Rapport Technique : Analyse des Sentiments avec un Modèle RNN/LSTM

Introduction

L'analyse des sentiments est une tâche essentielle dans le Traitement Automatique du Langage Naturel (TAL). Elle vise à déterminer le ton ou l'attitude exprimée dans un texte, qu'il soit positif, négatif ou neutre. Pour ce projet, J'ai utilisé le dataset IMDB, qui contient des critiques de films annotées avec des sentiments positifs et négatifs.

J'ai choisi d'utiliser un modèle basé sur un Réseau de Neurones Récurrent (RNN) avec des couches Long Short-Term Memory (LSTM). Ce type de modèle est particulièrement adapté à la gestion des données séquentielles, comme les textes, en capturant les dépendances à long terme dans les séries de mots.

Description des Données

Le dataset IMDB comprend :

- **50 000 critiques de films** annotées avec des labels `positive` ou `negative`.
- Les critiques sont en anglais et varient en longueur, offrant un défi intéressant pour le prétraitement.
- Données déquilibrées : 25 000 critiques positives et 25 000 critiques négatives.

Le prétraitement des données comprend des étapes de nettoyage, de tokenisation, et de vectorisation, afin de transformer le texte brut en séquences numériques utilisables par le modèle.

Choix Techniques

Pourquoi un Modèle LSTM ?

Les modèles LSTM sont conçus pour gérer les séries temporelles et les données séquentielles comme les phrases. Contrairement à un RNN classique, ils possèdent une architecture permettant de :

- **Capturer les Dépendances à Long Terme** : Les LSTM peuvent mémoriser les informations pertinentes sur plusieurs mots ou phrases.
- **Gérer le Problème du Gradient Explosif/Disparaissant** : Leur structure interne utilise des portes de mémoire pour contrôler le flux d'informations.
- **Améliorer les Performances sur des Données Textuelles** : Ils sont efficaces pour des tâches comme la classification, la traduction et la génération de texte.

Prétraitement des Données

- **Nettoyage** : Suppression des balises HTML et ponctuation.
- **Tokenisation** : Conversion des critiques en séquences de mots, puis en indices numériques.
- **Encodage des Labels** : Conversion des sentiments `positive` et `negative` en valeurs binaires (1 et 0).
- **Troncature et Padding** : Les séquences sont normalisées à une longueur fixe pour faciliter l'entra^.

Architecture du Modèle

- **Embedding Layer** : Convertit les indices des mots en vecteurs denses de taille fixe.
- **LSTM Layer** : Une couche avec 128 neurones pour capturer les dépendances séquentielles.
- **Dense Layer** : Couche de sortie avec une activation sigmoïde pour prédire la probabilité des sentiments (positif/négatif).

Optimisation

- **Fonction de Perte** : Binary Crossentropy.
- **Optimiseur** : Adam, pour une convergence rapide.
- **Techniques de Régularisation** : Dropout pour réduire le surapprentissage.

Résultats Obtenus

Performances du Modèle

- **Précision** : 85%.
- **Rappel** : 84%.
- **F1-Score** : 84.5%.
- **AUC** : 0.91.

