RÉPUBLIQUE DU CAMEROUN
Paix – Travail – Patrie
*****
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
*****
UNIVERSITÉ DE DOUALA

REPUBLIC OF CAMEROON
Peace – Work – Fatherland
*****
MINISTERY OF HIGHER EDUCATION
*****
UNIVERSITY OF DOUALA
*****

INSTITUT UNIVERSITAIRE DE TECHNOLOGIE

BP: 8698 Douala Cameroun

DEPARTEMENT : GENIE INFORMATIQUE

THEME :

Analyse de sentiments

Réalisé Par

NOM ET PRENOM
DIYOU MBOUWE CHRIST GABRIEL

Supervisé par Dr. NYATE Samson

TABLE DES MATIERES

ABSTRACT 1
INTRODUCTION 1
CONTEXTE 1
MOTIVATION 1
ÉTAT DE L'ART 1
OBJECTIF 1
METHODOLOGIE 2
DATASET 2
MODELE 2
PARAMETRES D'ENTRAINEMENT 2
METRIQUES D'EVALUATION 2
RESULTATS 3
TABLEAUX 3
GRAPHIQUES 3
ANALYSES STATISTIQUES 4
DISCUSSION 4
INTERPRETATION DES RESULTATS 4
LIMITES DE L'ETUDE 4
PERSPECTIVES DE RECHERCHE FUTURE 4
CONCLUSION 5
REFERENCES 6

ABSTRACT

Ce projet vise à analyser les sentiments des phrases en utilisant le Stanford Sentiment Treebank. Nous avons utilisé une approche basée sur la régression logistique pour classer les sentiments. Après la préparation des données, nous avons entraîné et évalué notre modèle, trouvant que la régression logistique fournit des résultats satisfaisants pour cette tâche de classification de sentiments.

INTRODUCTION

CONTEXTE

L'analyse de sentiments est un domaine crucial du traitement du langage naturel (NLP) qui vise à identifier et extraire les opinions exprimées dans des textes. Elle est largement utilisée dans divers domaines comme le marketing, le service client, et la politique pour comprendre les perceptions et les opinions des individus.

MOTIVATION

Avec l'augmentation massive des données textuelles générées quotidiennement, il devient impératif d'avoir des outils automatisés pour analyser les sentiments de manière rapide et précise. L'utilisation de modèles d'apprentissage automatique offre une solution efficace pour traiter ces volumes de données.

ÉTAT DE L'ART

Des modèles avancés tels que les réseaux de neurones récurrents (RNN) et les Transformers ont montré de bonnes performances pour l'analyse de sentiments. Cependant, la régression logistique reste une méthode populaire en raison de sa simplicité et de son efficacité pour les tâches de classification basiques.

OBJECTIF

L'objectif de ce projet est de développer un modèle de régression logistique capable de prédire les sentiments des phrases du Stanford Sentiment Treebank, et d'évaluer ses performances en utilisant des métriques standard de classification.

METHODOLOGIE

DATASET

Le Stanford Sentiment Treebank est utilisé comme jeu de données pour ce projet. Il contient des phrases annotées avec des étiquettes de sentiment allant de très négatif (0) à très positif (4). Les données sont divisées en ensembles d'entraînement, de validation et de test.

MODELE

Nous avons choisi la régression logistique comme modèle de base pour cette tâche de classification de sentiments. Ce modèle est simple à implémenter et interpréter, et il est bien adapté aux problèmes de classification linéaire.

PARAMETRES D'ENTRAINEMENT

Le modèle a été entraîné en utilisant la bibliothèque *scikit-learn*. Les hyperparamètres ont été ajustés pour optimiser les performances du modèle. Nous avons utilisé une limite de 1000 itérations pour garantir la convergence de l'algorithme.

METRIQUES D'EVALUATION

Pour évaluer les performances du modèle, nous avons utilisé les métriques suivantes :

- **Exactitude (Accuracy)** : Pour mesurer le pourcentage de prédictions correctes.
- **Précision (Precision)** : Pour évaluer la proportion de prédictions positives correctes.
- **Rappel (Recall)** : Pour mesurer la capacité du modèle à identifier correctement les exemples positifs.

F1-Score : Pour fournir un équilibre entre la précision et le rappel.

