📄 RAPPORT FINAL

CLASSIFICATION DE SENTIMENTS IMDB

Introduction

Ce projet vise à construire un système automatique capable de déterminer si une critique de film est positive ou négative. Pour cela, nous avons utilisé le dataset IMDB composé de 50 000 avis textuels. Le travail s'articule autour du nettoyage des données, de leur transformation en représentations numériques et de l'application d'un algorithme de Machine Learning : le K-Nearest Neighbors (KNN).

L'objectif est de transformer un texte brut en une prédiction exploitable reflétant le sentiment exprimé par l'utilisateur.

1. Présentation de l'Algorithme KNN

1.1 Définition

Le K-Nearest Neighbors (KNN) est un algorithme d'apprentissage supervisé utilisé pour la classification. Il adopte une logique directe : il ne construit pas de modèle complexe pendant l'entraînement. Au lieu de cela, il prend ses décisions en se basant sur les exemples déjà connus.

L'hypothèse est simple et intuitive : des données similaires appartiennent généralement à la même catégorie .

1.2 Principe de fonctionnement

Le processus du KNN se déroule en quatre étapes :

- **Représentation numérique des données** : ici, chaque critique de film est transformée en vecteur grâce au TF-IDF.
- **Mesure de la distance** entre la nouvelle critique et les critiques déjà enregistrées.
- **Sélection des K voisins** les plus proches.
- **Vote majoritaire** : la classe la plus fréquente parmi ces voisins est prédite.

➡️ Ce fonctionnement imite un raisonnement humain : on compare la nouvelle information à celles que l'on connaît déjà avant de décider.

1.3 Domaines d'utilisation

Le KNN est utilisé dans des tâches variées, notamment :

- analyse de sentiments,
- systèmes de recommandation basés sur la similarité,
- classification simple d'images,
- segmentation basique de clients.

Il est souvent employé comme point de départ pour comprendre les concepts fondamentaux du Machine Learning.

1.4 Avantages

- **Simple à comprendre et à expliquer** — Pas besoin d'un modèle mathématique complexe.
- **Pas d'entraînement coûteux** — Le travail se fait au moment de la prédiction.
- **Polyvalence** — Utilisable pour la classification et la régression.

1.5 Inconvénients

- **Lent sur de gros datasets** — Chaque prédiction nécessite de comparer des milliers d'exemples.
- **Sensible au choix de K** — Mauvaise valeur = perte de précision.
- **Peu efficace dans les espaces de grande dimension** — Comme avec TF-IDF, où chaque mot devient une variable.

2. Pipeline du Projet

2.1 Nettoyage des données

Les critiques IMDB contiennent des éléments inutiles pour l'analyse. Un nettoyage a été effectué comprenant :

- suppression des balises HTML,
- filtrage des caractères non alphabétiques,
- passage en minuscules,
- suppression des stopwords (mots fréquents sans valeur sémantique),
- stemming pour réduire les mots à leur racine.

Une nouvelle colonne `review_clean` stocke le texte prêt à être traité.

2.2 Vectorisation TF-IDF

Les modèles ne comprennent pas directement le langage humain. TF-IDF transforme chaque critique en un vecteur indiquant l'importance de chaque mot.

Plus un mot est pertinent dans un document par rapport au corpus entier, plus son poids est élevé.

2.3 Entraînement du modèle KNN

- Division du dataset en 80 % entraînement / 20 % test.
- Test de plusieurs valeurs de K pour choisir le meilleur paramètre.
- Entraînement final avec la valeur optimale.
- Prédiction sur les critiques du jeu de test.

3. Résultats

Le modèle KNN est parvenu à distinguer les avis positifs des avis négatifs de manière cohérente. Bien que ce ne soit pas l'algorithme le plus performant pour les données textuelles volumineuses, il constitue une base solide pour comprendre la logique de classification.

4. Améliorations possibles

Pour améliorer ce travail, plusieurs options sont envisageables :

- remplacer KNN par **Naive Bayes, SVM ou Logistic Regression**, plus adaptés aux données textuelles,
- ajouter une **lemmatisation** pour un traitement linguistique plus fin,
- développer une **interface utilisateur** (web ou desktop) pour exploiter le modèle.

Conclusion

Ce projet montre comment un système informatique peut analyser du texte et prédire automatiquement un sentiment. Bien que simple, le KNN offre une première approche opérationnelle du traitement automatique du langage naturel.

Ce travail pose les fondations nécessaires pour concevoir des modèles plus performants et plus intelligents.

FIN DU RAPPORT