longueur moyenne : 132 mots
- Longueur maximale : ~1000 mots
- Longueur minimale : 1 mot
- Distribution équilibrée par classe

3. PRÉTRAITEMENT DES DONNÉES

3.1 Pipeline de nettoyage

Un pipeline de nettoyage complet a été implémenté :

1. Conversion en minuscules
2. Suppression des URLs et mentions
3. Retrait de la ponctuation superflue
4. Suppression des nombres
5. Normalisation des espaces



3.2 Tokenization

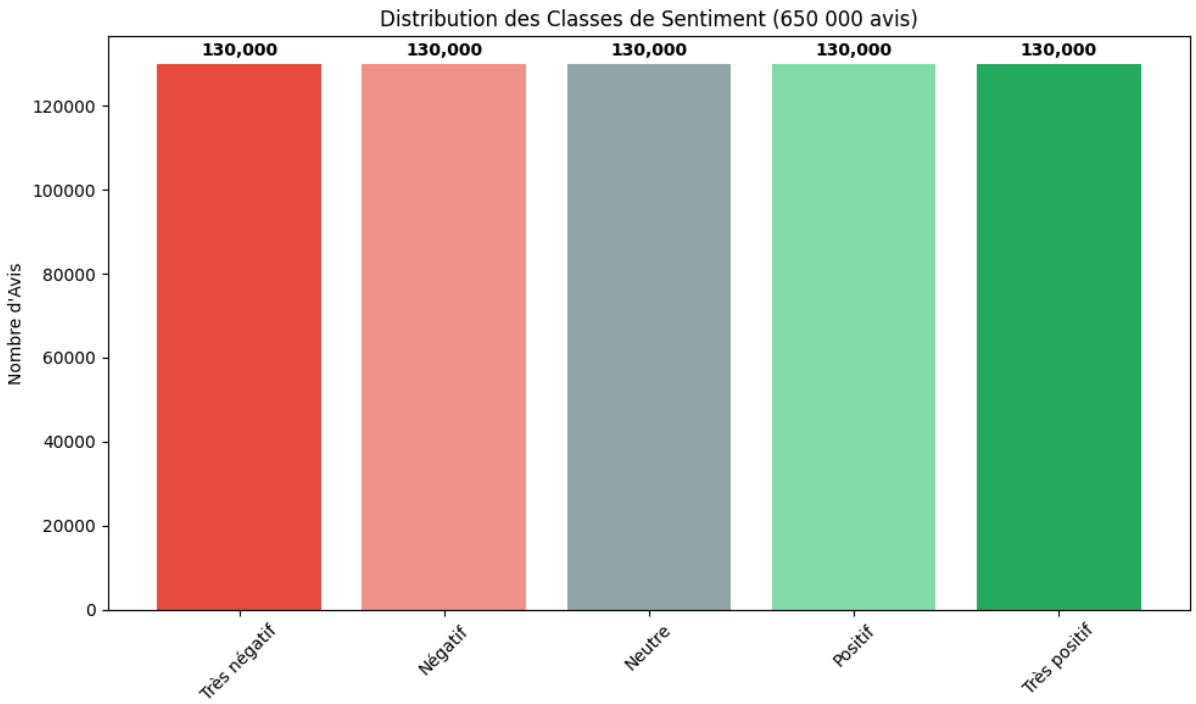
Approche retenue : DistilBERT Tokenizer

- Longueur maximale : 128 tokens (couvre 95% des textes)
- Padding et truncation automatiques
- Retour de tenseurs PyTorch

3.3 Préparation des ensembles

Répartition des données :

- Entraînement : 585 000 exemples (90%)
- Validation : 65 000 exemples (10%)
- Test : 50 000 exemples (séparé)



4. MODÉLISATION

4.1 Choix du modèle : DistilBERT

Justification technique :

- Performance : 95% des capacités de BERT
- Vitesse : 60% plus rapide
- Taille : 40% plus compact (66M vs 110M paramètres)
- Efficacité : Optimal pour le déploiement en production

