Compte Rendu — Analyse de Sentiments IMDB

Objectif

Créer un pipeline de classification de sentiments sur le dataset IMDB (50 000 reviews), en utilisant TF-IDF pour la vectorisation et KNN pour la classification.

Pipeline Développé

**1. Préparation des données**

• Nettoyage avancé du texte (HTML, caractères non alphabétiques, lowercase, stopwords, stemming).

• Construction d'une nouvelle colonne « review_clean ».

**2. Vectorisation TF-IDF**

• Extraction de 5 000 features significatives.

• Transformation en matrice sparse.

**3. Modélisation KNN**

• Split 80/20 train-test.

• Évaluation de K = 1 à 19.

• Sélection du meilleur K via l'accuracy max.

**4. Entraînement final**

• Modèle KNN ajusté avec le meilleur paramètre.

• Génération des prédictions sur le jeu de test.

**5. Évaluation**

• Accuracy globale.

• Matrice de confusion.

• Rapport de classification (precision, recall, f1-score).

Résultats

Le pipeline est complet et opérationnel. Le modèle KNN offre des performances correctes malgré la complexité des données textuelles.

Axes d'Amélioration

• Tester Naive Bayes, Logistic Regression, SVM.

• Ajouter une lemmatisation.

• Intégrer un front (Streamlit, Flask API).