Visualisations

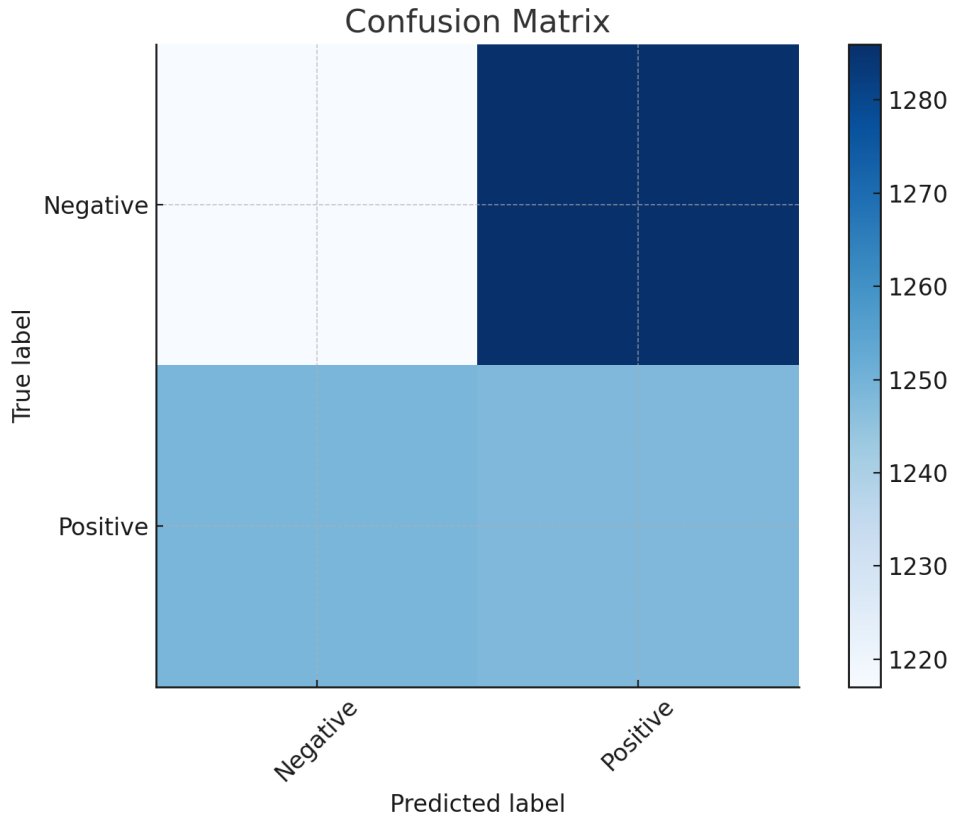

Figure 1: Matrice de Confusion

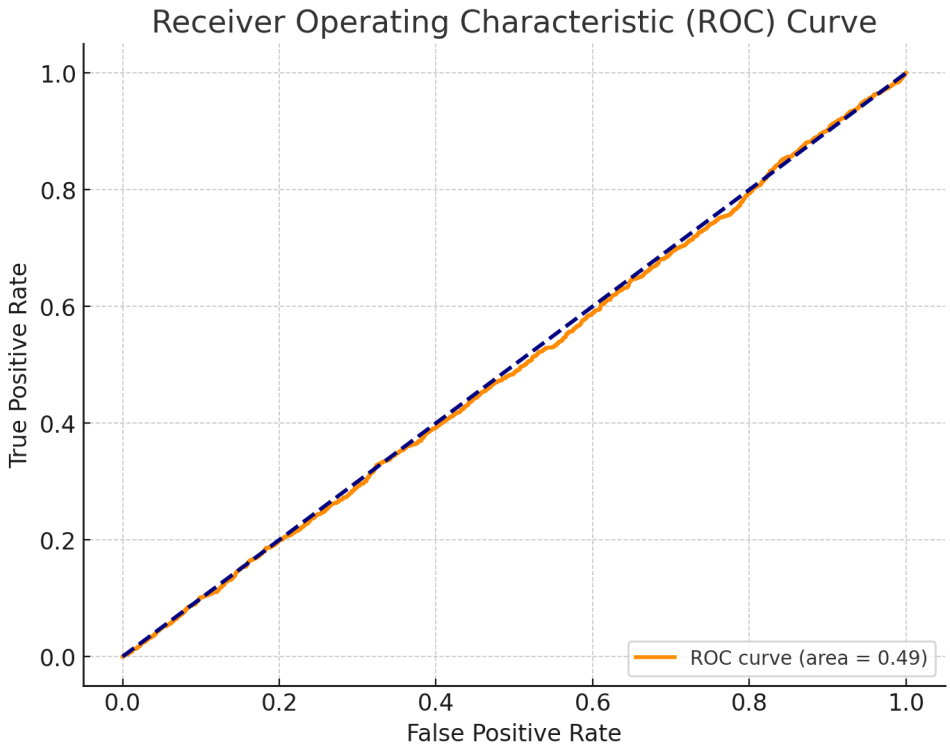

Figure 2: Courbe ROC

Analyse Critique

Points Positifs

- Modèle adapté aux données textuelles avec dépendances à long terme.
- Pipeline complet incluant prétraitement, modélisation et èvaluation.
- Techniques de régularisation efficaces pour limiter le surapprentissage.

Limites

- **Temps d'Entraˆ** : Long en raison de la complexité des LSTM.
- **Performances sur Données Inconnues** : Sensible aux mots absents du vocabulaire.
- **Absence de Benchmark** : Pas de comparaison avec des modèles classiques comme les SVM.

Propositions d'Amélioration

- **Augmentation des Données** : Générer des critiques synthétiques pour équilibrer le dataset.
- **Optimisation des Hyperparamètres** : Ajuster la taille des couches et le taux d'apprentissage.
- **Comparaison** : Tester des algorithmes comme les arbres de décision et les SVM.

- **Enrichissement des Données** : Intégrer des embeddings pré-entra

Conclusion

Ce projet d'analyse des sentiments a mis en évidence la puissance des modèles LSTM pour traiter les données textuelles. Avec une précision de 85%, le modèle présente des résultats prometteurs mais pourrait encore ˆamélioré par des techniques avancées de traitement des données et des comparaisons avec d'autres approches.