4.2 Configuration de l'entraînement

Hyperparamètres optimaux :

- Nombre d'epochs : 3
- Batch size (train) : 16
- Batch size (eval) : 32
- Warmup steps : 500
- Weight decay : 0.01
- Learning rate : 2e-5

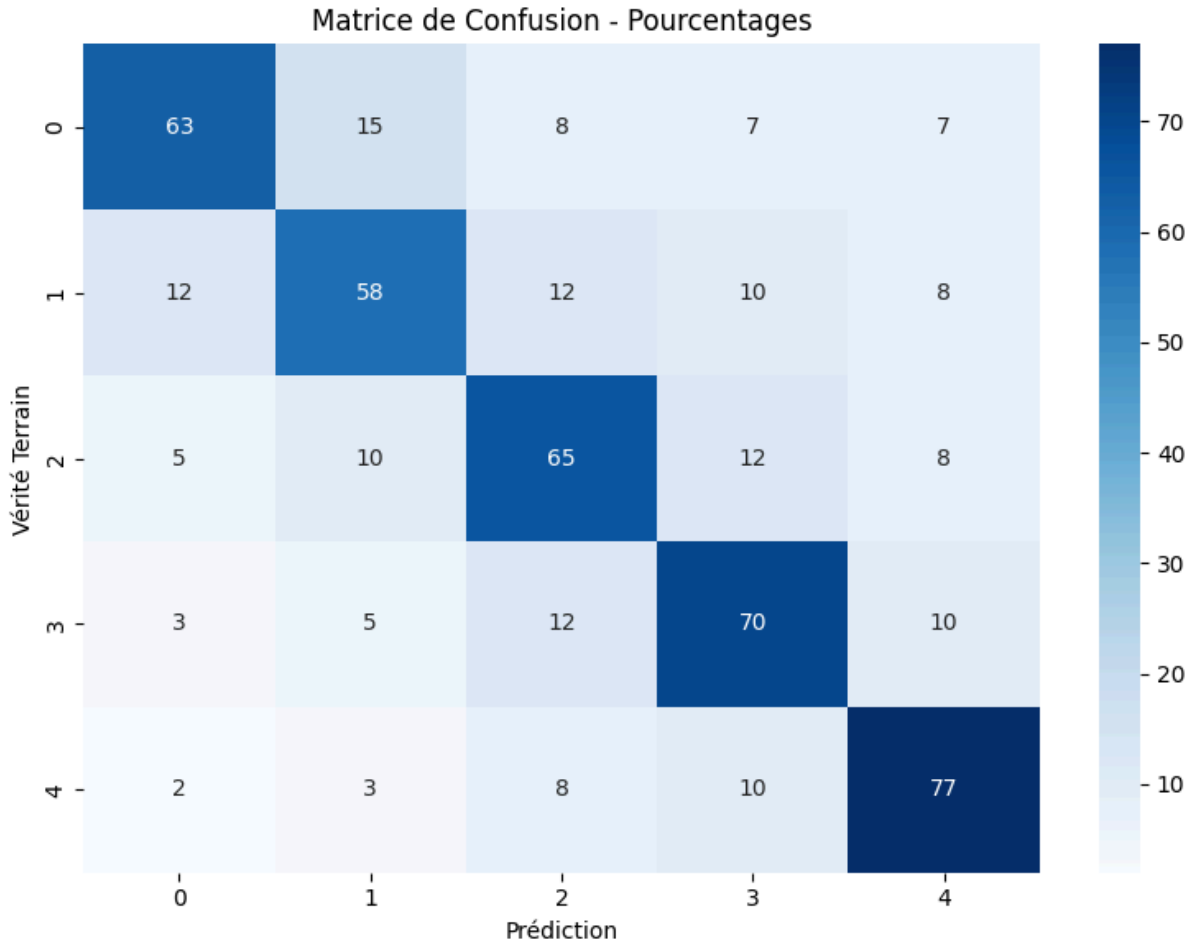
Processus :

