Liste des sujets de mastère de recherche en collaboration avec le laboratoire LARODEC (ISG – Université de Tunis)

Responsable : Dr. Aymen LOUATI

E-mail : aymen.louati@isikef.u-jendouba.tn

Téléphone : +216 95 335 375

Sujet 01

Analyse des sentiments des avis en ligne

Contexte et importance du sujet : Avec la croissance des plateformes numériques, les avis en ligne jouent un rôle essentiel dans la prise de décision des consommateurs. Ces avis influencent les ventes, la réputation des entreprises, et la fidélité des clients. L'analyse des sentiments permet de déterminer automatiquement si les avis sont positifs, négatifs ou neutres, ce qui aide les entreprises à mieux comprendre les opinions des utilisateurs et à ajuster leurs stratégies marketing et opérationnelles. Ce sujet est au croisement du traitement automatique du langage naturel (NLP), de l'apprentissage automatique et des applications commerciales concrètes.

Objectifs du projet :

1. Développer un modèle capable de classifier les avis en trois catégories : positif, négatif, et neutre.
2. Évaluer les performances des algorithmes classiques de machine learning sur cette tâche.
3. Analyser les points forts et faibles des différentes approches utilisées.
4. Fournir une base solide pour de futures améliorations ou extensions vers des modèles avancés.

Plan de travail détaillé :

1. Collecte et exploration des données :
    - Utiliser un jeu de données public comme :
        - Amazon ReviewsDataset (avis sur des produits).

        - IMDB Dataset (avis sur des films).
    - Effectuer une exploration initiale : distribution des sentiments, longueur moyenne des avis, vocabulaire le plus fréquent.
2. Prétraitement des données textuelles :
    - Nettoyage des données :
        - Suppression des caractères spéciaux, des espaces inutiles et des mots vides (stopwords).
    - Tokenisation :
        - Conversion des phrases en une liste de mots ou de n-grammes.
    - Vectorisation :
        - Utilisation de techniques comme TF-IDF ou Bag of Words pour représenter les avis sous forme de vecteurs numériques.
3. Développement des modèles d'apprentissage automatique :
    - Implémenter et comparer les algorithmes suivants :
        - Naïve Bayes (MultinomialNB).
        - SVM (Support Vector Machine).
        - Autres algorithmes possibles : LogisticRegression, Random Forest.
    - Comparer les performances sur un jeu de test en utilisant des métriques standard comme :
        - Accuracy, Precision, Recall, F1-score.
4. Évaluation et analyse des résultats :
    - Analyser les erreurs fréquentes pour identifier les limites du modèle.
    - Effectuer une validation croisée pour garantir la robustesse des résultats.
    - Visualiser les résultats via des graphiques (matrices de confusion, scores par classe, etc.).
5. Optimisation et expérimentation :
    - Explorer des techniques de régularisation pour améliorer les performances des modèles.
    - Comparer l'impact des différents paramètres d'hyperparamètres des algorithmes (par exemple, choix du noyau pour SVM).
6. Documentation et livrables :
    - Rédiger un rapport complet décrivant :

- Le processus de développement.
- Les résultats obtenus.
- Les conclusions et les recommandations pour des travaux futurs.

Livrables attendus :

1. Rapport détaillé documentant :

- Les étapes méthodologiques.
- Les résultats des tests et analyses.
- Les conclusions sur les performances des modèles.

2. Code source (en Python avec des bibliothèques comme Scikit-learn, NLTK, ou SpaCy).

Applications concrètes :

- Suivi de la satisfaction des clients pour améliorer les produits et services.
- Analyse automatisée de la réputation en ligne des entreprises.
- Application directe dans les plateformes d'e-commerce ou de services.

Références académiques et techniques :

1. **Traitement automatique du langage naturel :**
   - Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
   - Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing. Pearson.
2. **Machine Learning et modèles :**
   - Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.
3. **Datasets publics :**
   - Amazon Reviews Dataset : Lien Kaggle.
   - IMDB Dataset : Lien Stanford.
4. **Tutoriels et ressources pratiques :**
   - Documentations des bibliothèques : Scikit-learn, NLTK, SpaCy.