République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université 8Mai 1945 – Guelma
Faculté des sciences et de la Technologie
Département d'Electronique et Télécommunications

Mémoire de fin d'étude
pour l'obtention du diplôme de Master Académique

Domaine : **Sciences et Technologie**
Filière : **Télécommunications**
Spécialité : **Réseaux et Télécommunications**

DÉTECTION ET ANALYSE DE DIALECTE ALGERIEN UTILISÉ DANS LES MÉDIAS SOCIAUX

Présenté par :

MAIZI Randa

Sous la direction de :

Dr. ABAINIA Kheireddine

Septembre 2021

Résumé

Ce manuscrit se concentre sur la détection et l'analyse de dialecte Algérien dans les commentaires Facebook, pour cette étude un corpus DZDC12 a été développé à partir de pages et de groupes sur le réseau Facebook avec 2400 commentaires écrits avec l'arabes et le romain. Dans ce travail on va effectuer une série de tests basée sur des algorithmes de pointe (des outils de reconnaissance de langue et des classificateurs). Les résultats sont acceptables, mais l'algorithme de reconnaissance peut être bonifié par des recherches supplémentaires.

ملخص

تركز هذه المخطوطة على الكشف عن اللهجة الجزائرية وتحليلها في تعليقات فيسبوك، ومن أجل هذه الدراسة تم تطوير مجموعة DZDC12 من صفحات ومجموعات على شبكة Facebook مع 2400 تعليق مكتوب باللغتين العربية والرومانية. في هذا العمل، سنقوم بإجراء سلسلة من الاختبارات بناءً على أحدث الخوارزميات (أدوات التعرف على اللغة والمصنفات). النتائج مقبولة، ولكن يمكن تحسين خوارزمية التعرف من خلال مزيد من البحث.

Abstract

This manuscript focuses on the detection and analysis of Algerian dialect in Facebook comments, for this study a DZDC12 corpus was developed from pages and groups on the Facebook network with 2400 comments written with Arabic and Roman. In this work we will perform a series of tests based on state-of-the-art algorithms (language recognition tools and classifiers). The results are acceptable, but the recognition algorithm can be improved by further research.

Résumé i
ملخص ii
Abstract iii
Liste de figures iv
Liste de tableaux v
Liste des Acronymes vi
Introduction Générale vii
Chapitre 1 : Catégorisation des textes
1.1. Introduction 1
1.2. Définitions de la classification automatique 1
1.3. Historique 3
1.4. Systèmes de classification et vocabulaire utilisé 4
1.4.1. Notion de class dans les systèmes de classification 4
1.4.2. Supervisé ou clustering 5
1.4.3. Non supervisé ou catégorisation 6
1.4.4. Classification supervisée vs non supervisée 6
1.5. Différents contextes de classification 7
1.5.1. Classification bi-classe et multi-classe 7
1.5.1.1. Classification bi-classe 7
1.5.1.2. Classification multi-classe disjointe 8
1.5.1.3. Classification multi-classe 8
1.5.2. Catégorisation déterministe et floue 8
1.5.2.1. Catégorisation déterministe 8
1.5.2.2. Catégorisation floue ou basée sur le ranking 8
1.6. Catégorisation textuelle 10
1.6.1. Objectifs et intérêts 10
1.6.2. Démarche de la catégorisation textuelle 11
1.6.3. Quelques technologies utilisées dans la catégorisation textuelle 12
1.6.3.1 Les K plus proches voisins (KNN) 13
1.6.3.2 Les arbres de décision 14
1.6.3.3 Machines à supports de vecteur (SVM) 16
1.6.3.4 Réseaux neurones 17
1.6.3.5 Les réseaux bayésiens 19
1.7. Conclusion 19
Chapitre 2 : Détection de la langue
2.1. Introduction 20
2.2. Concepts fondamentaux des langues 20
2.2.1. Définition de la langue 20
2.2.2. Historique des langues 21
2.2.3. Classification des langues 23
2.2.3.1. Par nombre de langues 23
2.2.3.2. Par famille de langues 23
2.2.3.3. Par nombre de locuteurs dans chaque langue 24
2.2.3.4. Par dénomination 27
2.2.4. Dialectes 27
2.2.4.1. Définitions 27
2.2.4.2. Différence entre dialecte et langue formelle 28
2.2.5. Phénomènes linguistiques 28
2.2.5.1. Code-switching 29

2.2.5.2. Romanisation des langues non Latines 30
2.2.5.3. Argos et abréviations 31
2.3. Identification automatique de la langue 33
2.3.1. Identification de la langue des textes formels 33
2.3.1.1. Approches statistiques 33
2.3.1.2. Approches à base d'apprentissage 37
2.3.2. Identification de la langue des textes des réseaux sociaux 38
2.3.2.1. Approches statistiques 38
2.3.2.2. Approches à base d'apprentissage 40
2.4. Identification automatique des dialectes 41
2.4.1. Dialectes des langues Latines 41
2.4.2. Dialectes Arabes 42
2.4.3. Dialectes Algériens 44
2.5. Conclusion 45
Chapitre 3 : Méthodologie
3.1.Introduction 46
3.2.Algorithmes de pointe 46
3.2.1. Support Vector Machines (SVM) 46
3.2.2. Naïve bayes (NB) 47
3.2.3. Neural Network (NN) 49
3.2.4. Language Identification (LangId.py) 50
3.2.5. Language Detecting (LangDetect) 51
3.3. Conclusion 52
Chapitre 4 : Résultats et expérimentation
4.1.Introduction 53
4.2.Description du corpus 53
4.3.Installation et configuration 53
4.4.Résultats expérimentaux 54
4.5.Conclusion 57
Conclusion générale 58

Liste de figures

Figure 1.1 : Position de notre problème 1
Figure 1.2 : Etapes du processus de classification automatique de texte 2
Figure 1.3 : Exemple de système de classification d'emails 5
Figure 1.4 : Schéma illustratif de fonctionnement d'un outil de classification 7
Figure 1.5 : démarche de la catégorisation de textes 12
Figure 1.6 : Les K plus proches voisins 14
Figure 1.7 : L'arbre de décision 15
Figure 1.8 : Les vecteurs à support 16
Figure 1.9 : Architecture générale d'un réseau de neurones artificiels 18
Figure 2.1: Raisons du changement de code (Code-Switching) 29
Figure 2.2: Fréquences n-gram par rang dans un document technique 35
Figure 3.1 : Séparation linéaire et non linéaire dans l'espace de données d'entrée 46
Figure 3.2 : Un réseau de neurones de base 48
Figure 4.1 : Données affichées du corpus utilisé 54

Liste de tableaux

Tableau 2.1: Principales familles de langues du monde 24
Tableau 2.2: Classement des principales langues selon leur nombre de locuteurs 26
Tableau 2.3: Exemple de romanisation et translittération de quelques langues 31
Tableau 2.4: Exemple d'argot et abréviations 32

CT : Catégorisation de Texte

NLP : Natural Language Processing

AA : Apprentissage Automatique

TREC : Text Retrieval Conference

IR: Information Retrieval

ML: Machine Learning

KNN: k-Nearest Neighbours

SVM : Support Vector Machine

MSA: Multiple Sequence Alignment

NB: Naive Bayes

NN: Neural Network

LangId.Py: A standard Language Identification tool

LangDetect: Language Detection tool

Introduction Générale

Introduction générale

A l'heure actuelle, l'information et son analyse sont devenus un pilier de vie. Avec le développement du web, ces informations sont de plus en plus accessibles et plus précises sous forme numérique.

Ces nouveaux phénomènes telle que Facebook, Twitter, YouTube…etc sont apparu à travers le monde entier et qui autorisent les gens du monde de se communiquer, de diffuser des pensées et de partager des photos et des vidéos et pleins d'autre choses. Ils sont considérés maintenant comme une partie essentielle de notre vie quotidienne. Ces derniers sont accessibles pour toutes les tranches d'âges et pour tous les niveaux d'éducation ce qui rends l'apparition des bienfaits et des inconvénients obligatoires.

En Algérie, le réseau social le plus utiliser est bien Facebook, avec le dialecte algérien comme langue principale a utilisé, ce qui nous intéresse pour ce travail dont on cherche à détecter des dialectes et des sous dialectes algériens à partir des commentaires Facebook en ligne. En fait ce réseau est la bonne source pour pouvoir collecter une base de données sur le dialecte algérien.

Ce travail, consiste à faire l'identification des régions et des wilayas d'Algérie à partir du dialecte algérien collecter. Les différentes applications de PNL peuvent avoir cette tache comme étape essentielle. Les textes arabes écrivaient en arabe où en latine se typique par l'accent mis sur les mots des textes arabes.

Donc, dans ce travail nous allons traiter quelques difficultés spécifiques. On commence d'abord par traiter des textes arabizi qui s'écrivent irrégulièrement, puis le phénomène de changement de code arabe-français et enfin les sous dialectes prise du même dialecte et qui contient plusieurs linguistiques.

Chapitre 1 :

Catégorisation Des Textes

CHAPITRE 1 : CATEGORISATION DES TEXTES

1.1. Introduction

De nos jours, la classification de textes est un domaine de recherche mature, bien établi et très actif. La classification des textes consiste à trouver une liaison fonctionnelle entre un document textuel et un ensemble de catégories, en utilisant une technique d'apprentissage automatique. Pour effectuer cette dernière il est utile de disposer un ensemble de textes catégorisés (classés, étiquetés). Alors au titre de ce chapitre, nous allons définir la catégorisation automatique des textes ainsi que ses diversités.

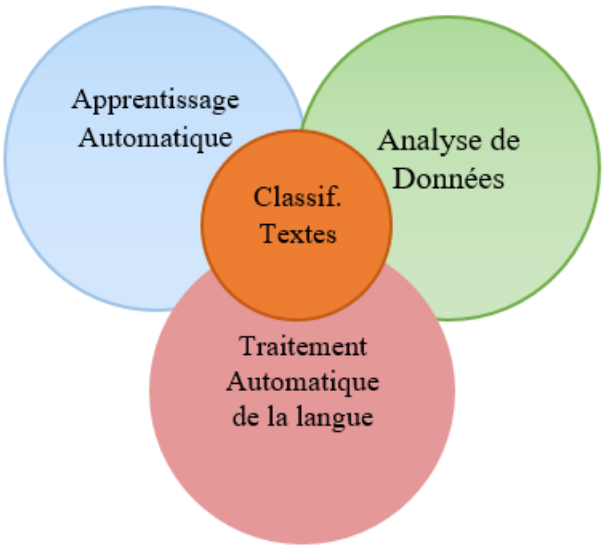


Figure 1.1 : Position de notre problème

1.2. Définitions de la classification automatique

La classification automatique est une méthode mathématique d'analyse de données, en informatique, il s'agit d'assigner un document a une ou plusieurs catégories ou classes de telle sorte que les textes d'une même classe soient le plus semblable. À l'aide du traitement du langage naturel (NLP), les classifieurs de texte peuvent analyser automatiquement le texte, puis affecter un ensemble de balises ou de catégories prédéfinies en fonction de son contenu. [1]

Le texte non structuré est partout, comme les e-mails, les conversations de chat, les sites Web et les médias sociaux, mais il est difficile d'extraire de la valeur de ces données à moins qu'elles ne soient organisées d'une certaine manière. Ceci est un processus difficile et coûteux, qui nécessite du temps et des ressources pour trier manuellement les données ou de créer des règles artisanales difficiles à maintenir.

Les classifieurs de texte avec NLP se sont avérés être une excellente alternative pour structurer les données textuelles de manière rapide, rentable et évolutive.

La classification de texte devient une partie de plus en plus importante des entreprises car elle permet d'obtenir facilement des informations à partir des données et d'automatiser les processus métier. Voici quelques-uns des exemples et des cas d'utilisation les plus courants de classification automatique de texte : [1]

- Analyse des sentiments : le processus de compréhension si un texte donné parle positivement ou négativement d'un sujet donné (par exemple : Texte orienté à des fins de surveillance d'une marque). [1]

- Détection de sujet : la tâche d'identifier le thème ou le sujet d'un morceau de texte (par exemple : savoir si un avis de produit concerne la facilité de son utilisation, le support client ou la tarification lors d'une analyse exhaustive des commentaires exprimés par clients). [1]

- Détection de langue : la procédure de détection de la langue d'un textlese donné (par exemple : savoir si un ticket de support entrant est transcrit en anglais ou dans une autre langue, et ainsi acheminer automatiquement les tickets vers l'équipe appropriée). [1]

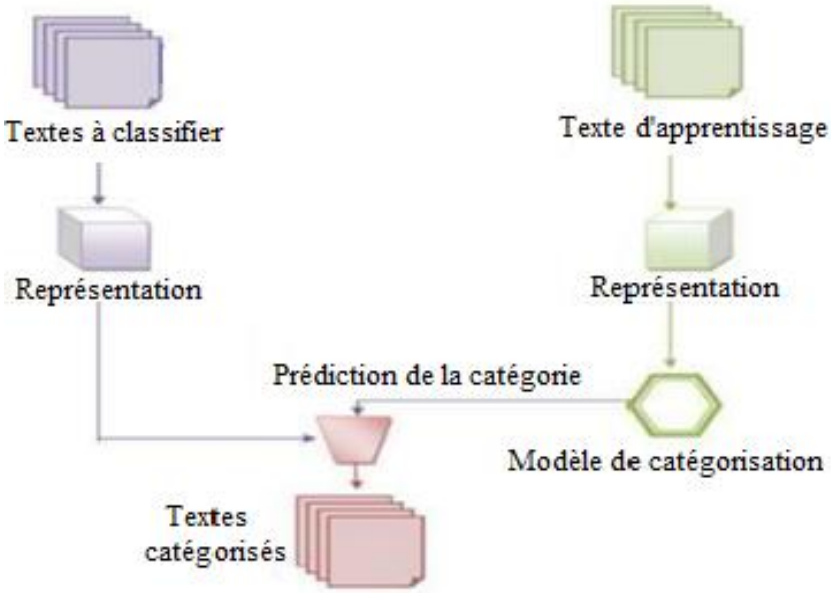


Figure 1.2 : Etapes du processus de classification automatique de texte

Source : Google image

1.3. Historique

En 1627, Gabriel Naudé propose un classement sous formes de grands thèmes, il a proposé en tout cinq thèmes sont : la théologie, la jurisprudence, l'histoire, la science et l'art. Le désir de maîtriser l'Univers se fait sentir dans la multiplication des encyclopédies. L'encyclopédie de Diderot qui a était parue entre les années 1751 et 1772 est ordonnée selon un ordre alphabétique avec des renvois associatifs, par contre l'association de Panckoucke qui a déclaré que cette dernière a parue entre 1776 à 1780 a été fait sous une organisation méthodique selon un ordre arborescent (Fayet & Scribe, 1997). Le système de classification par thème, apparu dès les débuts de l'écriture et institutionnalisé à Alexandrie conduisit à la création par Dewey, en 1876, d'un système de classification « universel ». Il s'agit d'une classification documentaire de type encyclopédique. [38]

Toutefois l'idée d'effectuer la classification de textes par des machines remonte au début des années 60 et qui a connu des progrès considérables à partir des années 90 avec l'apparition d'algorithmes beaucoup plus performants qu'auparavant.