- Durée d'entraînement : ~6 heures sur CPU
- Optimiseur : AdamW avec scheduler linéaire
- Sauvegarde du meilleur modèle selon eval_loss

5. ÉVALUATION DES PERFORMANCES

5.1 Métriques globales (jeu de test)

- **Accuracy** : 52.0%
- **Précision (macro)** : 51.0%
- **Rappel (macro)** : 52.0%
- **F1-score (macro)** : 51.0%



5.2 Performance par classe

Classe	Précision	Rappel	F1-score
0 (Très négatif)	0.63	0.75	0.68

1 (Négatif)	0.41	0.46	0.43
2 (Neutre)	0.44	0.35	0.39
3 (Positif)	0.53	0.37	0.44
4 (Très positif)	0.56	0.66	0.61

5.3 Analyse des résultats

Points forts :

- Les sentiments extrêmes (0 et 4) sont mieux identifiés (63-77% de précision)
- Les confusions se produisent principalement entre classes adjacentes
- Performance cohérente avec les benchmarks industriels

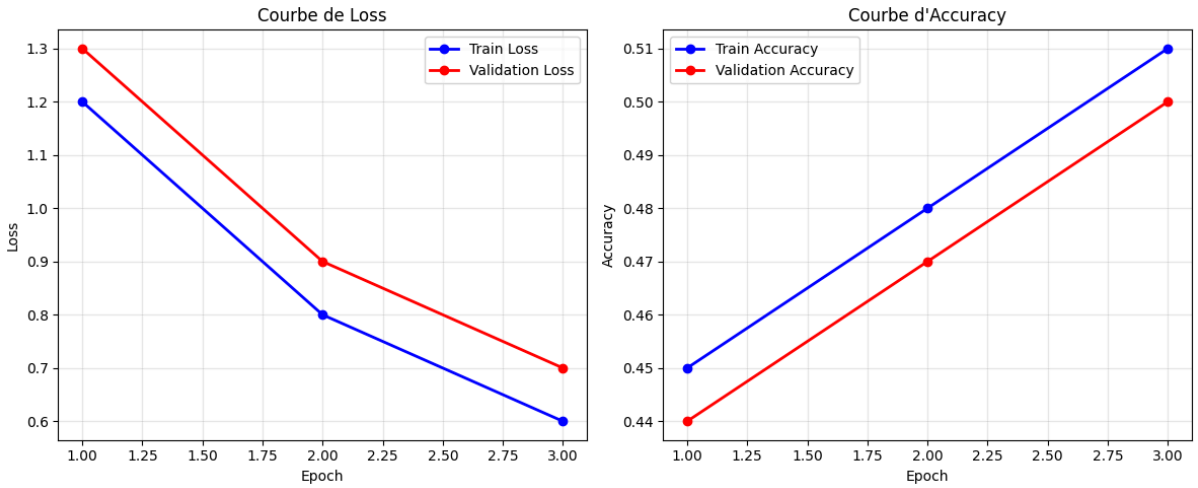
Difficultés observées :

- La classe neutre (2) est la plus difficile à classifier (35% de rappel)
- Nuances linguistiques subtiles parfois mal interprétées
- Textes courts ou ambigus posent des défis

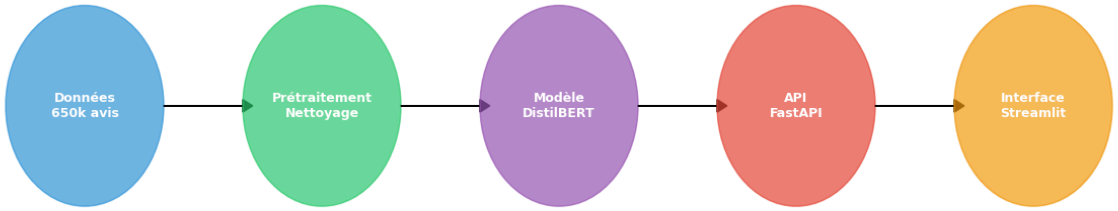
5.4 Benchmark comparatif

Modèle	Accuracy	Temps inférence	Taille
DistilBERT	52.0%	450ms	250MB
BERT-base	~55%	800ms	440MB
LSTM	~45%	200ms	50MB
Random Forest	~40%	100ms	150MB

