

République Algérienne Démocratique et Populaire  
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique  
Université Benyoucef BENKHEDDA-Alger1



Faculté des Sciences  
Département Informatique

---

Projet de Fin d'Etudes pour l'obtention du diplôme de  
Master en Informatique

Option : Analyse et Sciences de Données

Thème

---

Le text mining pour la classification  
automatique des documents textuels

---

**Encadré par :**

-Dr. BESSAI Fatma Zohra. CERIST

**Réalisé par :**

- ANNOU Farah

- BOUALBANI Lamis

2021/2022

# REMERCIEMENTS :

Avant tout propos, nous remercions Dieu le tout puissant qui nous a donné la sagesse et la santé pour faire ce modeste travail.

C'est avec un grand plaisir que nous exprimons notre gratitude et nos sincères remerciements à notre promotrice : **Mme Fatma Zohra BESSAI** pour son orientation et encadrement dans l'élaboration de ce projet de fin d'études .

Toutes nos reconnaissances sont adressées à tous les enseignants qui nous ont suivis infatigablement durant tout notre cursus universitaire.

Nous tenons à exprimer tout au fond de nos coeurs les reconnaissances à nos familles qui nous ont offert toujours un appui sûr par leurs soutiens et leurs encouragements. Nos vifs remerciements vont également à tous ceux qui ont contribué de loin ou de près à la réalisation de ce travail.

# *Dédicaces*

C'est avec profonde gratitude et sincères mots, que je dédie ce modeste travail de fin d'études.

*à ma très chère mère : Rabea ...*

Quoi que je fasse ou je dis, je ne saurai point te remercier comme il se doit. Ton affection me couvre, ta bienveillance me guide et ta présence à mes cotés a toujours été ma source de force pour affronter les différents obstacles, que dieu te garde pour moi.

*à mon très cher père : Rabah ...*

tu as toujours été à mes cotés pour me soutenir et m'encourager, tu as sacrifié ta vie pour ma réussite et tu m'as éclairé le chemin par tes conseils judicieux, j'espère qu'un jour, je pourrai te rendre un peu de ce que tu as fais pour moi, que dieu te prête bonheur et longue vie...

*à mon très cher frère Zakaria et ma soeur Chahinez ...*

Puisse Dieu vous donne santé, bonheur, courage, et surtout réussite.

Je dédie aussi ce travail à mes deux grandes mères, que Dieu leur donne une longue et joyeuse vie,

A toute ma famille, mes amis Farid, Lina, Hiba, Salma, Maria, Rania, Yasmine, Fares...

Tout les professeurs qui m'ont enseigner ,

A ma binôme de travail : **Farah** pour son soutien moral, sa patience tout au long de ce mémoire.

et à tous ceux qui nous sont chers.

## **Lamis...**

# *Dédicaces*

Avec l'expression de ma reconnaissance, je dédie ce modeste travail à ceux qui, quels que soient les termes embrassés, je n'arriverais jamais à leur exprimer mon amour sincère.

A l'homme, mon précieux offre du dieu, qui doit ma vie, ma réussite et tout mon respect :

mon cher père *Noureddine*.

A la femme qui a souffert sans me laisser souffrir, qui n'a jamais dit non à mes exigences et qui n'a épargné aucun effort pour me rendre heureuse : mon adorable mère *Sada*.

A ma chère soeur Sarah, et mes frères : Billel ,Amine, Abdelhakim et Oussama , ma belle soeur Louiza et ma petite nièce Eyline qui n'ont pas cessée de me conseiller, encourager et soutenir tout au long de mes études.

Que Dieu les protège et leurs offre la chance et le bonheur.

A ma grand mère , mes oncles et mes tantes. Que Dieu leur donne une longue et joyeuse vie.

A tous les cousins et les amis surtout Hanane et Maria qui m'ont soutenu. Merci pour leurs amours et leurs encouragements.

Sans oublier ma binôme **Lamis** pour son soutien moral, sa patience et sa compréhension tout au long de ce projet.

