



AVIS DE SOUTENANCE THESE DE DOCTORAT

Présentée par

Mr : MOHAMMED BEKKALI

Spécialité : Informatique

Sujet de la thèse : Contribution à l'amélioration de la représentation des textes courts en langue arabe : Analyse du contenu des réseaux sociaux.

Formation Doctorale : Sciences de l'ingénieur Sciences Physiques, Mathématiques et Informatique.

Thèse présentée et soutenue le vendredi 12 juillet 2019 à 15h30 au centre des conférences devant le jury composé de :

Nom Prénom	Titre	Etablissement	
Abderrahmane SBIHI	PES	Ecole Nationale des Sciences Appliquées de Tanger	Président
Abderrahim ELQADI	PES	Ecole Supérieure de Technologie de Salé	Rapporteur
Hassan BADIR	PES	Ecole Nationale des Sciences Appliquées de Tanger	Rapporteur
Abderrahim BENABBOU	PH	Faculté des Sciences et Techniques de Fès	Rapporteur
Noureddine CHENFOUR	PES	Faculté des Sciences Dhar El Mehraz de Fès	Examineur
Nour Eddine RAISS	PES	Faculté des Sciences Dhar El Mehraz de Fès	Examineur
Abdelhamid EL JIHAD	PH	Institut des études et de Recherches pour l'Arabisation Rabat	Examineur
Abdelmonaime LACHKAR	PES	Ecole Nationale des Sciences Appliquées de Tanger	Directeur de thèse

Laboratoire d'accueil : Laboratoire ISA.

Etablissement : Ecole Nationale des Sciences Appliquées de Fès.



Titre de la thèse : Contribution à l'amélioration de la représentation des textes courts en langue arabe : Analyse du contenu des réseaux sociaux.

Nom du candidat : Mohammed BEKKALI

Spécialité : Informatique

Résumé de la thèse

L'utilisation massive des appareils mobiles, tels que les smartphones et les tablettes, a permis une augmentation considérable de la génération de données par les utilisateurs. De nos jours, environ un milliard d'utilisateurs interagissent quotidiennement sur les réseaux sociaux en ligne, où ils partagent des informations et discutent d'une grande variété de sujets. Ces réseaux sociaux permettent aux utilisateurs de publier généralement un texte court, 280 caractères par exemple dans le cas de Twitter. Les textes courts se distinguent des documents traditionnels par leurs brièveté et le manque d'informations contextuelles, en outre la langue arabe complique encore plus ces problèmes vu la richesse de sa morphologie. Par conséquent, les techniques traditionnelles de représentation de texte seront inadéquates et peuvent avoir un impact négatif sur la performance de toute application de fouille de texte. Pour surmonter les problèmes liés au texte court, la majorité des propositions tendent à apporter plus de sémantique au contenu initial à partir de la collection de données elle-même ou bien provenant d'une base de connaissance externe.

Le but de cette thèse est d'analyser l'efficacité de l'enrichissement de la représentation de texte court en langue arabe, en particulier sur deux applications parmi les plus répondues, la catégorisation de texte court et l'analyse des sentiments (dite aussi fouille d'opinions). Nous avons exploité la théorie des ensembles approximatifs (*Rough Set Theory - RST*) afin d'enrichir et de renforcer les liens sémantiques entre les textes courts d'une même collection de données ; par la suite nous avons apporté plus de sémantique à partir du Web dans le but de cerner et d'identifier le contexte d'un texte court. Par la suite, nous avons mis l'accent sur l'enrichissement du texte court en la langue arabe en utilisant une représentation conceptuelle au lieu de celle basée sur les termes en intégrant des bases de connaissances externes. Dans ce sens nous avons proposé, une première contribution pour pallier au problème de désambiguïsons lexicale à travers une mesure de similarité dédiée pour le texte court, une deuxième contribution pour surmonter le problème de taux de couverture de la langue arabe dans les bases de connaissances à travers la modélisation par thèmes (*Topic Modeling*).

Finalement, pour tester et valider nos propositions, nous avons mené une série d'expérimentations, les résultats obtenus illustrent l'intérêt de nos contributions.

Mots clés : Langue Arabe ; Texte Court ; Réseaux Sociaux ; Enrichissement ; Représentation basé Termes ; Représentation basé Concepts ; Mesure de Similarité.