Jusqu'au début des années 80, pour construire un classifieur, il fallait consacrer d'importantes ressources humaines à cette tâche. Plusieurs experts éditaient des règles manuellement puis les affinaient au fur et à mesure des tests. L'avènement des de l'AA s'est donc traduit par un gain de temps conséquent. Il n'est plus nécessaire par exemple de reconfigurer tout le système en cas de changement d'arborescence. Ces évolutions technologiques et algorithmes avancées font aujourd'hui de la catégorisation un outil fiable. [38]

Au début des années 1990, les travaux proviennent essentiellement de la communauté de Recherche d'Information (RI). En effet, les méthodes de numérisation, les algorithmes de classification et les méthodologies de test ont été adaptés à la CT en particulier au cours des conférences TREC (Text REtrieval Conference. http://trec.nist.gov). [2]

Il y a dix ans la communauté d'Apprentissage Automatique (AA) a mis son intérêt à ce problème et le considérer comme domaine d'application à ces algorithmes de reconnaissance des formes. Maintenant, les méthodes de numérisation de texte sont extrêmement inspirées de la recherche d'information (RI) alors que les classifieurs les plus performants sont issus de l'apprentissage automatique.

Une autre communauté composée essentiellement de statisticiens et de linguistes, traite également le problème de la classification de textes en s'appuyant sur les méthodes d'analyse de données. Le but ici n'est pas de créer un système qui classe automatiquement des documents sans intervention humaine mais d'extraire des informations synthétiques du corpus. Les problématiques traitées ici sont par exemple l'étude des genres littéraires ou la détermination de l'auteur d'un texte. [38]

1.4. Systèmes de classification et vocabulaire utilisé

L'objectif de la CT est de classer de façon automatique les documents dans des catégories qui ont été définies de bonne façon, soit préalablement par un expert. Il s'agit alors de classification supervisée ou catégorisation. Soit de façon automatique, il s'agit alors de classification non supervisée ou encore clustering. Dans ce qui suit nous allons essayer de distinguer entre les différentes variantes de classification de textes et le vocabulaire utilisé on va donc l'aborder et le définit sous étapes précises.

1.4.1. Notion de classe dans les systèmes de classification

Le mot "sujet" veut dire souvent un synonyme de concept de classe d'un système de classification. Dans ce cas, catégoriser les documents consiste à les organiser selon différents thèmes, on cite par exemple : Earn, Ship, Trade.

Le problème de la classification évolue cependant en même temps que les besoins, et s'intéresse aujourd'hui à une variété de tâches pour lesquelles les catégories ne s'expliquent pas comme des thèmes : ainsi, par exemple, la tâche de classer les documents par auteur, par genre, et donc par style, par langue, Cela dépend même si le document exprime un jugement positif ou un jugement négatif, etc., pas nécessairement un seul sujet.

Comme nous le verrons ci-dessous, une classe n'est qu'une étiquette associée à un document.

Dans la figure 1.3, un système de classification d'emails est représenté où les classes peuvent être de différentes natures (thèmes, messages provenant de certaines personnes, messages d'un certain type, etc…). [3]

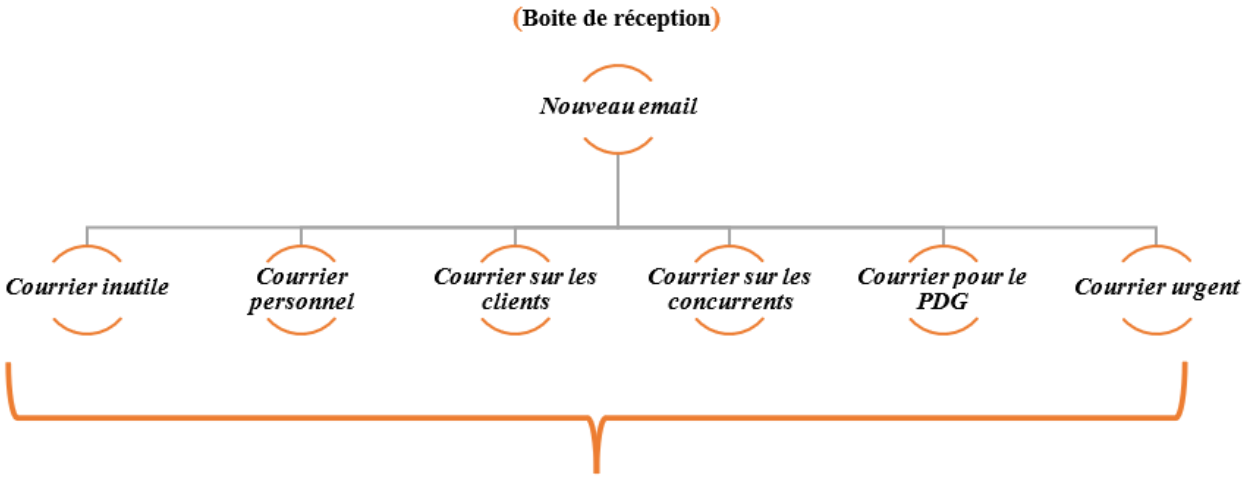


Figure 1.3 : Exemple de système de classification d'emails

Source : Mémoire de magister : Matallah Hocine

1.4.2. Supervisé ou clustering

La classification supervisée correspond au processus d'attribution d'une ou plusieurs catégories prédéfinies au texte. Elle correspond à la classification de l'apprentissage automatique et à la distinction en statistiques, tandis que la recherche d'informations utilise des termes plus stricts. Ce problème a récemment trouvé de nouvelles applications dans le domaine du traitement du langage, telles que : l'attribution de thèmes de recherche d'informations, l'aide à l'utilisateur pour l'indexation de documents, la veille technologique, le filtrage personnalisé des documents d'intérêt pour les internautes, la compréhension de leurs préférences thématiques, le routage de texte et la mise en réseau. L'amélioration de la recherche et l'organisation finale de plus en plus de sources de texte, en particulier les pages web d'aujourd'hui.

Ce problème utilise principalement des méthodes dérivées de l'apprentissage automatique et de nombreux algorithmes d'apprentissage supervisé ont été appliqués à cette méthode (Bayésien, K-plus proches voisins, Arbres de décision, Réseaux de neurones etc…). [4]

1.4.3. Non supervisé ou catégorisation

Quant aucun ensemble de catégories n'est donné au début, c'est un problème créé par le regroupement du texte en catégories avec une certaine cohérence interne. On est alors dans le contexte d'une classification non supervisée de l'apprentissage automatique. La classification non supervisée comprend la recherche automatique

d'une organisation cohérente pour un groupe de documents similaires afin d'établir un regroupement cohérent (Classe ou Cluster), ce qui correspond statistiquement au regroupement, qui est également un terme utilisé dans la recherche d'informations. Par conséquent, le clustering divise les objets (nous parlons ici de texte) en groupes sans connaître a priori leurs catégories d'appartenance. La technologie d'un tel regroupement constitue un champ de recherche très riche, qui a conduit à de nombreuses propositions, mais le recensement ne fait pas l'objet de ce document. [4]

1.4.4. Classification supervisée vs non supervisée

La classification supervisée comprend l'identification de la classe à laquelle appartient un objet en fonction de certaines caractéristiques descriptives. Cette approche permet d'affecter automatiquement les documents à des classes préexistantes. Le but est de trouver un lien fonctionnel entre le texte à classer et toutes les catégories, aussi appelé modèle prédictif. Afin d'estimer le modèle de prédiction, il doit exister un ensemble de textes préalablement étiquetés, appelé ensemble d'apprentissage, à partir duquel les paramètres du modèle de prédiction les plus efficaces peuvent être obtenus, c'est-à-dire que le moins d'erreurs sont générées dans la prédiction. Contrairement à la classification non supervisée (l'ordinateur doit découvrir le groupe de documents par lui-même), la classification supervisée suppose que la classification de documents existe déjà. Par exemple, la bibliothèque ou l'étude. [5]

L'objectif est de classer automatiquement les nouveaux documents. Par conséquent, apprenez d'abord le modèle ou le classificateur à partir de l'ensemble d'apprentissage composé de paires (objet, classe). Contrairement à la classification non supervisée, la classification supervisée permet de mesurer l'importance de chaque mot dans la classification de nouveaux documents. Par exemple, une métrique (gain d'information) calcule la catégorie type d'un terme au lieu des autres, et surtout : si un nouveau document le contient, le terme sera très discriminant. De nombreuses mesures similaires ont été développées. Enfin, contrairement à la classification non supervisée, il est facile d'évaluer ici les résultats d'une classification. Parmi les N exemples de documents classifiés, une partie des documents est utilisée pour la formation et le reste pour les tests. En phase de test, chaque document est soumis à l'algorithme de classification, on voit juste si la machine trouve la bonne classe. Bien sûr, le résultat de ce test n’est en rien garanti lorsque la machine aura à classer de nouveau document. [5]

1.5. Différents contextes de classification

Le contexte est l'ensemble des circonstances qui entourent un ou plusieurs événements. Cependant, cette définition très générale est insuffisante pour expliquer la multitude d'aspects qu'on peut relier à la notion de contexte et à ses effets particuliers.

Dans notre cas on distingue que le contexte de catégorisation est l'ensemble des circonstances qui entourent un ou plusieurs événements. Toutefois, cette définition large ne suffit pas à expliquer le large éventail d'implications que le concept de contexte et ses effets spécifiques peuvent avoir. Dans ce qui suit on va clarifier les différents contextes de classification (Classification bi-classe et multi-classe, Catégorisation déterministe et floue).

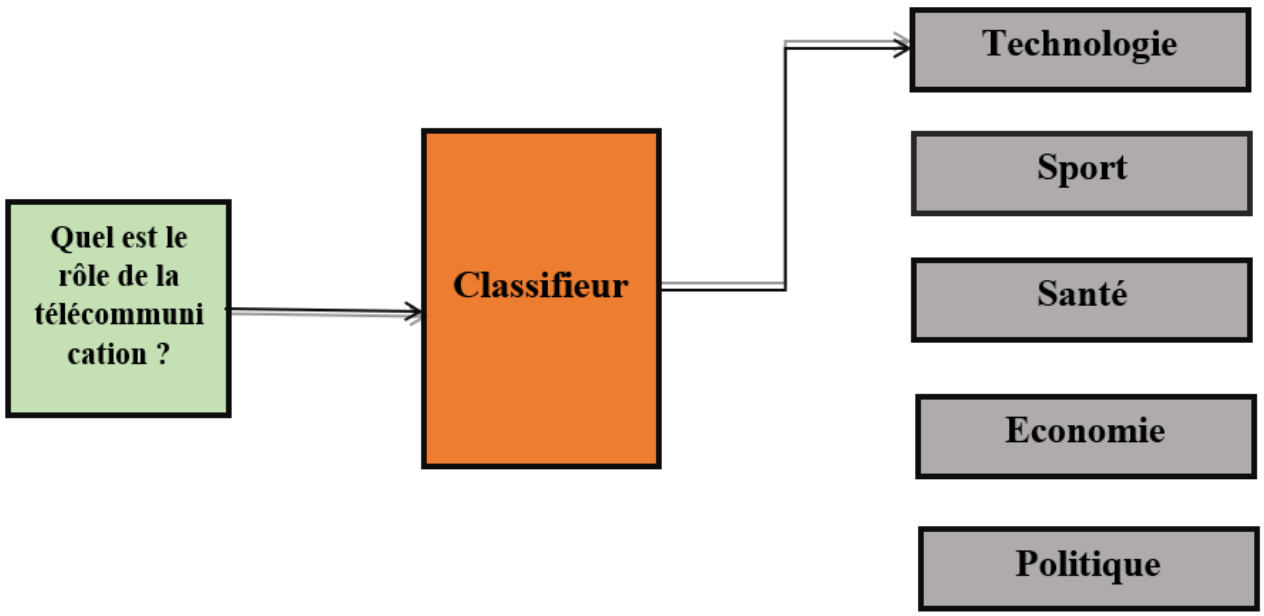


Figure 1.4 : Schéma illustratif de fonctionnement d'un outil de classification

1.5.1. Classification bi-classe et multi-classe

1.5.1.1. Classification bi-classe

Le système de classification répond à la question : "Le texte relève-t-il de la catégorie C ", pour cette requête la classification bi-classe correspond au filtrage.

C'est la question pour qu'on puisse poser la question est-ce un document autorisé pour les enfants ou non.

Cependant quand il s'agit d'effectuer une classification multi-classe qui permet de transmettre le document vers toutes les catégories les plus appropriés, on parle alors de routage. Cette classification multi-classes, selon le cas, peut être disjointes ou non. [2]

1.5.1.2. Classification multi-classe disjointe

C'est une taxonomie à n classes mais le document doit être affecté à une et une seule catégorie, nous avons trouvé ce type de catégorisation dans le routage des emails par exemple. La question "A quelle classe (au singulier) ce document appartient-il ? " doit être répondu par le système de classification multi-classes disjoint répond. [2]

Le but de la classification précédente est d'avoir une réponse claire à chaque texte (oui ou non, le texte T appartient à la catégorie C) [2], il peut être défini par classification déterministe. Plusieurs fonctions de classification sont utilisées, notamment : les règles de décision, les arbres de décision et SVM.

1.5.1.3. Classification multi-classe

Contrairement à la classification multi-classe disjointe, la classification multi-classe répond à la question : "A quelles classes (au pluriel) ce document appartient-il ? ", d'où dans ce système on peut conjoindre le texte à une ou plusieurs classes ou même sans classe. C'est le cas le plus général de la classification.

1.5.2. Catégorisation déterministe et floue

1.5.2.1. Catégorisation déterministe

La classification déterministe peut qualifier le but de la classification précédente qui le but de la classification précédente qui stipule de parvenir à une réponce bien détaillée pour chaque textes (oui ou non, le texte T appartient à la catégorie C).

Parmi les technologies de classification les règles de décision, les arbres de décision et les SVM sont employées. [2]

1.5.2.2. Catégorisation floue ou basée sur le Ranking

Pour classer un texte, on peut faire une évaluation simple de la catégorie la plus proportionnelle mais qu'à certains cas. Cela peut être nommé une classification ou un classement flou.