## **Farah...**

## Abstract

Faced with the increase in information available online and the number of electronic documents written in natural language, the categorization or automatic classification of texts is becoming more and more essential as a key technology in the management of intelligence within of the company.

The process of classifying a collection of texts consists of labeling each text with one or more predefined classes (categories) through a machine learning algorithm.

The objective of our work is to study the techniques of Text Mining and to propose an approach based on Text Mining and machine learning for the analysis and classification of textual documents. After a comparative study between the approach based on classical models (KNN, SVM, Naïve of Bayes, Logistic regression and Decision trees) and the neural approach (ANN, CNN, RNN and bidirectional LSTM), we have chosen the bidirectional LSTM neural network as a solution for the classification of textual documents.

**Keywords:** *Text mining, machine learning, deep learning, data analysis, categorization, classification, text documents.*

## Résumé

Face à l'accroissement de l'information disponible en ligne et le nombre de documents électroniques rédigés en langue naturelle, la catégorisation ou classification automatique de textes s'impose de plus en plus comme une technologie clé dans la gestion de l'intelligence ausein de l'entreprise.

Le processus de classification d'une collection de textes consiste à étiqueter chaque texte avec une ou plusieurs classes (catégories) prédéfinies par le biais d'un algorithme d'apprentissage automatique (Machine Learning).

L'objectif de notre travail est d'étudier les techniques du Text Mining et de proposer une approche basée sur le Text Mining et l'apprentissage automatique pour l'analyse et la classification de documents textuels. Après une étude comparative entre l'approche basée sur les modèles classiques (KNN, SVM, Naïf de Bayes, Régression logistique et les arbres de décision) et l'approche neuronale (ANN, CNN, RNN et LSTM bidirectionnel), nous avons choisi le réseau de neurone LSTM bidirectionnel comme solution pour la classification des documents textuels.

**Les mots clés :** *Fouille de texte, Apprentissage automatique, Apprentissage profond, Analyse de données, catégorisation, classification, documents textuels.*

## ملخص

في ظل زيادة حجم المعلومات المتاحة عبر الانترنت وعدد المستندات الالكترونية المكتوبة بلغة طبيعية، أصبح التصنيف او التصنيف التلقائي للنصوص بشكل متزايد تقنية رئيسية في ادارة الذكاء داخل الشركة. تتكون عملية تصنيف مجموعة من النصوص من تسمية الهدف من عملنا هو دراسة تقنيات التنقيب عن النص واقتراح نهج قائم على التنقيب على المعطيات النصية والتعلم الآلي لتحليل وتصنيف الوثائق النصية. بعد دراسة مقارنة بين النهج القائم على النماذج الكلاسيكية:

(KNN, SVM, Naïve de bayes, Régression logistique, Arbres de décision)

والنهج القائم على التعلم العميق:

(ANN, CNN, RNN, LSTM Bidirectionnel),

اخترنا كحل لتصنيف الوثائق النصية: LSTM Bidirectionnel

**الكلمات المفتاحية :** التعلم الآلي، التعلم العميق، التنقيب في النصوص، تحليل المعطيات، التصنيف، وثائق نصية

# Sommaire

<b>Introduction générale</b>	<b>1</b>
<b>1 Le traitement du langage naturel</b>	<b>3</b>
1.1 Introduction . . . . .	3
1.2 Définition . . . . .	3
1.3 L'apprentissage automatique pour le traitement de langage naturel . .	4
1.4 Le processus du traitement du langage naturel . . . . .	5
1.5 Les techniques de NLP . . . . .	7
1.5.1 Analyse syntaxique . . . . .	7
1.5.2 Analyse sémantique . . . . .	7
1.6 Les cas d'utilisation du NLP . . . . .	8
1.7 Les principaux modèles de NLP . . . . .	9
1.8 Conclusion . . . . .	9
<b>2 Data mining et Text mining</b>	<b>10</b>
2.1 Introduction . . . . .	10
2.2 Data mining . . . . .	11
2.2.1 Introduction . . . . .	11
2.2.2 Définition du Data Mining . . . . .	11
2.2.3 Multidisciplinarité du Data Mining . . . . .	12
2.2.4 Le processus du Data Mining . . . . .	12
2.2.5 Les méthodes du Data Mining . . . . .	14
2.2.6 Domaines d'application du Data Mining . . . . .	14
2.2.7 Conclusion . . . . .	16

