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

REMERCIMENTS :

Avant tout propos, nous remercions Dieux 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 chére 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 frére 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 intelligencewithin of the company.

The process of classifying a collection of texts consists of labeling each text with one or morepredefined 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 surles 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, Analysede 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