Nom et Prénom : LAGRARI FATIMA-EZZAHRA
Date de soutenance : 15/04/2023
Directeur de Thèse : EL MERABET YOUSSEF

Sujet de Thèse :

Text categorization and sentiment analysis using machine and deep learning methods

Résumé :

2,5 billions d'octets de données sont produites chaque jour sur internet : emails, vidéos, informations météorologiques, signaux GPS, transactions en ligne, cryptomonnaie, etc. Une partie de ces données est stockée sous forme de textes de manière non structurée. Les méthodes d'analyse de données conventionnelles ne parviennent pas à extraire des informations significatives de ces énormes quantités de données. Cela a motivé la communauté scientifique à rechercher de nouveaux algorithmes afin de pouvoir stocker, classer et analyser ce type de données d'une manière plus fiable.

Afin de contribuer à ce processus, le but de cette thèse est tout d'abord de donner un aperçu des approches classiques et approfondies de l'apprentissage et de leurs applications et de développer des modèles d'apprentissage qui peuvent automatiquement classer des données textuelles à grande échelle en fonction de différentes méthodes et être capables de prédire des cibles avec un niveau de précision élevé. Cependant, les approches d'apprentissage classiques et automatiques n'ont pas été à la hauteur en termes de précision, c'est pourquoi nous proposons deux modèles robustes basés sur des méthodes hybrides de classification automatique et d'apprentissage profond qui nous ont permis d'obtenir une meilleure précision de classification.

Le premier modèle est un modèle de catégorisation textuelle se basant sur des méthodes combinées d'apprentissage automatique avec une étape de réduction de dimensionnalité en utilisant l'algorithme génétique et des forêts aléatoires pour la catégorisation de documents. Le second Modèle a pour but d'analyser les sentiments sur le réseau social twitter en utilisant des méthodes d'apprentissage profond et plus précisément en utilisant une version optimale de l'algorithme BERT basé sur l'algorithme de Lion. Les deux modèles sont testés sur des jeux de données gratuits et publics et donnent de meilleurs résultats par rapport aux approches existantes dans la littérature.

Mots-clefs : classification textuelle, catégorisation, analyse sentimentale, apprentissage automatique, apprentissage profond, réduction de dimension, foret aléatoire, Bert, algorithme de lion

Abstract:

2.5 trillion bytes of data are produced every day in the internet: emails, videos, weather information, GPS signals, online transactions, crypto currencies etc. Part of this data is stored as texts in an unstructured way. Conventional data analysis methods fail in extracting meaningful information from these huge amounts of data. This motivated the scientific community to look for new algorithms in order to be able to store, classify and analyse this kind of data efficiently.

In order to contribute to this process, the objective of this thesis is to first give an overview of classical and deep learning approaches and then to develop optimal learning models that can automatically classify large-scale text data based on different methods. These models are able to predict targets with a high accuracy level. However, Machine learning approaches have fallen short in terms of accuracy, that's why we propose in this work two models based on hybrid machine learning methods and deep learning for automatic text classification to achieve better accuracy.

The first model combines a dimensionality reduction step using genetic algorithm with random forests for text classification of business documents and the second one uses an enhanced version of BERT (Bidirectional Encoder Representations from Transformers) algorithm with lion algorithm for sentiment analysis on twitter Dataset. These models are tested on free and public datasets and show better results compared to existing approaches.

Keywords: Text classification, categorization, sentiment analysis, machine learning, deep learning, feature selection, random forests, Bert, lion algorithm.