Le classement flou permettra aux utilisateurs d'être plus commode si le texte est près du sujet, et que si le texte n'a absolument rien à voir avec celui-ci dans le cas où ce dernier est incorrectement attribué à la classe.

Dans le cas où le thème est attribué par erreur à la classe, cette forme de catégorisation permettra à l'utilisateur d'être plus tolérant si le texte est "proche du thème" plutôt que si le contenu n'a rien à voir avec celui-ci.

Au lieu de lier catégoriquement un texte à une classe, le classement consiste à classer les classes par ordre de pertinence pour un texte particulier.

La catégorisation basée sur le Ranking est une forme de classement la plus facile et la plus réalisable en utilisant des méthodes d'évaluation de distance entre le texte et la catégorie, ainsi que quelques méthodes qui conclure la probabilité qu'un texte dans une classe.

Ludovic Denoyer dans (Denoyer, 2004), donne quelques exemples d'applications dans lesquelles ce système de classification est sollicité : [2]

- Le Ranking de pages Web pour une thématique définie par un internaute.
- Le Filtrage avec un rajustement de seuil de tolérance, le seuil étant ajusté par rapport aux scores de ranking.
- Proposer à un utilisateur un classement d'experts compétents pour évaluer un projet.

Dans ce cas spécifique, une fonction de score est définie de la manière suivante : [2]

SCORE : $D \times C \rightarrow [0,1]$

Cette fonction nous renseigne sur le degré d'appartenance d'un texte à une classe donnée. Ainsi, plus SCORE (d, c) est proche de 1, plus le document d est proche à être attribué à la classe c et inversement, plus cette valeur est proche de 0, plus le document est loin d'être attribué à la classe. Le calcul de cette fonction de score nous permet alors d'organiser les classes dans l'ordre pour y classer le texte et donc de savoir par exemple quelle est la classe la plus probable à être sélectionnée par rapport aux autres. [2]

Pratiquement, tous les algorithmes de classification calculent un score entre un texte et une classe. C'est le cas de toutes les approches probabilistes, particulièrement le classifieur Naïve Bayes. Toutefois, ces systèmes peuvent être aussi utilisés pour la classification déterministe. Dans ce cas, il est fondamental d'adopter une stratégie

transformant la fonction de score en une fonction de décision. Pour cela, la stratégie habituelle consiste à utiliser un seuil LC tel que : [2]

- Si SCORE (d, c) > L_c alors $\mathbf{D}$ (d,c) = vrai
- Sinon $\mathbf{D}$ (d,c) = faux

1.6. Catégorisation textuelle

Les classifications textuelles en domaines et en genres, qui représentent un enjeu pour la Recherche d'Information (RI), nécessitent de même un ensemble de descripteurs linguistiques adéquats. Dans les faits, domaines et genres sont associés à des niveaux linguistiques différents. Quand il s'agit de classification thématique ou domaniale, les textes sont souvent réduits à l'état de « sacs de mots ». Dans ce qui suit nous allons tout d'abord parler des objectifs et d'intérêts de la catégorisation textuelle, puis on va citer quelques cheminements et quelques technologies exploité à la catégorisation textuelle.

1.6.1. Objectifs et intérêts

Les objectifs et les intérêts des méthodes de catégorisation sont très variés. Elles peuvent être utilisées pour améliorer l'efficacité des moteurs de recherche de documents ou pour classer des articles sur la base de leurs références communes à d'autres articles afin de démontrer les liens entre eux. Nous pouvons citer différentes applications typiques qui sont :

- L'identification de la langue. [7]
- La reconnaissance d'écrivains et la catégorisation de documents multimédia. [7]
- Le classement automatique de différents communiqués de presse, ou messages sur des forums en différentes matières (« Les actualités de la région », « La bourse », « La culture » etc..).

A titre d'exemple : Une boite propose un système de tri d'informations dans des flots de dépêches d'agence de presse AFP ou Reuters etc… ou pages web. Chaque matin les nouvelles importantes sont faxées à différentes entreprises). [2]

- L'étiquetage de documents. [7]

- Indexation automatique sur des catégories d'index de bibliothèques : aide à la classification thématique des différentes rédactions dans une bibliothèque. [2]

- La gestion de bases documentaires (mémoire d'entreprise). Ce système peut être utilisé pour présenter l'information à l'utilisateur selon des catégories thématiques, ce qui facilite la navigation. [2]

- Sauvegarde automatique de fichiers dans des répertoires. [2]

- Les filtres internet en général, et en particulier les filtres anti-spams [2], le filtrage (déterminer si un document est pertinent ou non (décision binaire)) [7].

- Le classement automatique des emails, et particulièrement la redirection automatique de courriers des clients et fournisseurs en fonction de leur contenu vers les personnes compétentes dans une entreprise (Service commercial, livraison, service après-vente, approvisionnements, etc..) ou vers des répertoires prédéfinis dans un outil de Classification automatique de textes ou encore le tri de courriers électroniques dans différentes boîtes aux lettres personnelles et possibilité d'envoi de réponses automatiques. [2]

- Le routage des documents textuels (consistant à affecter un document à une ou plusieurs catégories parmi n). [7]

1.6.2. Démarche de la catégorisation textuelle

Pour réaliser l'opération de catégorisation automatique de textes comme nous l'avons défini, la démarche commune est la suivante : la première phase consiste donc à formaliser les textes afin qu'ils soient compréhensibles par la machine et utilisables par les algorithmes d'apprentissage.

La catégorisation des documents est la deuxième phase, cette étape est bien entendu décisive car c'est elle qui va permettre ou non aux techniques d'apprentissage de produire une bonne généralisation à partir des couples (Document, Classe). Pour améliorer la performance des modèles, une évaluation de la qualité des classifieurs et la comparaison des résultats fournis par les différents modèles est effectuée en fin de cycle. [38]

La démarche d'une approche standard de classification automatique de textes peut être résumée de la manière suivante :

- Eliminer les caractères de séparation, les signes de ponctuations, les mots vides, etc…
- Les termes restants sont tous des attributs.
- Un document devient un vecteur <terme, fréquence>.
- Entrainer le modèle de classification à partir des couples (Document, Classe).
- Évaluer les résultats du classifieur.

La figure ci-dessous (Figure 1.5) illustre la démarche de catégorisation de textes avec ses trois étapes qui peuvent être schématisées comme suit : [2]

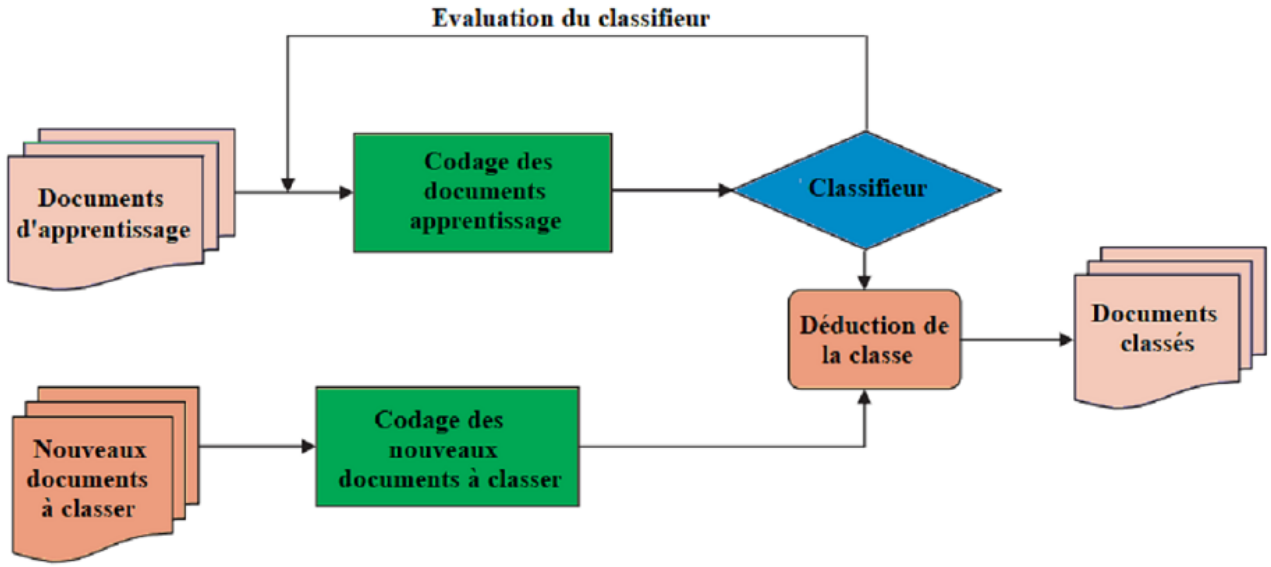


Figure 1.5 : démarche de la catégorisation de textes

Source : Mémoire de magister : MATALLAH Hocine

1.6.3. Quelques technologies utilisées dans la catégorisation textuelle

Historiquement, plusieurs générations consécutives d'algorithmes ont été utilisées dans la technologie de catégorisation automatique. Chaque génération apporte son lot d'améliorations par rapport à celle qui la précède. Nous pourrions mettre en avant les techniques sémantiques comme l'une des plus anciennes parmi les plus anciennes, dont le principal inconvénient était le coût humain et financier de la mise à niveau des systèmes de catégorisation.

Plusieurs entreprises ont mis sur le marché des solutions de catégorisation automatique basées sur des techniques purement statistiques et des techniques statistiques, basées sur des réseaux bayésiens pour résoudre ces contraintes. Depuis, de nombreuses générations d'algorithmes statistiques ont été créées, permettant des résultats bien plus significatifs. Nous présentons ici quelques approches du traitement statistique :

1.6.3.1 Les K plus proches voisins (KNN)

En intelligence artificielle, plus précisément en apprentissage automatique, la méthode des k plus proches voisins est une méthode d'apprentissage supervisé. En abrégé k-NN ou KNN, de l'anglais k-NearestNeighbors.

L'algorithme kNN suppose que des objets similaires existent à proximité. En d'autres termes, des éléments similaires sont proches les uns des autres dans cet algorithme on trouve la plupart du temps des points de données similaires. L'algorithme KNN repose sur le fait que cette hypothèse est suffisamment vraie pour que l'algorithme soit utile. L'algorithme kNN utilise l'idée de similitude (parfois appelée distance ou proximité) avec certaines notions de mathématique que nous aurions pu apprendre dans notre enfance, à savoir le calcul de la distance entre des points sur un graphique. [6]

- Charger les données.
- Initialiser K au nombre de plus proches voisins choisi.
- Pour chaque exemple dans les données :
  - Calculer la distance entre notre requête et l'observation itérative actuelle de la boucle depuis les données.
  - Ajouter la distance et l'indice de l'observation concernée à une collection ordonnée de données.
- Trier cette collection ordonnée contenant distances et indices de la plus petite distance à la plus grande (dans ordre croissant).
- Sélectionner les k premières entrées de la collection de données triées (équivalent aux k plus proches voisins).
- Obtenir les étiquettes des k entrées sélectionnées.
- Si régression, retourner la moyenne des k étiquettes.
- Si classification, retourner le mode (valeur la plus fréquente/commune) des k étiquettes.

Pour sélectionner la valeur de k qui convient à vos données, nous exécutons plusieurs fois l'algorithme KNN avec différentes valeurs de k. Puis nous choisissons le k qui réduit le nombre d'erreurs rencontrées tout en maintenant la capacité de l'algorithme à effectuer des prédictions avec précision lorsqu'il reçoit des données nouvelles (non vues auparavant). [6]

L'algorithme KNN a des avantages dont il est super simple et facile à mettre en œuvre, il n'est pas nécessaire de construire un modèle, d'ajuster plusieurs paramètres ou de faire des hypothèses supplémentaires aussi que cet algorithme est polyvalent. Il peut être utilisé pour la classification, la régression et la recherche d'informations. Comme il a des inconvénients qu'on va aussi citer est que l'algorithme ralentit considérablement à mesure que le nombre d'observations et/ou de variables dépendantes/indépendantes augmente. En effet, l'algorithme parcourt l'ensemble des observations pour calculer chaque distance.

Une particularité des algorithmes k-NN est d'être particulièrement sensible à la structure locale des données.

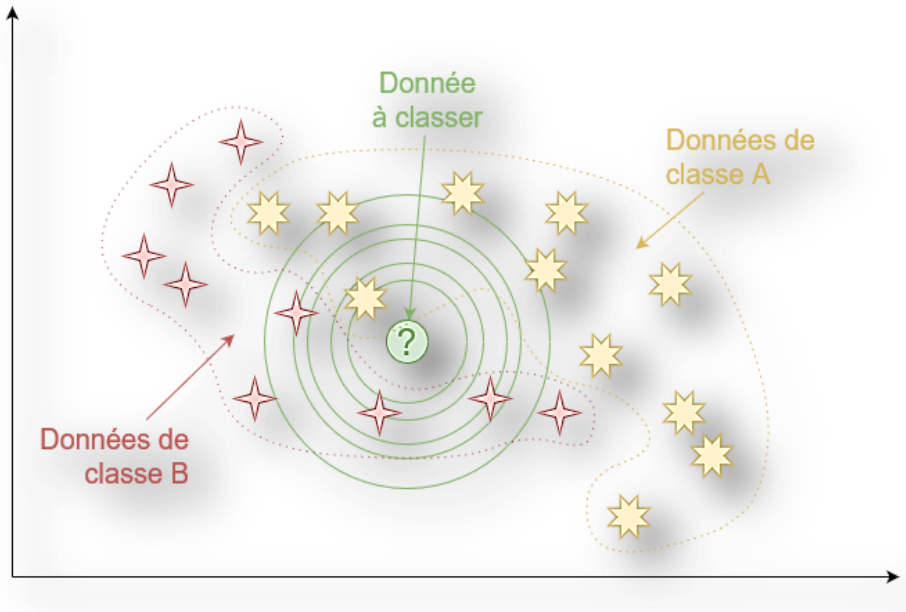


Figure 1.6 : Les K plus proches voisins

Source : Google image

1.6.3.2 Les arbres de décision

Les arbres de décision sont plus populaires des méthodes d'apprentissage. Les Algorithmes connus sont ID3 (Quinlan 1986) et C4.5 (Quinlan 1993). Ils sont également populaires pour la classification de document. Comme toute méthode d'apprentissage supervisée, les arbres de décision utilisent des exemples. Si l'on doit

classer des documents dans des catégories, il faut construire un arbre de décision par catégorie [5]. Pour déterminer à quelle(s) catégorie(s) appartient un nouveau document, on utilise l'arbre de décision de chaque catégorie auquel on soumet le document à classer. Chaque arbre répond Oui ou Non (il prend une décision). Concrètement, chaque nœud d'un arbre de décision contient un test (un IF...THEN) et les feuilles ont les valeurs «Oui» ou « Non». Chaque test regarde la valeur d'un attribut de chaque exemple. En effet, on suppose qu'un exemple est un ensemble d'attributs/valeurs. Pour des documents, chaque attribut peut être un mot et la valeur sera par exemple 0 ou 1 selon que ce mot appartient ou non au document. [9]