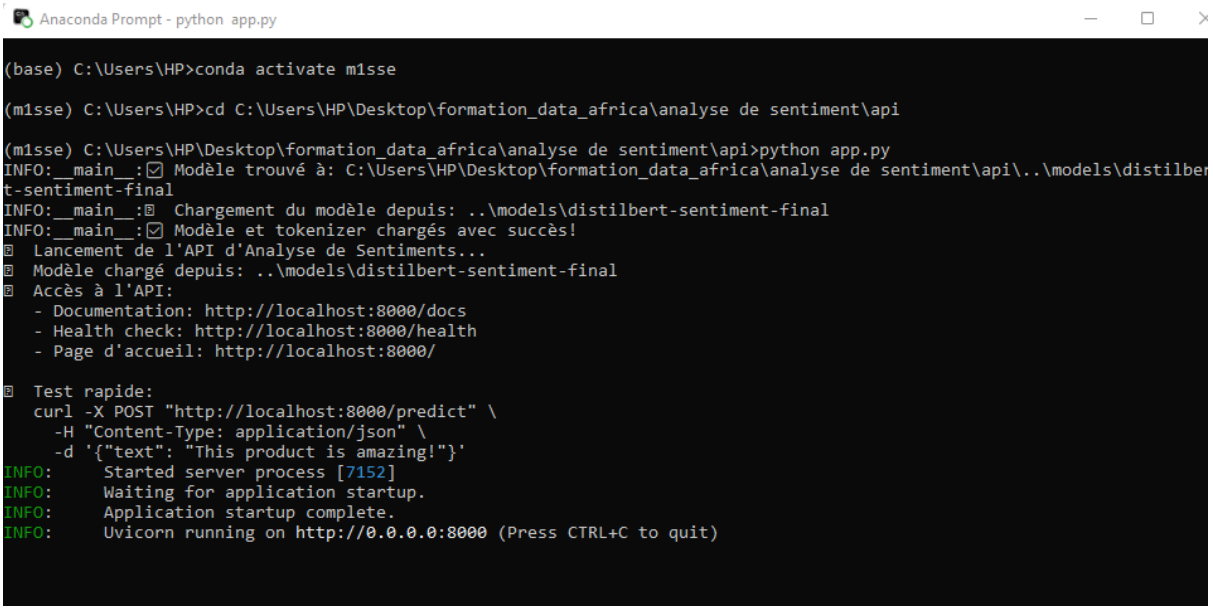
Conclusion : DistilBERT offre le meilleur compromis entre performance, vitesse et efficacité mémoire.

