



AVIS DE SOUTENANCE DE THESE

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

Mr **EL MOUSSAOUI Taoufiq**

Soutiendra : le Samedi 21/06/2025 à 10H00

Lieu : Centre des Etudes Doctorales - USMBA - Amphi 1

Une thèse intitulée :

«Machine Learning and Text Mining: Extraction de l'information à partir
d'un jugement »

En vue d'obtenir le **Doctorat**

FD : **Sciences et Technologies de l'Information et de la Communication**
Spécialité : **Informatique**

Devant le jury composé comme suit :

Nom et prénom	Etablissement	Grade	Qualité
EL BEQQALI Omar	Faculté des Sciences Dhar EL Mahraz, Fès	PES	Président
SABBANE Mohamed	Faculté des sciences, Meknès	PES	Rapporteur & Examineur
MERRAS Mostafa	Ecole Supérieure de Technologie, Meknès	MCH	Rapporteur & Examineur
AMAKDOUF Hicham	Faculté des sciences Dhar El Mahraz, Fès	MCH	Rapporteur & Examineur
BOUMHIDI Jaouad	Faculté des sciences Dhar El Mahraz, Fès	PES	Examineur
RIFFI Jamal	Faculté des sciences Dhar El Mahraz, Fès	MCH	Examineur
LOQMAN Chakir	Faculté des sciences Dhar El Mahraz, Fès	PES	Directeur de thèse



Résumé :

L'extraction d'information automatise l'identification et l'extraction d'informations structurées à partir de sources de données non structurées ou semi-structurées. Dans le domaine juridique, cela implique l'extraction d'éléments clés tels que les dispositions légales, les parties impliquées, les dates et les décisions, ainsi que la classification des documents et l'identification des relations entre les entités. En raison de la complexité du langage juridique—caractérisé par une terminologie spécialisée et des structures variées—l'extraction manuelle est à la fois chronophage et sujette à des erreurs. En s'appuyant sur le traitement du langage naturel et l'apprentissage profond, cette thèse aborde le défi de l'extraction d'information à partir de jugements juridiques marocains, visant à améliorer l'efficacité, à enrichir la recherche juridique et à accroître l'accessibilité et la transparence du système judiciaire.

La question de recherche centrale qui guide ce travail est la suivante : Comment l'information peut-elle être extraite efficacement des jugements juridiques, représentée sous une forme structurée et exploitée pour des applications pratiques? Pour y répondre, nous explorons des sous-questions liées aux techniques d'extraction, aux méthodes de représentation et aux applications pratiques.

Les contributions clés de cette thèse incluent : (1) quatre architectures d'apprentissage profond pour la classification de documents juridiques arabes, adaptables à différents types de textes juridiques ; (2) un modèle avancé de reconnaissance d'entités nommées en arabe, capable d'extraire à la fois des entités plates et imbriquées, tout en tenant compte des complexités linguistiques de la langue arabe ; (3) deux modèles d'apprentissage profond pour l'extraction d'information juridique en arabe, permettant une représentation structurée des éléments juridiques clés ; (4) une évaluation systématique des schémas d'annotation sur la performance de la reconnaissance d'entités nommées en arabe à travers des ensembles de données généraux et spécifiques à un domaine ; (5) un système basé sur des graphes de connaissances pour structurer les entités et les relations juridiques ; et (6) l'automatisation de trois processus juridiques marocains critiques, rationalisant les flux de travail et améliorant la précision des décisions.

Cette recherche fait progresser l'extraction d'information à partir de documents juridiques arabes et démontre son applicabilité à des processus juridiques réels, tels que la mise à jour des casiers judiciaires, la préservation de la confidentialité des données et l'indemnisation des accidents de la route. Les corpus juridiques arabes construits fournissent des ressources précieuses pour faire avancer les applications de traitement du langage naturel spécifiques au domaine, posant les bases pour des recherches futures sur la classification de documents juridiques arabes, la reconnaissance d'entités nommées en arabe et l'automatisation juridique pilotée par l'intelligence artificielle.

Mots clés :

Traitement du Langage Naturel, Fouille de Textes, Extraction d'Information, Reconnaissance d'Entités Nommées en Arabe, Apprentissage Profond, Classification de Documents Juridiques, Graphes de Connaissances, Automatisation Juridique.



Machine Learning and Text Mining : Extracting information from a judgment

Abstract :

Information extraction automates the identification and extraction of structured information from unstructured or semi-structured data sources. In the legal domain, this involves extracting key elements like legal provisions, involved parties, dates, and decisions, as well as classifying documents and identifying relationships between entities. Due to the complexity of legal language—characterized by specialized terminology and varied structures—manual extraction is time-consuming and error-prone. Leveraging natural language processing and deep learning, this thesis addresses the challenge of information extraction from Moroccan legal judgments, aiming to improve efficiency, enhance legal research, and increase the accessibility and transparency of the justice system.

The central research question guiding this work is: "How can information be effectively extracted from legal judgments, represented in a structured form, and leveraged for practical applications?" To answer this, we explore sub-questions related to extraction techniques, representation methods, and practical applications.

The key contributions of this thesis include: (1) four deep learning architectures for Arabic legal document classification, adaptable across legal text types; (2) an advanced Arabic named entity recognition model capable of extracting both flat and nested entities, addressing Arabic linguistic complexities; (3) two deep learning models for Arabic legal information extraction, enabling structured representation of key legal elements; (4) a systematic evaluation of annotation schemes on Arabic named entity recognition performance across general and domain-specific datasets; (5) a knowledge graph-based system for structuring legal entities and relationships; and (6) the automation of three critical Moroccan legal processes, streamlining workflows and enhancing decision-making accuracy.

This research advances information extraction from Arabic legal documents and demonstrates its applicability to real-world legal processes, such as criminal record updates, data privacy preservation, and road accident indemnifications. The constructed Arabic legal corpora provide valuable resources for advancing domain-specific natural language processing applications, laying the foundation for future research in Arabic legal document classification, Arabic named entity recognition, and artificial intelligence driven legal automation.

Key Words :

Natural Language Processing, Text Mining, Information Extraction, Arabic Named Entity Recognition, Deep Learning, Legal Document Classification, Knowledge Graphs, Legal Automation.