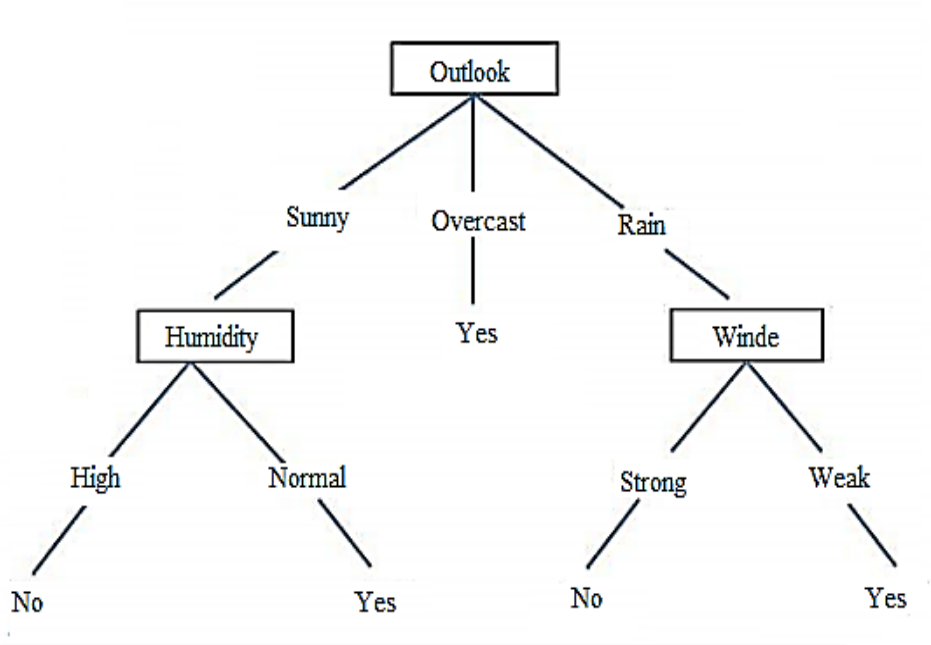


Figure 1.7 : L'arbre de décision

Source : Mémoire de Master : Ouali Choayb

Pour construire l'arbre de décision, il faut trouver quel attribut tester à chaque nœud. C'est un processus récursif. Pour déterminer quel attribut tester à chaque étape, on utilise un calcul statistique qui détermine dans quelle mesure cet attribut sépare bien les exemples Oui/Non. On crée alors un nœud contenant ce test, et on créer autant de descendants que de valeurs possibles pour ce test.

Cette méthode peut être utilisée dans plusieurs domaines tels que : Les études (pour comprendre les critères prépondérants dans l'achat d'un produit, l'impact des dépenses publicitaires), les ventes (pour analyser les performances par région, par enseigne, par vendeur), l'analyse de risques (pour détecter les facteurs prédictifs d'un comportement de non-paiement), Le domaine médical (pour étudier les rapports

existant entre certaines maladies et des particularités physiologiques ou sociologiques).[10]

Exemple : si on teste la présence d'un mot, les valeurs possibles sont Présent/Absent. A chaque fois, on aura donc deux descendants pour chaque nœud.

1.6.3.3 Machines à supports de vecteur (SVM)

Il s'agit d'une classe récente de méthodes d'apprentissage automatique. Les Machine à Vecteurs de Support (SVM) ont été introduites par [Vapnik, 1995, 1998]. Cette classe de méthodes est basée sur la minimisation de risque structurel. Les SVM cherchent une surface de décision « épaisse » pour séparer les points de l'ensemble d'apprentissage en deux classes. La décision prise est fondée sur les vecteurs de support (traduction de support vectors) sélectionnés pour définir la frontière entre les classes.

Les machines à support de vecteurs (SVM) sont à l'origine de nouvelles méthodes de catégorisations, bien que les premières publications sur le sujet datent des années 60. Avant d'aborder le principe de fonctionnement général des SVM voici quelques notions de base : [5]

- **Hyperplan :** est un séparateur d'objets des classes. De cette notion, nous pouvons dire qu'il est évident de trouver une mainte d'hyperplans mais la propriété délicate des SVM est d'avoir l'hyperplan dont la distance minimale aux exemples d'apprentissage est maximale, cet hyperplan est appelé L'hyperplan optimal, et la distance appelée marge.

- **Vecteurs Support :** ce sont les points qui déterminent l'hyperplan tels qu'ils soient les plus proches de ce dernier.

Voici un schéma représentatif de ces notions :

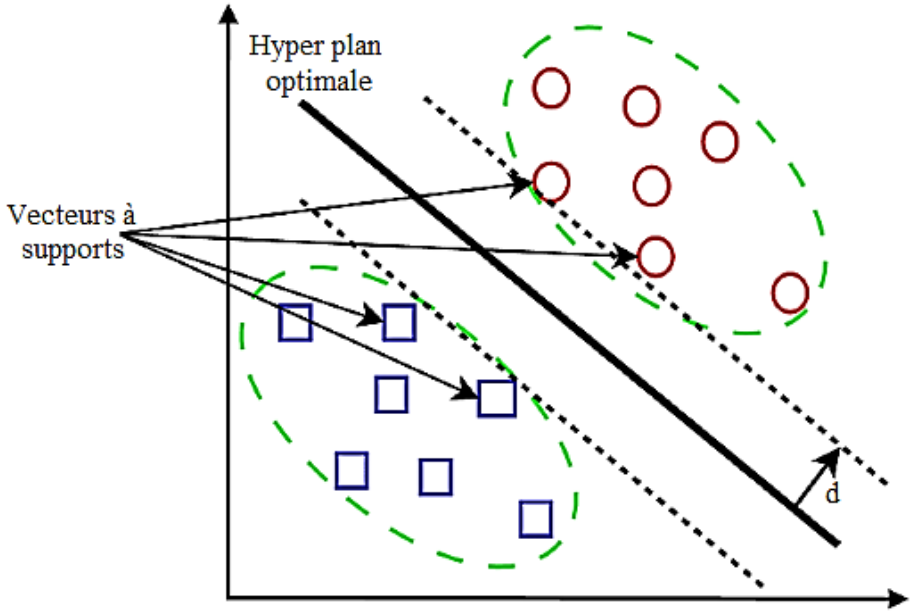


Figure 1.8 : Les vecteurs à support

Source : Google image

Le principe des SVM consiste en une stratégie de minimisation structurelle du risque, mais le problème revient à trouver une frontière de décision qui sépare l'espace en deux régions, à trouver l'hyperplan qui classifie correctement les données et qui se trouve le plus loin possible de tous les exemples. On dit qu'on veut maximiser la marge qui veut dire la distance du point le plus proche de l'hyperplan. Dans le cas de la catégorisation des textes, les entrées sont des documents et les sorties sont des catégories. En considérant un classificateur binaire, on voudra lui faire apprendre l'hyperplan qui sépare les documents appartenant à la catégorie et ceux qui n'en font pas partie [11].

Les SVM conviennent bien pour la classification de textes parce qu'une dimension élevée ne les affecte pas puisqu'ils se protègent contre le surapprentissage. Autrement dit, il affirme que peu d'attributs sont totalement inutiles à la tâche de classification et que les SVM permettent d'éviter une sélection agressive qui aurait comme résultat une perte d'information. On peut se permettre de conserver plus d'attributs. Également, une caractéristique des documents textuels est que lorsqu'ils sont représentés par des vecteurs, une majorité des entrées sont nulles. Or, les SVM conviennent bien à des vecteurs dits clairsemés. Un autre aspect positif des SVM est qu'aucun ajustement de paramètres manuel n'est requis, car ils ont l'habileté de trouver automatiquement des paramètres adéquats. [12]

1.6.3.4 Réseaux neurones

Le réseau de neurones artificiels (abrégé réseau de neurones) est l'un des algorithmes les plus utilisés de l'apprentissage automatique. Cet algorithme s'adapte à l'apprentissage non supervisé, supervisé pour la régression comme la classification (Specht, 1991). Il est généralement non probabiliste mais Specht a proposé une version qui l'est (Specht, 1990). Le réseau de neurones est l'algorithme d'apprentissage à la base du deep learning.

Le réseau de neurones artificiels est une tentative de formalisation des réseaux de neurones naturels et notamment du neurone formel. Sa création n'est pas due à l'informatique uniquement mais aussi aux sciences du vivant et de l'Homme. (Lettvin et al., 1959) est l'article fondateur du réseau de neurones. Le réseau de neurones peut être vu comme un graphe contenant des nœuds répartis par couche. Un réseau de neurones est constitué de trois couches au moins. La première couche est la couche d'entrée. Elle contient autant de nœuds que de variables définissant le vecteur de l'individu statistique. La seconde est la couche dite cachée. Il y a au moins une couche cachée mais il peut y en avoir de nombreuses (c'est notamment le cas dans les algorithmes de deep learning). Enfin, la dernière couche correspond à la couche de sortie. Dans le cas d'une régression, il n'y a qu'un nœud de sortie, dans le cas d'une classification, il y a autant de nœuds de sortie que de classes. [39]

Le principe général d'une approche neuronale est présenté ci-dessous. :

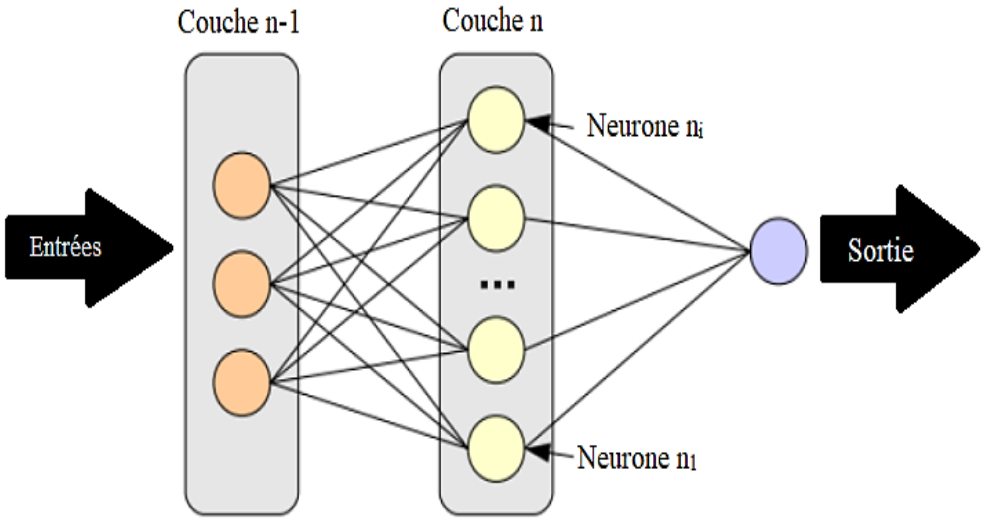


Figure 1.9 : Architecture générale d'un réseau de neurones artificiels

Source : Mémoire de Master en Informatique : Rimouche, Hachemi

Un réseau de neurones artificiels est composé d'une ou de plusieurs couches se succédant dont chaque entrée est la sortie de la couche qui la précède comme illustré sur la figure 1.8.

1.6.3.5 Les réseaux bayésiens

La méthode de classification Bayésienne naïve est une méthode par apprentissage sur un corpus. Le classement se fait en fonction de probabilités de suite de mots construits sur le corpus. La représentativité du corpus d'apprentissage par rapport aux données qu'il faudra exploiter est cruciale.

Cet algorithme utilise la règle de Bayes pour estimer la probabilité $P(c_j|d_i)$, basé sur la probabilité de la classe, et la probabilité que les mots qui composent d_j appartiennent à c_j. On suppose que les probabilités d'apparition des mots sont indépendantes bien que ce ne soit pas le cas. Cette stratégie n'est pas forcément optimale, quand il y a des erreurs dans l'estimation des probabilités, dues à un ensemble d'exemples trop restreint. Les expérimentations sur les réseaux Bayésiens ne se sont pas révélées concluantes.

Les réseaux bayésiens sont efficaces pour classer de manière binaire (spam ou non spam par exemple) des contenus sur des échantillons de plusieurs centaines d'éléments. Par contre, lorsqu'il s'agit de multiplier les classements possibles et donc, de réduire l'échantillon d'apprentissage, ils deviennent relativement peu pertinents : entre 20 et 40% d'efficacité seulement. [8]

1.7. Conclusion

Ces dernières années, la classification des textes a constitué un domaine de recherche de premier plan, tant pour les entreprises que pour les particuliers. Elle sert à la fois les entreprises et les particuliers.

Cette activité est due en partie à la forte demande des consommateurs pour cette technologie. Elle gagne en importance dans de nombreux cas où le traitement des documents textuels électroniques est difficile. La catégorisation de textes a fondamentalement progressé au cours des dix dernières années, grâce à l'avènement de techniques dérivées de l'apprentissage automatique, qui ont considérablement augmenté les taux de classification correcte.

Par cette rétrospective, nous avons donc tenté au titre ce chapitre de définir au mieux « La catégorisation de textes » et d'énumérer les concepts essentiels sur la base desquels, celle-ci trouve son fondement et sa logique probabiliste.

Chapitre 2 :

Détection de La Langue

CHAPITRE 2 : DETECTION DE LA LANGUE

2.1. Introduction

Une langue est un ensemble de signes (caractères) qui permettent aux gens de communiquer entre eux. L'identification des caractères (Alphabet) dont les mots et les phrases du texte à étudier constituent la base de la reconnaissance linguistique (également appelée reconnaissance de la langue ou du dialecte). L'idée de base est de trouver, d'identifier et de reconnaître les termes populaires d'une langue.

Dans ce chapitre on va décrire toutes les extensions qui concernent la détection de la langue, commençant par tous les concepts fondamentaux des langues (définition, classification des langues, dialectes, phénomènes linguistiques...), en parlant aussi de l'identification automatique de la langue ainsi que l'identification automatique des dialectes.

2.2. Concepts fondamentaux des langues

2.2.1 Définition de la langue

La langue est un ensemble de signes vocaux ou graphiques, choisis par chaque masse ou groupe d'hommes, par chaque nation, pour mettre de l'ordre dans la diversité des pensées, dans le but de les exprimer et de les communiquer aux autres hommes par le moyen de la voix.