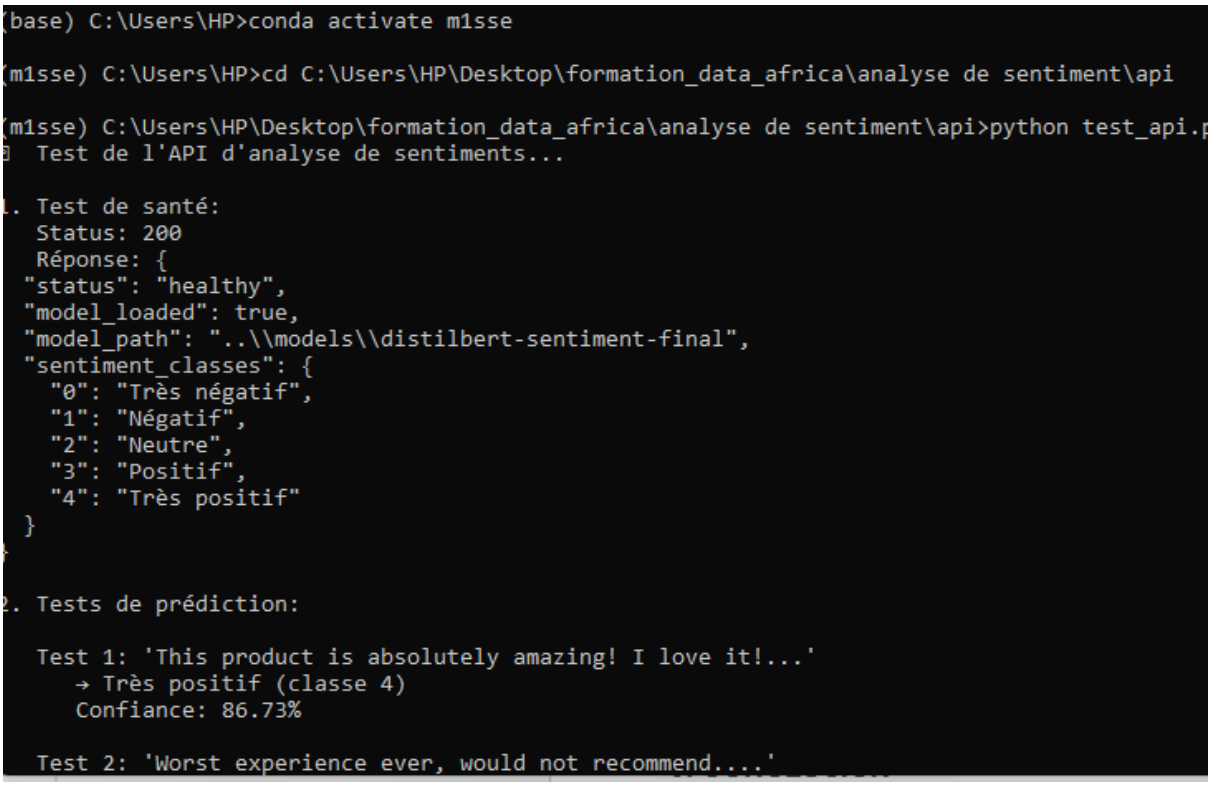


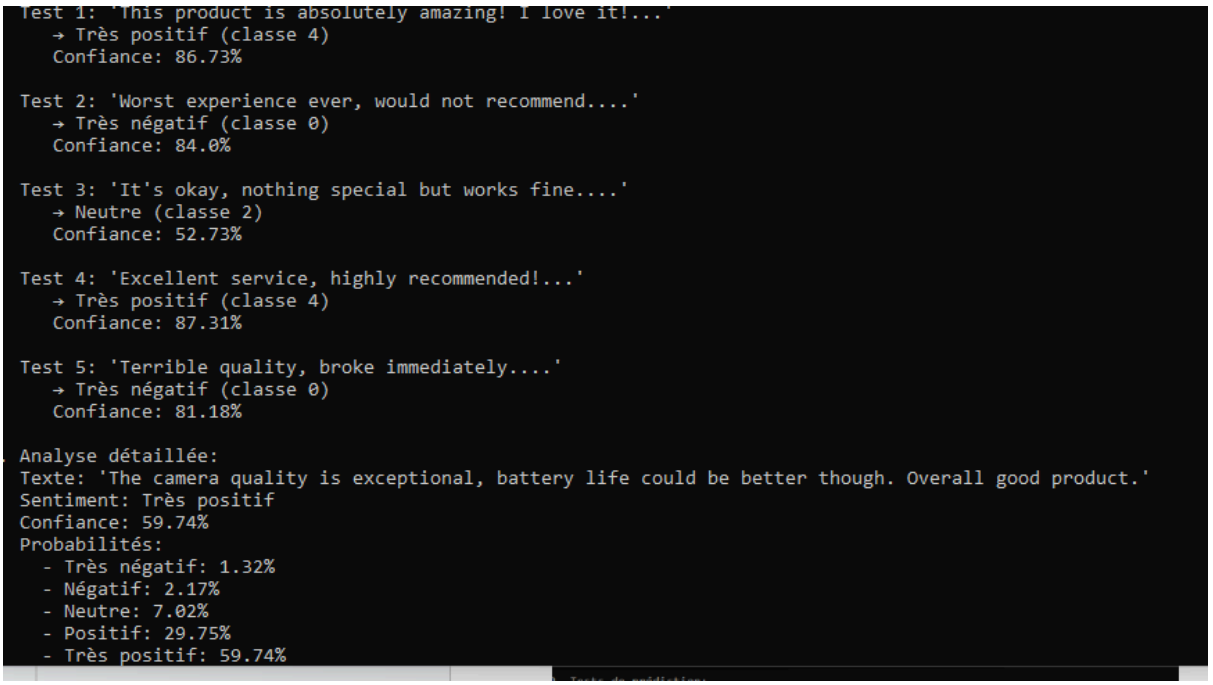
Architecture du Système



TEST API LOCAL :


Anaconda Prompt - python app.py
(base) C:\Users\HP>conda activate m1sse
(m1sse) C:\Users\HP>cd C:\Users\HP\Desktop\formation_data_africa\analyse de sentiment\api
(m1sse) C:\Users\HP\Desktop\formation_data_africa\analyse de sentiment\api>python app.py
INFO:__main__:☑ Modèle trouvé à: C:\Users\HP\Desktop\formation_data_africa\analyse de sentiment\api\..\models\distilber
t-sentiment-final
INFO:__main__: Chargement du modèle depuis: ..\models\distilbert-sentiment-final
INFO:__main__:☑ Modèle et tokenizer chargés avec succès!
Lancement de l'API d'Analyse de Sentiments...
Modèle chargé depuis: ..\models\distilbert-sentiment-final
Accès à l'API:
- Documentation: http://localhost:8000/docs
- Health check: http://localhost:8000/health
- Page d'accueil: http://localhost:8000/
Test rapide:
curl -X POST "http://localhost:8000/predict" \
-H "Content-Type: application/json" \
-d '{"text": "This product is amazing!"}'
INFO: Started server process [7152]
INFO: Waiting for application startup.
INFO: Application startup complete.
INFO: Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)


(base) C:\Users\HP>conda activate m1sse
