جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Madame Ikram EL KARFI

Soutiendra publiquement sa thèse de Doctorat en Informatique

Le Lundi 26 Février 2024 à 14h30 au Grand Amphi à l'ENSIAS

Intitulé de la thèse

Arabic Sentiment Analysis using Advanced Machine Learning Techniques

Président :

Pr. Rachid OULAD HAJ THAMI, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :

Pr. Sanaa EL FKIHI, PES, ENSIAS, Université Mohammed V de Rabat

Rapporteurs :

Pr. Nour-eddine EL FADDOULI, PES, EMI, Université Mohammed V de Rabat

Pr. Najima DAOUDI, PES, École des Sciences de l'Information, Rabat

Pr. Abdelalim SADIQ, PES, Faculté des Sciences, Université Ibn Tofail, Kenitra

Examinateur :

Pr. Rdouan FAIZI, PES, ENSIAS, Université Mohammed V de Rabat

Résumé: La révolution numérique et technologique a contribué à ce que l'Internet soit considéré comme l'une des technologies les plus évolutives et les plus croissantes. Le nombre total des internautes n'a cessé d'augmenter. Au niveau mondial, les internautes communiquent, partagent du contenu et expriment leurs opinions ou leurs sentiments sur Internet sur un large panel de sujets dans des groupes de discussion, des blogs, des forums et d'autres sites Web publics. Par conséquent, il est indispensable de développer des systèmes capables de détecter et d'analyser les opinions dans des domaines aussi variés que la politique, la santé et le marketing.

Depuis quelques années, l'analyse des sentiments est devenue un sujet de recherche en plein essor. Ce sujet de recherche consiste à identifier le sentiment exprimé dans une expression subjective. Une opinion est une expression subjective décrivant des pensées et des émotions personnelles. Ces pensées et ces émotions peuvent être associées à un certain sentiment. Les sentiments les plus étudiés sont : positif, neutre ou négatif. Bien que plusieurs travaux aient abordé le sujet d'analyse des sentiments, la plupart de ces travaux se sont concentrés sur le traitement des textes en anglais, et donc les études fournies en arabe nécessitent encore plus d'attention afin de tirer profit de la richesse de la langue arabe.

Dans cette perspective, nous avons exposé dans cette étude un jeu de données dédié à l'analyse de sentiments en arabe. Ce jeu de données est composé des critiques de livres en arabe collectées sur Goodreads associées à leurs étiquettes correspondantes. Par la suite, nous avons effectué une étude sur les principales recherches portant sur l'analyse des sentiments en arabe à l'aide de l'apprentissage profond. Selon les conclusions de cette étude, il a été révélé que différents modèles ont été utilisés. Il s'agit notamment des réseaux de neurones convolutifs (CNN), de la mémoire longue à court terme (LSTM), des réseaux neuronaux récurrents bidirectionnels (Bi-RNN), des réseaux de neurones profonds (DNN), des unités récurrentes fermées (GRU), ainsi que des réseaux de neurones basés sur l'attention. Ces modèles ont permis de comprendre l'analyse des sentiments dans cette langue morphologiquement complexe à différents niveaux, à savoir au niveau de la phrase, du document et de l'aspect. Au niveau de la phrase, il a été constaté que l'approche LSTM surpasse les autres modèles d'apprentissage profond tels que le CNN sur la plupart des jeux de données. Au niveau du document, CNN a fourni des meilleurs résultats, suivi de LSTM. Cependant, au niveau de l'aspect, il a été démontré que LSTM est le modèle le plus fréquemment utilisé par les chercheurs.

Dans cette étude, nous avons évalué l'impact de la combinaison du modèle d'apprentissage profond BiLSTM et du modèle ARAGPT pour l'analyse des sentiments en arabe. En fusionnant ces deux puissantes approches du traitement du langage naturel, nous visons à améliorer les performances de l'analyse des sentiments dans le contexte de l'arabe. Le modèle BiLSTM est reconnu pour sa capacité à saisir les dépendances à long terme dans une séquence, tandis que le modèle AraGPT excelle dans la génération de texte et la compréhension contextuelle. En

tirant parti de leurs forces respectives, nous cherchons à obtenir une analyse plus précise et nuancée des sentiments exprimés dans les textes arabes. Les résultats préliminaires de cette étude sont prometteurs, ouvrant ainsi de nouvelles perspectives pour une meilleure compréhension des émotions et des opinions dans cette langue complexe. En outre, pour tirer avantage de la performance des modèles de langage basés sur les transformers, nous avons expérimenté deux modèles basés sur des transformers, à savoir ARABERT et CAMeLBERT. Également, un modèle d'ensemble a été mis en place pour obtenir des performances plus raisonnables.

Mots-clés: Analyse de sentiments en arabe ; apprentissage automatique ; apprentissage ensembliste ; apprentissage profond ; BERT ; fouille de données ; Intelligence artificielle ; Modèle de langage ; NLP; réseaux de neurones ; traitement de texte ; Transformers.

Abstract: The digital and technological revolution has helped the Internet to become one of the most changing and fast-growing technologies. The total number of Nombre d'Internet users has been growing steadily. Globally, Internet users communicate, share content, and express their opinions or sentiments on a wide range of topics in discussion groups, blogs, forums, and other public websites. Hence, the need to detect and analyze opinions in different domains, such as politics, health, and marketing.

In recent years, sentiment analysis has gained momentum as a research area. This task aims to identify the opinion expressed in a subjective statement. An opinion is a subjective expression describing personal thoughts and feelings. These thoughts and feelings can be assigned a certain sentiment. The most studied sentiments are positive, neutral, and negative. Although several works addressed sentiment analysis, most of these works have focused on treating texts in English. Thus, studies provided in the Arabic language still require more attention to take better advantage of the richness of this language.

To this end, in this study, we have introduced an Arabic sentiment analysis dataset consisting of Arabic book reviews collected from Goodreads and associated with their related labels. Then, we have provided an overview of the major studies that have addressed sentiment analysis in Arabic using deep learning. Based on the findings of the review we have conducted, it has been revealed that various models have been used. Some examples of these are convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, bidirectional recurrent neural networks (Bi-RNNs), deep neural networks (DNNs), gated recurrent units (GRUs), as well as attention-based neural networks. These models have explored sentiment analysis in this morphologically complex language at various levels, namely sentence, document, and aspect levels. At the sentence level, it has been found that LSTM approaches outperform other deep learning models such as CNN on most datasets. At the document level, CNN

has been reported to give the best results, followed by LSTM. However, at the aspect level, it has been demonstrated that LSTM is the most commonly used model by researchers.

In this study, we evaluated the impact of combining the deep learning model BiL- STM (Bidirectional Long Short-Term Memory) with the AraGPT (Generative Pre-trained Transformer) model for sentiment analysis in Arabic. By merging these two powerful natural language processing approaches, our goal is to enhance the performance of sentiment analysis in the context of Arabic. The BiLSTM model is known for capturing long-term dependencies in a sequence, while the AraGPT model excels in text generation and contextual understanding. Leveraging their respective strengths, we aimed to achieve a more accurate and nuanced analysis of sentiments expressed in Arabic texts. The preliminary results of this study are promising, opening new perspectives for a better understanding of emotions and opinions in this complex language.

Furthermore, to make use of the power of transformer language models, we have experimented with two transformer-based models, namely AraBERT and CAMeLBERT. In addition, an ensemble model has been implemented to achieve a more reasonable performance.

Keywords: Arabic sentiment analysis; Artificial intelligence; BERT; data mining; deep learning; ensemble learning; language model; neural networks; NLP; text mining; machine learning; transformers.