L'étude des langues se fait au moyen des grammaires et des dictionnaires. Ceux-ci nous apprennent les mots isolés, celles-là nous apprennent à connaître quelle suite de modifications les mots sont susceptibles de subir pour exprimer telle ou telle circonstance épisodique ; comment les mots, lorsqu'on les réunit pour en former des phrases, des périodes, des discours, se combinent, se précèdent, se suivent, se transposent, et sous quelles modifications ils apparaissent lorsqu'il s'agit d'exprimer tel ou tel rapport entre eux.

La première partie se nomme « Lexicologie » et la deuxième, « Syntaxe ». Les deux réunies forment la grammaire, qui contient les règles du langage, la connaissance, la découverte et la fixation de ces règles.

Les langues que l'on parle aujourd'hui ont à peine quelques siècles d'existence. Comme tout se détruit et se renouvelle dans la nature, les langues ne sont pas exemptes de cette loi générale, elles naissent, vieillissent, périssent comme tout le reste, et sont remplacées par d'autres qui mettent plus ou moins de temps à se perfectionner, suivant que la civilisation est plus ou moins avancée. Il en est des

langues comme des arts, des sciences et de la littérature. C'est au sein de l'aisance et de la liberté qu'elles acquièrent cette richesse d'expression, cette pureté de style, cette énergie, qui les rend propres à transmettre nos connaissances à la postérité. [13]

2.2.2 Historique des langues

La question de l'origine des langues a toujours suscité de nombreuses hypothèses et mis à contribution les travaux tant des anthropologues, que des archéologues, des généticiens, des linguistes, etc. En 1865 la Société de linguistique de Paris avait informé ses membres dans ses règlements qu'elle ne recevrait « aucune communication concernant [...] l'origine du langage ». Mais la question a continué néanmoins à hanter les linguistes et la recherche d'une langue mère unique s'est poursuivie, si tant est qu'une telle langue ait existé. Dans L'Homme de paroles (Fayard, 1996), le linguiste français Claude Hagège réfute le mythe d'une langue commune unique :

> **Contrairement à l'idée courante, il est très probable que l'immense diversité des idiomes aujourd'hui attestés ne se ramène pas à une langue originelle unique pour toute l'humanité. S'il y a unicité, c'est celle de la faculté de langage propre aux hominiens et non celle de la langue elle-même. À l'origine, donc, une seule espèce (monogénisme de la lignée), mais non un seul idiome (polygénisme des langues).**

Néanmoins, l'idée d'une langue mère relève d'un fantasme ancien. Dès le Moyen Âge, on croyait à l'existence d'une langue originelle de l'humanité, jusqu'à ce que la colère de Dieu intervienne après l'épisode de la tour de Babel. Pendant longtemps, on a cru que l'hébreu était la langue d'Adam et d'Ève, d'autres ont pensé au latin ou au grec. Pour leur part, les musulmans ont toujours cru que la première langue de l'humanité était l'arabe.

A partir du XIX^{e} siècle, certains linguistes ont persisté dans ce type de recherche, ils ont été suivis par des spécialistes de la génétique des populations. L'un des livres les plus connus sur le celui fut celui de l'Américain Merritt Ruhlen (né en 1944) dans L'origine des langues (1994, mai 1997 en français). Ses travaux proposant une origine commune (la Monogenèse) ont alimenté une controverse vieille de plusieurs siècles. Pour établir des ressemblances entre toutes les langues du monde, la méthode de Ruhlen consiste à procéder à des comparaisons entre des lexiques de référence (en l'occurrence : 27 formes orthographiques associées aux formes phonétiques) pour un grand nombre de langues choisies parmi des familles communément

acceptées. Il s'agit du système de « Comparaison multilatérale » proposé auparavant par le linguiste américain Joseph Greenberg (1915-2001).

Quoi qu'il en soit, Merritt Ruhlen a avancé la thèse d'une proto-langue mère originelle et commune à toutes les superfamilles, qui aurait vécu vers 50 000 ans avant notre ère. Selon lui, le premier mot prononcé par l'homme serait la monosyllabe : Tik (« doigt ») ou AQWa (« eau »), appartenant à 32 familles de langues et proto-langues reconnues par la majorité des linguistes.

Cela étant dit, les critiques portant sur la méthodologie de Ruhlen sont innombrables. Non seulement on peut se demander si les ressemblances relevées par Ruhlen sont dues ou non au hasard, mais on met en doute la capacité des sons humains à se maintenir sur des dizaines de milliers d'années. Malgré tout, nombreux sont ceux qui reconnaissent au moins à Merritt Ruhlen le mérite d'avoir raison sur le fond : toutes les langues pourraient avoir une source unique, sauf que nous n'en savons strictement rien.

L'origine des langues reste toujours une énigme pour la science.

Cependant, si le moment de l'émergence du langage demeure encore une énigme pour la science et divise les linguistes, il est généralement admis que l'aptitude au langage se soit inscrite il y a environ 2,2 millions d'années dans le code génétique de l'Homo habilis, dont la capacité à fabriquer des outils témoigne déjà d'une grande complexité de l'organisation neurologique. On croit que cette aptitude n'aurait été utilisée que bien plus tard par l'Homo erectus, sinon par l'Homo sapiens, selon les plus prudents. Les langues, dans leur sens moderne, ne seraient apparues qu'entre 80 000 à 60 000 avant notre ère, en Afrique de l'Est ou au Proche-Orient, alors que nos ancêtres, les Homo sapiens, n'étaient plus que quelques milliers d'individus. À supposer qu'ils aient pu parler, on peut se demander s'ils parlaient une langue commune « Théorie de la monogénèse », auquel cas les quelque 6000 langues actuelles, descendraient de cette langue parlée il y a : 60 000 à 80 000 ans. On peut aussi imaginer que des langues existaient bien avant cette date et que les langues ne se soient développées qu'après la dispersion des différents groupes d'Homo sapiens « Théorie de la polygénèse ». Dans l'état actuel des choses, les outils de la science et de la linguistique comparée ne nous permettent pas d'en savoir davantage.

Pour sa part, le linguiste américain **Noam Chomsky** croit qu'il est possible qu'il y ait eu une langue d'origine unique, mais nous n'en savons strictement rien :

We don’t come from Adam and Eve. Get your facts right. The story of Adam and Eve is completely false. Get out to the world and teach yourself some reality. As for the origin of languages, it is possible that languages have single origin. But, wedon’t have clearevi-denceyet.	**[Nous ne venons pas d'Adam et Ève. Vérifiez vos sources. L'histoire d'Adam et Ève est complètement fausse. Sortez de votre monde et renseignez-vous sur une certaine réalité. Quant à l'origine des langues, il est possible que les langues aient une origine unique. Mais nous n'avons encore aucune preuve évidente.]**

La théorie néodarwinienne de l’évolution, plaide en faveur du polygénisme, c'est-à-dire que plusieurs couples humains seraient à l’origine de l’humanité. C’est au sein d’une espèce que prennent place les mutations génétiques, lesquelles séparent les espèces entre elles. Toutefois, cette théorie recèle encore beaucoup de zones d’ombre. On ne se surprendra pas que, dans ces conditions, la question sur l'origine des langues ne soit pas encore résolue. [14]

2.2.3 Classification des langues

Il est impossible d'estimer combien de personnes parlent chacune des quelque 7 000 langues parlées dans le monde. Chaque année, la base de données Ethnologue, la référence en matière de données statistiques sur les langues du monde, tente d'atteindre cet objectif. Les locuteurs natifs et les locuteurs de langue seconde sont inclus dans leur total. Les facteurs suivants peuvent être utilisés pour déterminer le classement :

2.2.3.1 Par nombre de langues

D’après les organisations du monde tous les spécialistes sont en désaccord sur le nombre de langue parlées dans le monde entier. Néanmoins, depuis des années passées, un chiffre semble mettre à peu près tout les monde en accord : il y en aurait environ 7000 langues vivantes. En fonction des études, cette donnée est très variable. Certaines indiquent moins de 5 000 langues quand d’autres disent 10 000 [40]. Selon l 'ONU reconnaît 141 langues officielles. [15]

2.2.3.2 Par famille de langues

Une famille de langues est un regroupement de plusieurs langues linguistiquement reliées, descendantes d’une langue-ancêtre commune (appelée protolangue). La plupart des langues du monde appartient à une famille définie et les

langues qu'on ne peut regrouper avec d'autres sont généralement appelées isolats. Seules les langues créoles ne sont ni des isolats, ni des membres d'une famille linguistique, mais constituent un modèle à part. En linguistique, ce type de relation est qualifié de « parenté généalogique ou génétique ». La linguistique comparée, comme son nom l'indique, est une branche de la linguistique qui compare les langues pour déterminer leurs liens de parenté. Cela est possible en comparant leur phonologie, leur grammaire et leur vocabulaire, même dans les cas où il n'existe pas de traces écrites de leurs ancêtres. [41]

Une famille de langues peut être divisée en plusieurs sous-groupes : par exemple, le polonais et le slovaque sont tous les deux des langues slaves occidentales, une subdivision des langues slaves, qui font elles-mêmes partie de la grande famille des langues indo-européennes.

Plus les langues sont éloignées, plus il peut être difficile de déterminer si elles ont un lien de parenté. Par exemple, aucun linguiste ne doute que l'espagnol et l'italien appartiennent à la même famille, mais l'existence des langues altaïques est controversée et n'est pas acceptée par tous les linguistes. Il est, à l'heure actuelle, impossible de savoir si toutes les langues descendent d'un ancêtre commun. Si une langue humaine originelle a existé, celle-ci n'est plus parlée depuis des dizaines de milliers d'années, sinon plus. [41]

	Grandes familles de langues	**Pays/Région**
1	Afro-Asiatiques	Corne de l'Afrique, Afrique du Nord, Sahara, Moyen-Orient, Malte
2	Nigéro-Congolaises	Afrique de l'Ouest, Afrique centrale, Afrique australe
3	Indo-Européenne	Europe, Asie mineure, monde iranien, Asie centrale, Inde du Nord
4	Ouraliennes	Hongrie et pays voisins, Finlande, Estonie, Lettonie, Russie
5	Dravidiennes	Sous-continent indien
6	Sino-Tibétaines	Asie
7	Austroasiatiques	Est de l'Inde, Asie du Sud-Est
8	Tai-Kadai	Thaïlande, Laos, Birmanie, Cambodge, Viêt Nam, Chine, Inde
9	Hmong-Mien	Chine, Laos, Viêt Nam, Birmanie
10	Austronésiennes	Asie du sud-est maritime, Océanie, Madagascar, Taïwan et Îles Andaman

Tableau 2.1: Principales familles de langues du monde

2.2.3.3 Par nombre de locuteurs dans chaque langue

Le nombre total de locuteurs représente le nombre de locuteurs de langue première additionné au nombre de locuteurs de langue seconde. Cependant, cet indicateur est délicat à appréhender statistiquement et il convient donc de prendre en considération plusieurs facteurs, dont les suivants : [17]

- Le nombre total de locuteurs est très souvent supérieur au nombre total de personnes recensées au sein d'une zone géographique donnée en raison doublons statistiques résultants du multilinguisme. [17]
- Dans le cas d'une situation bilinguisme parental, un enfant peut acquérir plusieurs langues natives simultanément lorsque les personnes chargées de son éducation. [17]
- Les citoyens d'un état ayant une seule langue officielle. [17]
- Les citoyens des états ayant plusieurs langues officielles ne sont pas systématiquement multilingues. [17]

De manière sûre, certaines langues sont systématiquement mises en tête de liste (les dix plus parlées) :

- Par le nombre de pays (Anciennes colonies, le plus souvent) : Anglais, Français, Arabe, Espagnol, Portugais.
- Par la population du pays : Mandarin, Hindi, Russe, Bengali, Indonésien.

Aussi, l'estimation du nombre total de locuteurs d'une langue demeure approximative et dépend également de nombreux autres paramètres statistiques. De plus, il n'est pas évident de distinguer linguistiquement une langue d'un « dialecte ». En effet, certaines langues dont le chinois, l'arabe et l'italien sont considérées comme des langues uniques où comme des familles de langues. A cela s'ajoute le facteur politique qui complique plus encore toute évaluation du nombre de locuteurs : des états parlant une même langue désirent marquer leur identité nationale en lui attribuant des noms différents. [17]

Enfin, un classement serré perd une partie de sa pertinence lorsque les marges d'erreur entre candidats se recoupent.

Rang	Langue	Famille	Langue mater-nelle	Rang L1	Langue seconde	Rang L2	Total
1	Anglais	Indo-Européenne	369,7 Millions	3	898,4 millions	1	1,268 milliard
2	Manda-rin	Sino-Tibétain	921,5 Millions	1	198,7 millions	4	1,120 milliard
3	Hindi	Indo-Européenne	342,0 Millions	4	295,3 millions	2	637,3 millions
4	Espa-gnol	Indo-Européenne	463,0 Millions	2	74,9 millions	9	537,9 millions
5	Fran-çais	Indo-Européenne	77,3 Millions	15	199,3 millions	3	276,6 millions
6	Arabe	Chamito-Sémitique	-	-	274,0 millions	-	274,0 millions
7	Bengali	Indo-Européenne	228,5 Millions	5	36,8 millions	13	265.2 millions
8	Russe	Indo-Européenne	153,6 Millions	7	104,3 millions	6	258,0 millions
9	Portu-gais	Indo-Européenne	227,9 Millions	6	42,2 millions	15	252,2 millions
10	Indoné-sien	Austroné-sienne	43,6 Millions	24	155,4 millions	5	190,0 millions

Tableau 2.2: Classement des principales langues selon leur nombre de locuteurs

Source : Ethnologue (23e édition, 2020)

2.2.3.4 Par dénomination

Avec le nombre énorme des langues et de leurs familles on distingue que chaque langue à des noms multiples et à ne pas confondre avec les noms de nations. D'une part, il y a une différence entre les noms de lieux et les exonymes. Locuteurs de la langue « inuit » et les « esquimaux » amérindiens.

D'autre part, il existe également des langues très proches et largement compréhensibles les unes des autres, mais qui sont différentes et portent des noms différents, comme le tchèque et le slovaque, le macédonien et le bulgare. En outre, il peut y avoir différents alphabets, histoires et différents noms pour la même langue parfaitement compréhensible, comme le croate et le serbe, ou l'hindi et l'ourdou. La même langue qui est parfaitement comprise par tous les locuteurs, utilise le même alphabet et a la même histoire, mais peut changer de nom selon le pays où elle est parlée, par exemple Moldavie-Romain.

