AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

Mme (elle) : **EL-ALAMI Fatima-Zahra**

Soutiendra **: le 23/01/2021** à **10h**

Lieu : **Salle de réunion géologie**

Une thèse intitulée :

Deep learning models for arabic text categorization: Application to offensive language detection

En vue d'obtenir le Doctorat

FD : Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité : Informatique

Devant le jury composé comme suit :

	NOM ET PRENOM	GRADE	ETABLISSEMENT
Président	Pr. SATORI Khalid	PES	Faculté des Sciences Dhar El Mahraz - Fès
Directeur de thèse	Pr. EN NAHNAHI Noureddine	PH	Faculté des Sciences Dhar El Mahraz - Fès
Co-Directeur	Pr. OUATIK EL ALAOUI Said	PES	Ecole Nationale des Sciences Appliquées - Kenitra
Rapporteurs	Pr. FRIKH Bouchra	PES	Ecole Supérieure de Technologie - Fés
	Pr. DRISSI EL MALIANI Ahmed	PH	Faculté des Sciences - Rabat
	Pr. GAHI Youssef	PH	Ecole Nationale des Sciences Appliquées - Kenitra
Membres	Pr. AOURAGH Si Lhoussain	PH	Faculté des Sciences Juridiques, Economiques et Sociales - Salé
	Pr. ZINEDINE Ahmed	PES	Faculté des Sciences Dhar El Mahraz - Fès

Résumé :

Chaque jour, une énorme quantité de données textuelles en langue Arabe est générée sur le Web. Ces données contiennent des connaissances importantes de différents types (par exemple, scientifiques, médicales ou financières). Par conséquent, le développement des systèmes de catégorisation automatique est devenu incontournable et cruciale afin d'explorer et retrouver l'information pertinente. Des approches classiques basées sur l'apprentissage automatique ont été proposées pour effectuer la catégorisation du texte arabe. La plupart de ces approches sont basées sur les représentations sac des mots ou latentes. Cependant, ces dernières représentations ont des inconvénients, tels que le manque de sémantique et la dimensionnalité élevée de l'espace de représentation.

Les modèles basés sur l'apprentissage profond surpassent les approches classiques basées sur l'apprentissage automatique et surmontent leurs lacunes dans diverses tâches de classification de texte, y compris l'analyse des sentiments, la catégorisation des sujets et la réponse aux questions. Dans ce travail, nous abordons la tâche de catégorisation de texte arabe dans une perspective d'apprentissage profond et, dans ce contexte, nous allons entamer les questions de recherche suivantes :

1. Quelles architectures d'apprentissage en profondeur sont les plus appropriées pour la catégorisation de texte en langue arabe ?
2. Comment l'apprentissage profond peut-il aider à traiter une langue difficile comme l'arabe ?
3. Dans quels scénarios l'apprentissage profond est-il faisable pour les représentations textuelles arabes ?

Afin de répondre à ces questions de recherche, nous apportons quatre contributions au niveau de la représentation et la catégorisation de texte en langue arabe. La première contribution consiste à représenter des documents arabes à l'aide d'un Auto-encodeur profond. Notre choix est justifié par le fait que l'Auto-encodeur permet à reproduire une représentation textuelle précise avec dimension réduite. De plus, notre méthode a permis de considérer la sémantique des documents en combinant à la fois la sémantique implicite et explicite pour impulser les performances de la catégorisation. Dans la deuxième contribution, nous explorons et évaluons plusieurs modèles neuronaux profonds ainsi que leurs combinaisons pour la catégorisation de texte en langue arabe. De plus, nous intégrons un processus de retrofitting pour capturer davantage la sémantique du texte. Dans la troisième contribution, nous étudions l'apprentissage par transfert basé sur Bidirectional Encoder Representations from Transformers, à savoir BERT, pour la catégorisation de texte en langue arabe. L'idée sous-jacente consiste à représenter le texte arabe à l'aide de AraBERT, puis effectuer le fine-tuning sur la tâche de catégorisation de texte. Les résultats indiquent que l'apprentissage par transfert est une stratégie puissante pour améliorer l'efficacité de la catégorisation. Dans la dernière contribution, nous proposons une méthode de détection de langage offensive dans un contexte multilingue (MOLD). L'idée principale est de représenter les tweets à l'aide de modèles BERT et par la suite les affiner sur la tâche MOLD. De plus, pour gérer le multilinguisme, nous explorons différentes approches, telles que l'approche multilingue conjointe et l'approche basée sur la traduction.

La validation expérimentale de toutes les contributions apportées dans le cadre de cette thèse est effectuée en utilisant les collections OSAC et SOLID.

Mots clés :

Catégorisation de Texte, Représentation Textuelle, Traitement Automatique de la Langue Arabe, Apprentissage Profond, Représentations Distribuées des Mots et des documents, Réseaux Neuronaux

DEEP LEARNING MODELS FOR ARABIC TEXT CATEGORIZATION: APPLICATION TO OFFENSIVE LANGUAGE DETECTION

Abstract:
Every day, an enormous amount of Arabic textual data is generated on the Web. These data contain virtually important knowledge of different types (e.g., scientific, medical, or financial). Therefore, the ability to automatically and efficiently classify these data to explore the knowledge inside is becoming crucial. Classical machine learning-based approaches were proposed to tackle the Arabic text categorization. Most of these approaches are based on the bag-of-words or latent representations. However, these latter representations suffer from several drawbacks, such as the lack of semantic and curse of dimensionality.
Deep learning-based models have surpassed classical machine learning-based approaches and overcame their shortcomings in various text classification tasks, including sentiment analysis, topic categorization, and question answering. In this work, we tackle the task of Arabic text categorization from the deep learning perspective and, within this context, we address the following research questions:

1. Which deep learning architectures are most appropriate for Arabic text categorization?
2. How can deep learning help in dealing with a challenging language like Arabic?
3. In which scenarios is deep learning feasible for Arabic text representations?

In pursuit of answering these research questions, we propose four contributions to improve Arabic content representation and categorization. The first contribution consists of representing Arabic documents using a deep Autoencoder. This latter is motivated by the fact that the Autoencoder allows to produce precise and low dimensional text representation. Moreover, we combine both implicit and explicit semantics to boost the categorization performances. In the second contribution, we explore and evaluate several deep neural models and their combinations for Arabic text categorization. Additionally, we investigate the retrofitting technique to capture further text semantics. The third contribution investigates the transfer learning based on Bidirectional Encoders Representation from Transformers, namely AraBERT, for Arabic text categorization. It relies on representing Arabic text using Arabic BERT and then performing the fine-tuning on text categorization task. Results indicate that transfer learning is a powerful strategy to boost categorization efficiency. The last contribution is an offensive language detection method in a multilingual context (MOLD). The main idea is to represent tweets using different BERT models and then fine-tuning these representations on the MOLD task. Moreover, to handle multilingualism, we explore different approaches, such as the joint-multilingual approach and the translation-based one.
The experimental validation of all the proposed contributions is performed using OSAC collections and the semi-supervised offensive language identification dataset (SOLID).

Key Words:
Text Categorization, Text Representation, Arabic Natural Language Processing, Deep Learning, FastText, Word and Document Embeddings, Neural Networks, Social Media.