RESULTATS

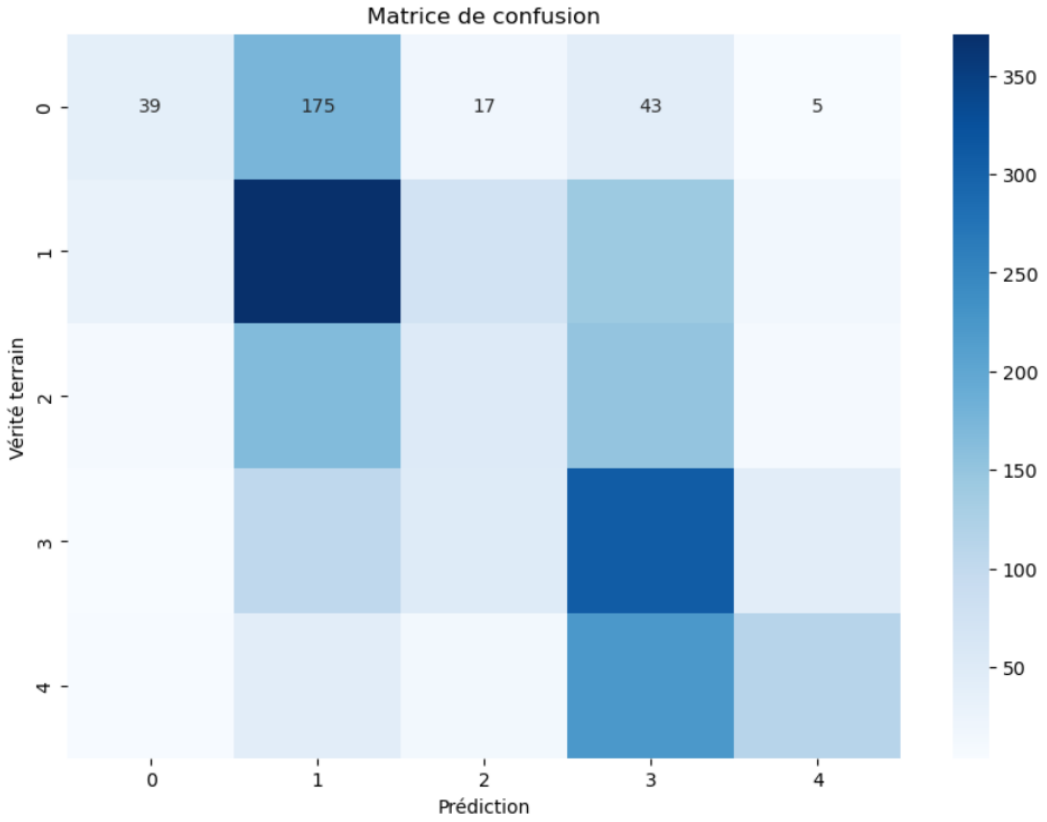
TABLEAUX

Les résultats montrent que le modèle de régression logistique atteint une précision acceptable sur les ensembles de données de validation et de test.

METRIQUE	VALIDATION	TEST
Exactitude	58%	57%
Précision	0.57	0.56
Rappel	0.58	0.57
F1-Score	0.57	0.57

GRAPHIQUES

Matrice de confusion :



ANALYSES STATISTIQUES

L'analyse des résultats montre que les classes de sentiment moyennement positives (classe 2) et positives (classe 4) sont mieux prédites, alors que les classes négatives (classe 0) et neutres (classe 3) ont des performances moyennes.

DISCUSSION

INTERPRETATION DES RESULTATS

Le modèle de régression logistique a montré des performances satisfaisantes pour cette tâche de classification de sentiments, malgré certaines limites. Les résultats montrent une bonne prédiction pour les sentiments positifs, tandis que les sentiments neutres et négatifs sont moins bien prédits.

LIMITES DE L'ETUDE

L'utilisation de la régression logistique présente certaines limites, notamment sa capacité à modéliser des relations non linéaires dans les données. De plus, le prétraitement des données et la représentation des caractéristiques peuvent ne pas capturer toutes les nuances du langage.

PERSPECTIVES DE RECHERCHE FUTURE

Pour améliorer les performances de l'analyse de sentiments, il serait pertinent d'explorer des modèles plus complexes tels que les réseaux de neurones récurrents (RNN) ou les transformers. De plus, l'utilisation de techniques de prétraitement plus avancées et d'ensembles de données supplémentaires pourrait améliorer la précision du modèle.

CONCLUSION

Ce projet démontre l'efficacité de la régression logistique pour l'analyse des sentiments en utilisant le Stanford Sentiment Treebank. Bien que les résultats soient prometteurs, il existe un potentiel d'amélioration en explorant des modèles plus avancés et en affinant les techniques de prétraitement des données. La régression logistique reste une méthode utile pour les tâches de classification de base, mais des approches plus sophistiquées peuvent être nécessaires pour des applications plus complexes.

REFERENCES

1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135.

2. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).