C'est pourquoi les linguistes préfèrent utiliser des dénominations scientifiques marquées par le suffixe phone, comme lorsque l'on parle d'anglophones ou de francophones nonobstant leurs nationalité, origine ou histoire.

2.2.4 Dialectes

2.2.4.1 Définitions

Un dialecte (du bas latin : dialectus, du grec : διάλεκτος / diálektos, de διαλέγομαι / dialégomai « parler ensemble »), en gros une variante d'une langue propre à un groupe déterminé d'utilisateurs. Il existe des dialectes de n'importe quelle langue naturelle d'une population et d'une aire géographique spécifiques.

Le Larousse définit le dialecte comme « un ensemble de parlers qui présentent des particularités communes et dont les traits caractéristiques dominants sont sensibles aux usagers ». Pourtant, pour les linguistes, le dialecte est bel et bien une langue. [18] Plus simplement c'est une forme particulière d'une langue, intermédiaire entre cette langue et le patois, parlée et écrite dans une région d'étendue variable et parfois instable ou confuse, sans le statut culturel ni le plus souvent social de cette langue, à l'intérieur ou en marge de laquelle elle s'est développée sous l'influence de divers facteurs sociaux, politiques, religieux. [42]

2.2.4.2 Différence entre dialecte et langue formelle

Le langage est un système de symboles qui permet une compréhension mutuelle entre les groupes humains. Au niveau linguistique, il n'y a pas de frontières claires

entre les langues, les dialectes, les dialectes ou les dialectes : on parle plutôt d'un continuum linguistique. Cette expression signifie plus précisément qu'au sein d'un groupe linguistique, il existe une gamme de dialectes mutuellement intelligibles, chacun avec des différences qui n'empêchent pas la compréhension mutuelle.

Ainsi, il y a un continuum linguistique entre les différents dialectes de l'arabe, qui est un cas limite en raison de son extension spatiale : l'intercompréhension n'est pas évidente entre deux dialectes éloignés géographiquement (l'arabe du Maroc et celui du Yémen par exemple). Ce qui différencie la langue du dialecte ou du parler est le degré de reconnaissance officielle de leur statut, décrétée par l'État ou une autre forme de pouvoir dominant (une Église par exemple). Pour reprendre l'exemple de l'arabe, c'est le statut social, religieux et intellectuel de l'arabe littéral (véhiculaire) qui le différencie des formes vernaculaires de cette langue. D'une certaine manière, une langue est un dialecte qui a réussi à s'imposer aux autres.

Une situation, comme en France, ou la langue officielle, le français, a remplacé comme langue maternelle la plupart des dialectes (à l'exception de certaines régions et des langues créoles parlées outre-mer), est plutôt l'exception. Dans une grande partie des pays du monde, une ou plusieurs langues officielles se superposent à plusieurs langues vernaculaires ou régionales.

Le multilinguisme est alors une compétence partagée par une majorité d'habitants qui parviennent à passer d'une langue à une autre en fonction des contextes et des situations. Si on change d'échelle, cette situation se retrouve pour la France à l'échelle de l'Union européenne (UE), dans laquelle le français n'est qu'une langue parmi les 24 langues officielles de l'Union. En revanche, à la différence d'autres mosaïques linguistiques comme le Canada, la Russie, l'Inde ou la Chine, l'Union européenne ne possède pas de langue véhiculaire partagée par une grande majorité de locuteurs. [19]

2.2.5 Phénomènes linguistiques

Grâce à Internet, de nouvelles expressions et de nouveaux vocabulaires sont devenus courants. Tout le monde peut créer un nouveau terme ou une nouvelle expression et le diffuser sur l'internet. Il suffit que la création soit reprise par quelques individus pour qu'elle soit reconnue comme un moyen de communication et d'expression. « Urbandictionary » par exemple, est un site web dédié à ce type de comportement.

2.2.5.1 Code-Switching

Le code-switching (l'alternance codique) désigne le passage d'une langue à une autre dans une même conversation. Par exemple : « Cette fête a l'air fun, let's go ! ». Comme vous aurez remarqué, le recours aux mots anglais s'est répandu ces dernières années en France, notamment parmi les jeunes qui veulent donner un aspect « cool » et cosmopolite à leurs discussions quotidiennes. De nos jours, les nouveaux termes se répandent comme une traînée de poudre grâce aux réseaux sociaux. Il n'est donc pas étonnant que les anglicismes comme « friend-zone » soient utilisés partout dans le monde ; ils deviennent partie du langage de la jeunesse avant qu'on ne puisse créer une traduction. [20]

Néanmoins le code-switching signerait plutôt une grande richesse de communication et de compétence du bilingue sain, ce que nous allons démontrer par la suite. Cependant, il est intéressant de noter que ce phénomène est mal vu par les bilingues eux-mêmes ! C'est ainsi que des formes isolées de code-switching sont jugées inacceptables alors que c'est n'est plus le cas une fois remises en contexte. [42]

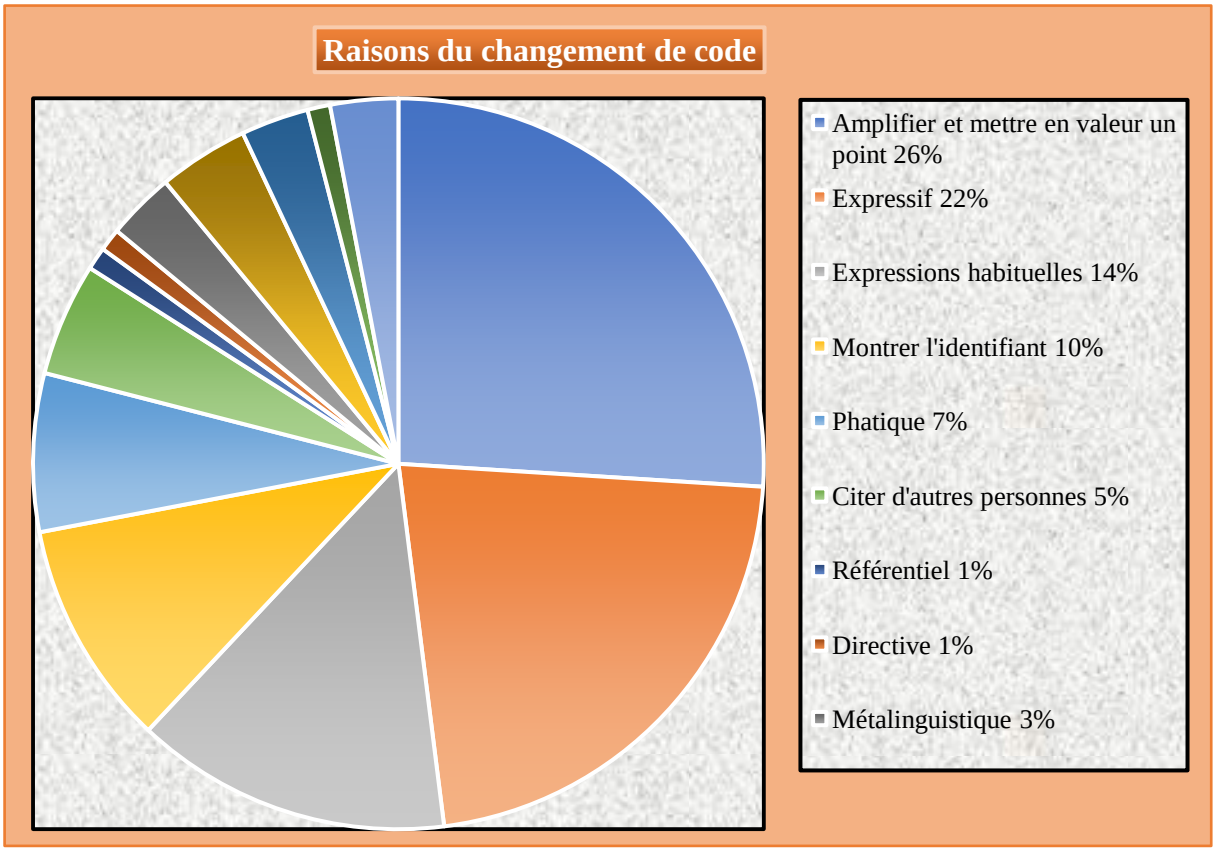


Figure 2.1: Raisons du changement de code (Code-Switching)

Source : Google image

2.2.5.2 Romanisation des langues non Latines

On entend par romanisation tout système permettant la conversion d'écritures non latines en alphabet latin, ou un système pour le faire. Les méthodes de romanisation comprennent la translittération, pour représenter le texte écrit, et la transcription, pour représenter la parole, et les combinaisons des deux. Les méthodes de transcription peuvent être subdivises en transcription phonémique, qui enregistre les phonèmes ou les unités de signification sémantique dans la parole, et transcription phonétique plus stricte, qui enregistre les sons de la parole avec précision.

Il existe de nombreux systèmes de romanisation cohérents ou normalisés. Ils peuvent être classés selon leurs caractéristiques. Les caractéristiques d'un système particulier peuvent le rendre mieux adapté à diverses applications, parfois contradictoires, y compris la récupération de documents, l'analyse linguistique, la lisibilité facile, la représentation fidèle de la prononciation. [21]

- Langue source ou langue du donateur : Un système peut être adapté pour romaniser le texte d'une langue particulière, ou d'une série de langues, ou pour n'importe quelle langue dans un système d'écriture particulier. Un système spécifique à la langue préserve généralement les caractéristiques linguistiques telles que la prononciation, tandis que le système général peut être préférable pour cataloguer les textes internationaux. [21]

- Langue cible ou récepteur : La plupart des systèmes sont destinés à un public qui parle ou lit une langue particulière. (Les systèmes dits internationaux de romanisation du texte cyrillique sont basés sur des alphabets d'Europe centrale comme l'alphabet tchèque et croate.) [21]

- Simplicité : Étant donné que l'alphabet latin de base à un plus petit nombre de lettres que de nombreux autres systèmes d'écriture, des digraphes, des diacritique-sou des caractères spéciaux doivent être utilisés pour les représenter tous en écriture latine. Cela affecte la facilité de création, de stockage et de transmission numériques, de reproduction et de lecture du texte romanisé. [21]

- Réversibilité : Indique si l'original peut ou non être restauré à partir du texte converti. Certains systèmes réversibles permettent une version simplifiée irréversible. [21]

Langue	Description de la langue	Lignes directrices sur translittération/ romanisation
Chinese (Mandarin)	Le mandarin (Chine) utilise des caractères logo-syllabiques qui représentent des mots sans la présence de voyelles ou de consonnes. Le mandarin est l'une des langues les plus parlées avec 845 millions de locuteurs.	Translittération et romanisation de chinois
Arabe	L'alphabet arabe a 28 lettres monocamérales qui n'ont pas de définition entre majuscules et minuscules. L'arabe est parlé par 175 millions de personnes et c'est l'une des langues officielles de l'ONU.	Translittération et romanisation de l'arabe
Arménien	L'arménien a sa propre écriture, et son alphabet se compose de 36 lettres. L'arménien compte 10 millions de locuteurs et c'est la langue maternelle des hauts plateaux arméniens.	Translittération et romanisation arméniennes
Hindi	L'hindi utilise l'écriture devanagari, et sa forme moderne d'alphabets a été développée au 15ème siècle. Avec près de 500 millions de locuteurs, l'hindi est la quatrième langue la plus parlée.	Translittération et romanisation hindi, marathi et népalais
Japonais	La langue japonaise moderne utilise trois scripts, kanji (caractères chinois), hiragana et katakana. Parlée par 127 millions de personnes, sa forme écrite a été développée au VIIIe siècle.	Translittération et romanisation japonaises

Tableau 2.3: Exemple de romanisation et translittération de quelques langues

Source : Thomas T. Pedersen, Translittération des écritures non romaines et Bibliothèque du Congrès (Tables de romanisation)

2.2.5.3 Argot et abréviations

L'argot est un langage de convention imaginé par les voleurs, les vagabonds et les diverses classes de gens hors de la société ou de la loi, pour communiquer entre eux sans être compris par ceux qui n'y sont pas initiés. Ce qui caractérise l'argot, c'est précisément la nécessité d'une initiation au sens des mots dont il se compose, qu'ils

soient forgés à plaisir ou que, tirés de la langue vulgaire, ils aient reçu une acception nouvelle.

L'argot est une chose aussi ancienne que la société. Du jour où il y a eu des hommes en lutte permanente avec la loi, ils ont dû recourir à un langage conventionnel destiné à soustraire la complicité de leurs tentatives ou de leurs actes au reste des hommes. Il y a des mots chez tous les peuples pour désigner cette langue de convention. [22]

L'abréviation est la réduction d'un mot par retranchement de lettres. Elle est généralement utilisée pour gagner de l'espace et du temps. Il n'existe pas de règles précises pour la formation des abréviations, mais il est possible de remarquer certaines régularités. [23]

Donc il y a lieu de distinguer trois grandes catégories d'abréviations :les abréviations proprement dites, les sigles et les acronymes et les symboles.

Argot	**Signification**	**Argot**	**Signification**
Bjr	Bonjour	**2m1**	Demain
Bcp	Beaucoup	**Dac**	D'accord
Jpp	J'en peux plus	**Jsp**	J'en ne sais pas
Msg	Message	**Svp**	S'il vous plait
Dsl	Désolé	**Mdr**	Mort de rire
Avc	Avec	**B8**	Bonne nuit
Cad	C'est-à-dire	**Re**	Je suis de retour

Tableau 2.4: Exemple d'argot et abréviations

2.3. Identification automatique de la langue

2.3.1 Identification de la langue des textes formels

2.3.1.1 Approches statistiques

La catégorisation de texte est une tâche fondamentale dans le traitement des documents, permettant l'extraction automatisée d'énormes flux de documents sous forme électronique.

L'une des difficultés rencontrées dans le traitement de certaines catégories de documents est la présence de différents types d'éléments textuels, tels que les fautes d'orthographe et de grammaire dans les e-mails et les erreurs de reconnaissance des caractères dans les documents qui passent par l'OCR. La catégorisation de texte doit fonctionner de manière fiable sur toutes les entrées et doit donc tolérer un certain niveau de ces types de problèmes.

Nous décrivons ici une approche basée sur N-gram (forme d'approche statistique) à la catégorisation de texte qui est tolérante des erreurs textuelles. Le système est petit, rapide et robuste.

Ce système a très bien fonctionné pour la classification des langues, obtenant en un seul test un taux de classification correct de 99,8% sur les articles d'où les textes sont formels.

Un type fondamental de processus de document est la catégorisation de texte, dans laquelle un document entrant est affecté à un cate-gory préexistant.