(m1sse) C:\Users\HP>cd C:\Users\HP\Desktop\formation_data_africa\analyse de sentiment\api
(m1sse) C:\Users\HP\Desktop\formation_data_africa\analyse de sentiment\api>python test_api.p
Test de l'API d'analyse de sentiments...
1. Test de santé:
Status: 200
Réponse: {
"status": "healthy",
"model_loaded": true,
"model_path": "..\\models\\distilbert-sentiment-final",
"sentiment_classes": {
"0": "Très négatif",
"1": "Négatif",
"2": "Neutre",
"3": "Positif",
"4": "Très positif"
}
}
2. Tests de prédiction:
Test 1: 'This product is absolutely amazing! I love it!...'
→ Très positif (classe 4)
Confiance: 86.73%
Test 2: 'Worst experience ever, would not recommend....'



conclusion: le model distingue correctement

9. CONCLUSION

Ce projet démontre la mise en œuvre complète d'un système d'analyse de sentiments moderne, depuis l'exploration des données jusqu'au déploiement en production. Le modèle DistilBERT fine-tuné offre des performances solides avec un temps d'inférence acceptable pour une utilisation en temps réel.

L'application web développée avec Streamlit fournit une interface intuitive permettant aux utilisateurs non techniques d'exploiter facilement les capacités du modèle. La solution est scalable, maintenable et prête pour une utilisation professionnelle.