

# INTRODUCTION GENERALE

Aujourd'hui, nous vivons dans un monde où l'information est disponible en grande quantité tout en étant de qualité très diverse. Internet s'enrichit continuellement de nouveaux contenus. Par exemple, les entreprises emmagasinent de plus en plus de données, le courriel devient un moyen de communication extrêmement populaire, des documents autre fois manuscrits sont aujourd'hui disponibles sous format numérique. Mais toute cette information complexe serait sans intérêt si notre capacité à y accéder efficacement n'augmentait pas elle aussi. Pour cela, nous avons besoin d'outils permettant de chercher, classer, conserver, mettre à jour et analyser les données accessibles. Il est ainsi nécessaire de proposer des systèmes afin d'accéder rapidement à l'information désirée, réduisant ainsi l'implication humaine. Un des domaines qui tente d'apporter des améliorations et de réduire la tâche de l'humain est la classification automatique de documents [01].

L'information contenue dans les documents se présente dans différentes langues [02] et selon l'apparition d'internet et le nombre important de collections de documents multilingues, il est devenu indispensable aux utilisateurs du web de trouver les documents pertinents, quelles que soient leurs langues. Ce qui a donné naissance à un nouveau domaine qui est le domaine de catégorisation des textes multilingue [03].

La catégorisation automatique de textes ; est un problème qui intéresse les chercheurs depuis relativement longtemps. On retrouve des travaux portant sur ce sujet depuis au moins le début des années 1960.

La recherche dans ce domaine est toujours très pertinente, car les résultats obtenus aujourd'hui sont encore sujets à amélioration. Pour certaines tâches, les classificateurs automatiques performant presque aussi bien que les humains, mais pour d'autres, l'écart est encore grand. Au premier abord, l'essentiel du problème est facile à saisir. D'un côté, on est en présence d'une banque de documents textuels et de l'autre, d'un ensemble prédéfini de catégories. L'objectif est de rendre une application informatique capable de déterminer de façon autonome, dans quelle catégorie classer chacun des textes, à partir de leur contenu.

Le domaine de la fouille de textes (text mining) s'est développé pour répondre à volonté à la gestion par contenu des sources volumineuses de textes. A l'heure actuelle, de nombreux logiciels de classification de textes sont disponibles, ils ont fait l'objet de publications et leurs champs d'application s'élargissent de jour en jour. En général, ces systèmes sont basés sur des algorithmes d'apprentissage automatique (approche statistique, approche syntaxique et approche connexionniste).

Nous nous intéressons ici plus particulièrement aux algorithmes d'apprentissage et nous avons utilisé deux types d'algorithmes : les arbres de décision pour la catégorisation des textes (politique, sport, économique et art) et l'algorithme de Naïve Bayésien pour l'identification de la langue de ces textes. Pour pouvoir utiliser de tels algorithmes, il est nécessaire de transformer les données, initialement en format texte, en une représentation numérique. Nous avons choisi pour ce faire, la méthode de sélection des termes les plus pertinents. Une fois ce prétraitement terminé, nous pouvons effectuer la classification à l'aide de nos algorithmes [04].

Nous résumons la présentation de notre travail ainsi :

Chapitre I : présente une introduction à notre domaine d'étude via l'explication de sujet du texte mining.

Chapitre II : résume d'une façon claire et abrégée la catégorisation automatique des textes et ses domaines d'applications.

Chapitre III: expose les différents algorithmes d'apprentissage utilisé pour la catégorisation de textes.

Chapitre IV : met l'accent sur la classification de textes en utilisant l'algorithme arbre de décision.

Chapitre V : expose l'architecture du logiciel conçu et son fonctionnement, ainsi que son implémentation et quelques exemples de démonstration.