Le routage des articles de presse à partir d'un fil de presse est une application pour un tel système. Trier les archives papier numérisées en serait une autre. Ces applications présentent les caractéristiques suivantes :

- La catégorisation doit fonctionner de manière fiable malgré les erreurs textuelles.

- La catégorisation doit être efficace, avec le moins de temps possible de stockage et de traitement, en raison du volume même des documents à traiter.

- La catégorisation doit pouvoir être reconnue lorsqu'un document donné ne correspond à aucune catégorie ou lorsqu'il se situe entre deux catégories. En effet, les limites des catégories ne sont presque jamais clairement tranchées.

i. Section 1 : présente des mesures de similarité fondées sur les N-grammes :

Un N-gram est une tranche de N caractères d'une chaîne plus longue. Bien que dans la littérature, le terme puisse inclure la notion de tout ensemble de caractères co-occurrence dans une chaîne (par exemple, un N-gramme composé du premier et du troisième caractère d'un mot).

En règle générale, on tranche la chaîne en un ensemble de N-grammes de dépassement. Dans ce système, nous utilisons simultanément des N-grammes de plusieurs longueurs différentes. Nous ajoutons également des blancs au début et à la fin de la chaîne afin d'aider à faire correspondre le début du mot et la fin du mot situations. Ainsi, le mot « TEXT » serait composé des N-grammes suivants :

Bi-grammes : _T, TE, EX, XT, T_

Tri-grammes : TE, TEX, EXT, XT_, T_ _

Quad-grammes : _TEX, TEXTE, EXT_, XT_ _, T_ _ _

En général, une chaîne de longueur k, rembourrée avec des blancs, aura k+1 bi-grammes, k+1tri-grammes, k+1 quad-grammes, et ainsi de suite.[24]

ii. Section 2 : traite la catégorisation des textes à l'aide des statistiques de fréquence n-gramme :

Les langues humaines ont invariablement des mots qui se produisent plus fréquemment que d'autres. L'une des façons les plus courantes d'exprimer cette idée est devenue connue sous le nom de loi de Zipf[25], que nous pouvons reformuler comme suit :

Le nième mot le plus courant dans un texte de langage humain se produit avec une fréquence inversement proportionnel à n.

L'implication de cette loi est qu'il y a toujours un ensemble de mots qui domine la plupart des autres mots de la langue en termes de fréquence d'utilisation. En outre, il y a un seuil continu de dominance du plus fréquent au moins. La nature lisse des courbes de fréquence nous aide d'une certaine manière, car elle implique que nous n'avons pas à nous soucier trop de seuils de fréquence spécifiques. Cette même loi tient, au moins approximativement, pour d'autres aspects de langage humain.

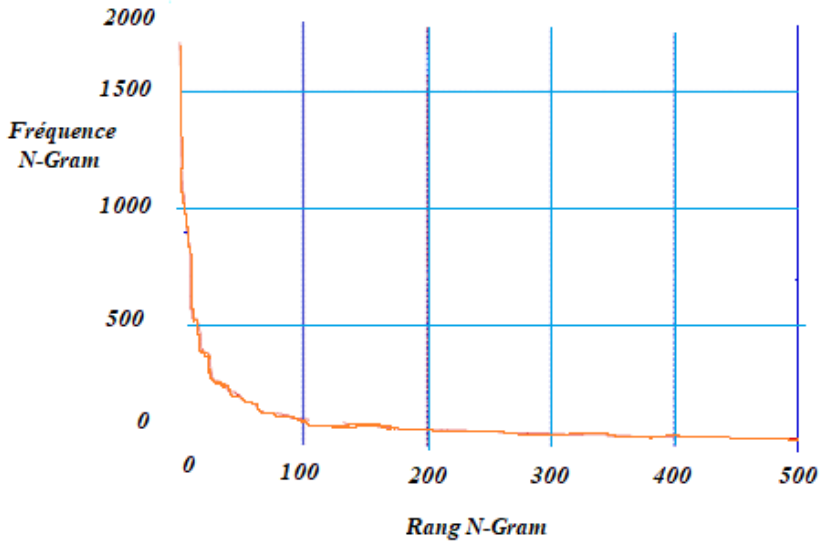


Figure 2.2: Fréquences n-gram par rang dans un document technique

iii. Section 3 : présente certaines conclusions et indique des orientations pour la poursuite des travaux :

La méthode de la fréquence N-gramme fournit un moyen inexpensif et très efficace de classer les documents. Essentiellement, cette approche définit une méthode de « Catégorisation par exemple ».

La collecte d'échantillons et la construction de profils peuvent même être traitées de manière largement automatique. En outre, ce système est résistant à divers problèmes d'OCR, il dépend des propriétés statistiques de N-gramme et non sur une occurrence particulière d'un mot.

Bien que le système existant ait déjà fait ses bons résultats, il y a encore de la place pour des travaux supplémentaires :

- Actuellement, le système utilise un certain nombre de N-grammes différents, dont certains dépendent en fin de compte davantage de la langue du document que les mots comprenant son contenu. En omettant les statistiques pour les N-grammes qui sont extrêmement courantes parce qu'elles sont essentiellement des caractéristiques de la langue, il peut être possible d'obtenir une meilleure discrimination à partir des statistiques qui restent. Il est également possible que le système comprenne des statistiques supplémentaires pour les N-grammes plus rares, ce qui lui permette d'obtenir une couverture supplémentaire.

- Il semble clair que la qualité de l'ensemble de documents affecte la performance de catégorisation des sujets. Nous aimerions expérimenter avec des ensembles de documents qui ont une cohérence et une qualité supérieures dans l'ensemble.

Par exemple, il serait intéressant de tester cette technologie sur un ensemble de résumés techniques pour plusieurs domaines différents. En divisant l'ensemble pour chaque zone en portions d'entraînement et de test, puis en calculant le profil de chaque zone à partir de l'ensemble d'entraînement, nous pourrions répéter cet expérimente d'une manière plus contrôlée.

- Les scores de correspondance bruts produits par le système sont en grande partie inutiles en eux-mêmes, sauf pour imposer un ordre relatif global des correspondances pour les différents profils. Pour corriger cela, nous devons concevoir un bon schéma de normalisation, qui produirait une sorte de mesure absolue de la qualité d'une correspondance particulière. Cela permettrait au système de rejeter certains documents au motif que leurs scores normalisés étaient si faibles que les documents ne correspondaient pas du tout à de bonnes correspondances.

Les scores normalisés permettent également au système de déterminer si un document particulier se mentait entre deux classifications en raison de sa nature interdisciplinaire. Une idée connexe serait de voir dans quelle mesure le système pourrait prédire quels articles sont recoupés dans différents groupes précisément en raison de leur contenu interdisciplinaire.

- Ce type de mesure de similarité de document convient parfaitement au filtrage et au routage de documents. Tout ce qu'un utilisateur doit faire est de créer un ensemble représentatif de documents qui couvrent les sujets pertinents, puis de calculer un profil global. À partir de là, il est simple et peu coûteux de calculer le profil de chaque document entrant, de le faire correspondre au profil global de l'utilisateur et d'accepter ceux dont les scores de correspondance sont suffisamment bons.

- Actuellement, ce système ne gère que les langues directement représentables en ASCII. La nouvelle norme ISO-6048/UNI-CODE ouvre la possibilité d'appliquer l'idée de fréquence N-gram à toutes les langues du monde, y compris les langues idéographiques.[24]

2.3.1.2 Approches à base d'apprentissage

En ce qui concerne le processus d'apprentissage, il existe deux critères d'optimisation principaux : une probabilité maximale (ml) et une information mutuelle maximale (MMI). En ML, la probabilité d'une séquence d'observation donnée OI, appartenant à une catégorie donnée (langue), est maximisée, compte tenu des paramètres de modèle de la catégorie. Le critère ML peut être exprimé mathématiquement.

Il n'y a aucun moyen connu de déterminer analytiquement les paramètres de modèle qui optimisent P (OI | Y). Néanmoins, les paramètres de modèle tels que P (OI |) sont optimisés localement peuvent être choisis, à l'aide d'une procédure itérative, comme la méthode Baum-Welch ou une méthode à base de gradient. Contrairement à ML, où une HMM d'une seule catégorie à la fois est maximisée, en gardant le HMMS pour d'autres catégories intactes à cette époque, dans le MMI, le concept de formation discriminatoire est api. C'est-à-dire que le HMMS de toutes les catégories est formé simultanément, de sorte que les paramètres du modèle approprié soient mis à jour pour améliorer sa contribution aux observations, tandis que les paramètres des modèles alternatifs sont mis à jour pour réduire leurs contributions. Dans notre expérimentation, le critère ML est utilisé.

L'apprentissage général procède comme suit : Initialement, tous les symboles de chaque état standard sont définis pour être également susceptibles, car les différentes valeurs de T, sont présentées au DHMM correspondant et sont segmentés de manière uniforme, à temps utile, parmi ses (standard) États SI. Les vecteurs affectés à un état sont considérés comme générés par cet état. L'avoir fait pour chaque séquence d'observation OI, des estimations de probabilité de probabilité maximales sont trouvées pour les paramètres de modèle Y = (A ; B, PI) ; de sorte que p (oi | y) est maximisé. L'alignement de Viterbi est utilisé ensuite pour la re segmentation des vecteurs d'observation et la recommutassions des estimations de la matrice de probabilité de transition A et la probabilité de sortie.

Matrice de distribution B, jusqu'à ce que la convergence soit obtenue, par exemple lorsqu'elle a atteint un p (OI | Y), ou lorsqu'une limite d'itération supérieure est atteinte. Les probabilités de transition à chaque itération sont estimées en comptant le nombre de fois que chaque transition est effectuée dans les alignements et la normalisation Viterbi. Plus tard, il est possible d'utiliser une phase de réévaluation pour attribuer chaque vecteur d'observation à chaque état proportionnellement à la probabilité du modèle dans cet état lorsque le vecteur a été

observé. La probabilité d'occupation de l'état est calculée ébouriffée au moyen de l'algorithme avancé. L'ensemble du processus s'appelle une réestimation de Baum-Welch ou d'un algorithme d'une optimisation des attentes).

Les preuves expérimentales ont montré que l'application de la réestimation Baum-Welch laisse les résultats presque non modifiés et a ensuite été omis. Ceci est probablement dû à l'utilisation excessive de l'alignement de Viterbi.

À l'étape d'identification, le décodage Viterbi est utilisé. Le but de cet algorithme est de trouver le meilleur chemin des transitions de l'état q = (q1 ; q2 ; .... Qt), compte tenu des vecteurs d'observation o et des paramètres de modèle. La probabilité de journalisation d'un chemin est calculée en prenant la sommation des probabilités de sortie du journal et des probabilités de transition du journal. Pour les modèles LR sans saut, le décodage Viterbi ne détermine pas l'ordre de la succession d'États, qui est connu en raison de la nature séquentielle du modèle, mais le nombre de fois que chaque état est utilisé comme générateur. International, l'algorithme de passage de jeton est utilisé. Pour la mise en œuvre de TLI, une structure parallèle des DHMM entraînée est construite. Pour chaque DHMM, il existe un résultat de probabilité calculé pour le meilleur chemin trouvé par l'algorithme Viterbi.

Le résultat d'identification est trouvé en choisissant, parmi les DHMM, celui qui est plus susceptible d'avoir produit la séquence d'observation des tests OTST. Cette approche est également utilisée dans la reconnaissance de mots isolée à l'aide de HMMS. Le HMM qui optimise la probabilité d'OTST lorsque les paramètres du HMM (Y) sont donnés sont choisis et, étant donné que chaque HMM représente une langue, cela est déclaré comme la langue identifiée. Pour la mise en œuvre de l'apprentissage et des phases d'identification. [26]

2.3.2 Identification de la langue des textes des réseaux sociaux

2.3.2.1 Approches statistiques

LangID est le problème de mapper un document sur la ou les langues dans lesquelles il est écrit. La technique la plus connue classe les documents en fonction des statistiques d'ordre de classement sur des séquences de n-grammes de caractères entre un document et un profil de langue global (Cavnar et Trenkle, 1994). D'autres approches statistiques appliquées à LangID incluent des modèles de Markov sur des profils de fréquence n-grammes (Dunning, 1994), des produits scalaires de vecteurs de fréquence de mots (Darnashek, 1995) et des noyaux de chaînes dans des machines à vecteurs de support (Kruengkrai et al., 2005). Contrairement aux méthodes

purement statistiques, des modèles à motivation linguistique pour LangID ont également été proposés, tels que l'utilisation de listes de mots vides (Johnson, 1993), où un document est classé en fonction de son degré de chevauchement avec des listes pour différentes langues. D'autres approches incluent la corrélation des mots et des parties du discours (POS) (Grefenstette, 1995), la tokenisation inter-langue (Giguet, 1995) et les modèles de classe grammaticale (Dueire Lins et Gonc¸alves, 2004). LangID des chaînes courtes a récemment suscité l'intérêt de la communauté des chercheurs. Hammarstrom (2007) décrit une méthode qui augmente un dictionnaire avec une table d'affixes et le teste sur des données synthétiques dérivées d'un corpus biblique parallèle. Ceylan et Kim (2009) comparent plusieurs méthodes d'identification de la langue des requêtes des moteurs de recherche de 2 à 3 mots. Ils développent une méthode qui utilise un arbre de décision pour intégrer les sorties de plusieurs approches LangID différentes. Vatanen et al. (2010) se concentrent sur des messages de 5 à 21 caractères, en utilisant des modèles de langage n-grammes sur des données tirées de la DUDH dans un classificateur naïf de Bayes. Carter et al. (2013) se concentrent spécifiquement sur LangID dans les messages Twitter en augmentant les méthodes standard avec des a priori LangID basés sur les messages précédents d'un utilisateur et le contenu des liens intégrés dans les messages, et c'est également la méthode utilisée dans TwitIE (Bontcheva et al., 2013).

Tromp et Pechenizkiy (2011) présentent une méthode pour LangID de messages texte courts au moyen d'une structure graphique, étendant le modèle de texte standard « sac » pour inclure des informations sur l'ordre relatif des jetons. Bergsma et al. (2012) examinent LangIDpour créer des collections twitter spécifiques à une langue, constatant qu'une méthode compressive formée sur des données hors domaine de Wikipedia et des corpus de texte standard fonctionne mieux que les identificateurs de langue standard qu'ils ont testés.