2.3	Text Mining . . . . .	16
2.3.1	Introduction . . . . .	16
2.3.2	Définition du Text Mining . . . . .	16
2.3.3	Les avantages du Text Mining . . . . .	17
2.3.4	Processus de Text Mining . . . . .	17
2.3.5	Techniques liées au Text Mining . . . . .	18
2.3.6	Relation entre Text Mining et apprentissage automatique . . . . .	19
2.3.7	Méthodes du Text Mining . . . . .	19
2.3.8	Conclusion . . . . .	19
2.4	Conclusion . . . . .	20
<b>3</b>	<b>La classification de textes</b>	<b>21</b>
3.1	Introduction . . . . .	21
3.2	Apprentissage automatique . . . . .	21
3.2.1	L'apprentissage automatique supervisé . . . . .	23
3.2.2	L'apprentissage automatique non supervisé . . . . .	24
3.2.3	L'apprentissage par Renforcement . . . . .	25
3.3	La classification Automatique de Texte . . . . .	26
3.3.1	Définition . . . . .	26
3.3.2	Définition Formelle de la classification de textes . . . . .	27
3.3.3	Automatisation de la classification . . . . .	27
3.3.4	Les approches de la classification automatique des textes . . . . .	28
3.4	Points forts de la classification de textes . . . . .	31
3.5	Le processus de catégorisation de textes . . . . .	33
3.6	Problèmes rencontrés lors de la classification de textes . . . . .	48
3.7	Conclusion . . . . .	51
<b>4</b>	<b>Les algorithmes d'apprentissage automatique</b>	<b>52</b>
4.1	Introduction . . . . .	52
4.2	Les algorithmes les plus utilisés dans la classification de textes . . . . .	52
4.2.1	K plus proche voisins . . . . .	53
4.2.2	Les arbres de décision . . . . .	55
4.2.3	La classification bayésienne . . . . .	56



4.2.4	Machine à vecteurs de support . . . . .	57
4.2.5	La régression logistique . . . . .	59
4.2.6	Les réseaux de neurones artificiels . . . . .	61
4.3	Critères d'évaluation des modèles de classification automatique . . . . .	71
4.4	Conclusion . . . . .	74
<b>5</b>	<b>Conception du modèle de classification de textes</b>	<b>75</b>
5.1	Introduction . . . . .	75
5.2	Approche proposée . . . . .	75
5.2.1	Le prétraitement des données . . . . .	76
5.2.2	Le choix du meilleur modèle de classification . . . . .	79
5.3	Conclusion . . . . .	82
<b>6</b>	<b>Implémentation et résultats</b>	<b>83</b>
6.1	Introduction . . . . .	83
6.2	Présentation du corpus d'expérimentation . . . . .	83
6.2.1	Le dataset . . . . .	83
6.2.2	Statistiques et Visualisations . . . . .	84
6.3	Evaluation des classifieurs . . . . .	90
6.3.1	Le prétraitement des données . . . . .	91
6.3.2	Création et évaluation des classifieurs . . . . .	91
6.4	Discussion des résultats . . . . .	114
6.5	Implémentation du système de classification des documents textuels . . . . .	116
6.5.1	Outils de développement . . . . .	116
6.5.2	Environnement de travail . . . . .	120
6.5.3	Présentation du système de classification des tweets . . . . .	121
6.6	Conclusion . . . . .	131
	<b>Conclusion générale et perspectives</b>	<b>132</b>
	<b>Annexes</b>	<b>134</b>