Goldszmidt et al. (2013) proposent une méthode basée sur des statistiques de classement, utilisant un processus d'amorçage pour acquérir des données d'entraînement dans le domaine à partir de messages Twitter non étiquetés. Des travaux récents ont également mis l'accent sur LangID au niveau du mot plutôt qu'au niveau du document (Yamaguchi et Tanaka-Ishii, 2012 ; King et Abney, 2013), y compris des recherches sur l'identification de la langue de chaque mot dans les communications multilingues en ligne (Nguyen et Dogruoz , 2013 ; Ling et al., 2013). Dans cet article, nous nous concentrons sur les messages monolingues, car bien qu'il soit plus simple, LangID des messages Twitter monolingues est loin d'être résolu. Dans la section 1, nous avons discuté de certains travaux à ce jour sur LangID

sur les données Twitter. Certains auteurs ont publié des ensembles de données d'accompagnement : l'ensemble de données utilisé par Tromp et Pechenizkiy (2011) a été rendu disponible dans son intégralité, composé de 9066 messages dans 6 langues d'Europe occidentale.

D'autres auteurs ont publié des identifiants de message avec des étiquettes de langue associées, notamment Carter et al. (2013), avec 5000 identifiants dans 5 langues d'Europe occidentale, et Bergsma et al. (2012), fournissant 13190 identifiants dans 9 langues de 3 familles de langues (arabe, cyrillique et devanagari). À ce jour, seul l'ensemble de données de Tromp et Pechenizkiy (2011) a été utilisé par d'autres chercheurs (Goldszmidt et al., 2013). Avec l'aimable coopération des auteurs, nous avons obtenu les ensembles de données complets de Carter et al. (2013) et Bergsma et al. (2012), nous permettant de présenter l'évaluation empirique la plus complète de LangID des messages Twitter à ce jour. Cependant, l'ensemble total de langues couvertes est encore très petit. Dans la section 2.1, nous présentons notre propre ensemble de données annotées manuellement, en ajoutant le chinois (zh) et le japonais (ja) aux langues qui ont des données annotées manuellement. [27]

2.3.2.2 Approches à base d'apprentissage

Dans cette démonstration, nous utilisons un apprenant naïf multinomial de Bayes. Par souci de concision, nous ne donnons qu'une brève esquisse de la technique, elle est décrit de manière beaucoup plus détaillée par McCallum et Nigam (1998). L'essentiel de la méthode est de calculer la probabilité qu'une instance appartienne à une classe Ci à partir d'un ensemble fermé C donné, et donc d'affecter la classe la plus probable à un document D, constitué d'un vecteur de n caractéristiques x1...xn :

$$c = \text{argmax}_{Ci \in C} P(Ci|D)$$

Le théorème de Bayes nous permet d'exprimer ceci comme :

$$c = \text{argmax}_{Ci \in C} [P(D|Ci)\ P(Ci)] / [P(D)]$$

Où P(D) est une constante de normalisation indépendante de Ci. Ainsi, pour la classification, il suffit d'estimer P(D|Ci) et P(Ci). C est modélisé comme une distribution catégorielle sur les classes, et donc P(Ci) est obtenu via une estimation du maximum de vraisemblance. Afin d'estimer P(D|Ci), nous faisons l'hypothèse naïve que chaque terme est conditionnellement indépendant.

La raison pour laquelle nous sélectionnons Bayes naïf multinomial est qu'il est relativement léger et qu'il s'est avéré très précis lorsqu'il est combiné à la sélection de caractéristiques (McCallum et Nigam, 1998, Manning et al., 2008).

Pour établir la généralisabilité de nos résultats, nous avons également expérimenté un classificateur prototype le plus proche basé sur la divergence d'asymétrie (Lee, 2001), basé sur les résultats de Baldwin et Lui (2010a). Cependant, lorsqu'il est combiné avec la sélection de caractéristiques, nous avons constaté qu'il était systématiquement surclassé par le classificateur naïf de Bayes, et omet donc les résultats de cet article. Nous avons également expérimenté avec un apprenant SVM à noyau linéaire, mais encore une fois, nous omettons les résultats car nous avons trouvé qu'il était comparable en précision à Bayes naïf lorsqu'il était combiné à la sélection de fonctionnalités, mais beaucoup plus lent à recycler. [28]

2.4. Identification automatique des dialectes

2.4.1 Dialectes des langues Latines

Le traitement des données dialectales constitue un défi pour les applications NLP. Lorsqu'il s'agit d'une langue non standard, les systèmes sont formés pour reconnaître les variations orthographiques et syntaxiques pour un traitement ultérieur dans des applications telles que la traduction automatique.

Dans le cas de l'allemand, un certain nombre d'études ont été publiées sur le développement d'outils et de ressources NLP pour le traitement de la langue non standard (Dipper et al., 2013), le traitement de la variation orthographique sur les données dialectales et la réalisation de la normalisation orthographique (Samardzi˘ c et al., 2015), et l'amélioration des performances des tagueurs POS pour les données dialectales (Hollenstein et Aepli, 2014).

L'identification des dialectes suisses allemands, sujet de la tâche partagée du GDI, a fait l'objet de quelques études récentes. Les méthodes d'identification des dialectes allemands se sont avérées particulièrement importantes pour la validation des méthodes appliquées à la compilation de corpus de dialectes allemands (Scherrer et Rambow, 2010a ; Scherrer et Rambow, 2010b ; Hollenstein et Aepli, 2015). [29]

L’identification du dialecte allemand est encore moins étudiée que l’identification du dialecte arabe. Scherrer et Rambow (2010) décrivent un système d’identification des dialectes fondé sur un lexique suisse-allemand généré par

des automates qui associe les formes de mots à leurs externe géographiques. Au test time, ils divisent une phrase en mots et cherchent leurs extensions géographiques dans le lexique. Hollenstein et Aepli (2015) présentent un système d'identification des dialectes suisse-allemand basés sur des trigrammes de caractères. Ils entraînent un modèle trigramme langage pour chaque dialecte et notent chaque phrase de test par rapport à chaque modèle. Le dialecte prédit est choisi en fonction de la plus faible perplexité. Bien que Samardzic et coll. (2016) présentent un corpus qui peut être utilisé pour le GDI, ils ne traitent pas de cette tâche dans leur article. Néanmoins, leur corpus a été utilisé pour évaluer les participants à la tâche partagée GDI du Défi DSL 2017. [30]

Les travaux présentés ici, concernent également des études sur la discrimination entre des groupes de langues : variétés de langues et dialectes similaires, tels que les langues slaves du sud (Ljubesi˘ c et al..., 2007), les variétés portugaises (Zampieri et Gebre, 2012), les variétés anglaises (Lui et Cook, 2013), les dialectes roumains (Ciobanu et Dinu, 2016), les variétés chinoises (Xu et al., 2016), et les éditions passées de la tâche partagée DSL (Zampieri et al., 2014 ; Zampieri et al., 2015 ; Malmasi et al., 2016c). [30]

2.4.2 Dialectes Arabes

L'identification du dialecte arabe est une tâche relativement nouvelle NLP avec seulement une poignée d'œuvres pour l'habiller (Biadsy et al., 2009 ; Zaidan et Callison- Burch, 2011 ; Elfardy et Diab, 2013 ; Darwish et al., 2014 ; Zaidan et Callison-Burch, 2014 ; Mal- masi et al, 2015).

Bien qu'il n'ait pas reçu trop d'attention, la tâche est très importante pour les outils de NLP arabe, car la plupart de ces outils n'ont été que des conceptions pour l'arabe standard moderne. Bi-adsy et al. (2009) décrivent une approche phonotactique qui identifie automatiquement le dialecte arabe d'un locuteur donné un échantillon de parole. Alors que Bi-adsy et al. (2009) se concentrent sur l'identification du dialecte arabe parlé, d'autres ont essayé d'identifier le dialecte ara-bic de textes donnés (Zaidan et Callison- Burch, 2011 ; Elfardy et Diab, 2013 ; Darwish et coll., 2014 ; Malmasi et coll., 2015).

Zaidan et Callison-Burch (2011) introduisent l'ensemble de données de commentaires en ligne en arabe (AOC) de 108K phrases la-beled, 41% d'entre elles ayant un contenu dialectal.

Ils utilisent un modèle linguistique d'identification dialectale automatique sur leurs données collectées. El-fardy et Diab (2013) proposent une approche supervisée pour l'identification du dialecte au niveau de la peine entre l'égyptien et la MSA. Leur système surpasse l'approche présentée par Zaidan et Callison-Burch (2011) sur le même ensemble de données. Zaidan et Callison-Burch (2014) extendent leurs travaux précédents (Zaidan et Callison- Burch, 2011) et effectuent plusieurs expériences ADI en utilisant des mots et des caractères p-grammes.

Différents de la plupart des travaux précédents, Darwish et al. (2014) ont constaté que les modèles d'uni grammes de mots ne se généralisent pas bien aux sujets invisibles. La suggestion que les caractéristiques lexicales, morphologiques et phonologiques peuvent capturer des informations plus pertinentes pour les dialectes discriminants.

Comme le corpus de l'AOC n'est pas contrôlé pour le biais du sujet, Malmasi et coll. (2015) affirment également que les modèles formés sur ce corpus peuvent ne pas se généraliser à d'autres données car ils impliquent et ils capturent des indices topiques. Ils effectuent des expérimentes ADI sur le Corpus parallèle multi dialectes arabes (MPCA) (Bouamor et al., 2014) en utilisant des modèles de mots et de caractères p-grammes afin d'évaluer l'influence du biais de sujet. Fait intéressant, Malmasi et al. (2015) trouvent que le caractère p-grammes est « dans la plupart des scénarios la meilleure caractéristique unique pour cette tâche », même dans un cadre inter-corpus.

Leurs conclusions concordant avec celles d'Ionescu et de Popescu (2016b) dans la tâche partagée de l'ADI du Défi DSL 2016(Malmasi et coll., 2016), car ils se sont emparés de la deuxième place en utilisant uniquement les caractères p-grammes des transcriptions de la reconnaissance automatique de la parole (ASR).

Toutefois, l'ensemble de données sur les tâches partagées de l'ADI de 2017 (Ali et coll., 2016) contient : les fichiers audios originaux et certaines fonctionnalités audios de bas niveau, appelées i-vectors, ainsi que les transcriptions ASR du discours arabe collectées dans le domaine Broadcast News.

Nos expériences indiquent que les fonctionnalités audios produisent une performance beaucoup meilleure, probablement parce qu'il y a beaucoup d'erreurs ASR (peut-être plus dans les segments de discours dialectaux) qui rendent l'identification du dialecte arabe à partir des transcriptions ASR beaucoup plus difficile. [31]

2.4.3 Dialectes Algériens

Les dialectes algériens ont été étudiés récemment pour les documents parlés, où certains auteurs ont abordé l'identification des accents. Par exemple, Djellab et al. ont proposé un nouveau corpus vocal pour la reconnaissance des accents régionaux algériens (AMCASC) [3].

Une étude récente concernant les sous-dialectes algériens a été menée par Bougrine et al. [5], dans laquelle les auteurs ont créé un nouveau corpus de parole arabe parallèle. Plus précisément, le corpus a été créé via un enregistrement direct, et les auteurs ont préparé une série de questions à poser à 109 participants de 17 villes différentes.

D'autre part, Amazouz et al. ont étudié le phénomène de commutation de code arabe-français dans des documents parlés, et ils se sont penchés en particulier sur l'identification du français, de l'arabe et des parties commutées par code [6].

Quelques travaux ont été menés sur l'arabe dialectal algérien pour les documents écrits. Par exemple, Adouane et Dobnik ont abordé l'identification de la langue dans des documents multilingues algériens, où l'identification a été effectuée au niveau du mot [7]. MSA, dialecte, français, anglais, mots empruntés, etc. De même, Cotterell et al. ont créé un corpus d'échange de codes algérien arabizi-français, le premier corpus de ce genre [8].

Les auteurs ont collecté un grand nombre de pages Web liées à différents sujets, extraites du site Web d'un journal algérien, et les textes ont été annotés au niveau des mots après avoir été prétraités.

Guellil et al. se sont penchés sur la traduction automatique de l'arabizi algérien [9], où ils se sont concentrés sur la traduction de textes arabizi en MSA en utilisant une approche neuronale.

Les auteurs ont aussi créé un corpus parallèle manuellement en traduisant des phrases MSA en arabizi, mais le corpus ne contenait pas de véritables textes en arabizi.

Dans un travail connexe, Guellil a construit un lexique de mots arabes pour l'analyse des sentiments en incorporant différentes possibilités d'écriture de certains mots [10]. L'identification de la langue des textes arabes a été abordée à nouveau par Adouane et al. Les auteurs ont testé et comparé trois pipelines, à savoir TextCat,

SVM et PPM (Prediction by Partial Matching), et les résultats ont montré que le SVM est plus précis que les deux autres.

L'arabe dialectal algérien a été étudié récemment afin de créer un traducteur parole-parole entre l'arabe dialectal algérien et le MSA [13-14]. Les auteurs ont construit un corpus en transcrivant des discours dialectaux et en traduisant les textes en MSA. Ainsi, deux dialectes algériens ont été étudiés, à savoir les dialectes d'Alger et d'Annaba, et pour lesquels les auteurs ont mis en évidence l'inflexion des verbes et le vocabulaire utilisé dans ces dialectes.

Enfin, Rahab et al. se sont penchés sur l'identification de la polarité des sentiments dans les commentaires de journaux algériens, en recueillant des commentaires écrits en arabe dans un quotidien bien connu [15-16]. Comme tâche de prétraitement, les auteurs ont converti manuellement tous les mots dialectaux en MSA avant d'appliquer le SVM et le NB. [32]

2.5. Conclusion

Dans ce chapitre, nous avons présenté une revue de littérature sur les travaux précédents sur la détection automatique des langues sous ses divers aspects, où nous avons mis en évidence certains travaux sur les langues latines et la langue arabe. En outre, nous nous sommes concentrés sur certains travaux réalisés sur les langues dialectales, et l’arabe dialectal en particulier.

De plus, nous avons discuté et défini le langage dialectal sur les médias sociaux, ses catégories et ses impacts sociaux sur les individus. Dans le chapitre suivant, nous présenterons la méthodologie adoptée pour identifier automatiquement la langue sur l’arabe dialectal algérien.

Chapitre 3 :

Méthodologie

3.1. Introduction

Avec l'utilisation croissante des plateformes en ligne, la classification des textes est devenue de plus en plus importante. Nous pouvons utiliser la classification automatique ou la classification manuelle pour effectuer de telles tâches. Après l'étude générale de la catégorisation et la détection de la langue qui ont fait l'objet du premier et du deuxième chapitre de ce mémoire, on s'est consacré dans ce troisième chapitre à connaitre les classificateurs utilisés dans la classification automatique aussi qu'on va avoir une idée à propos de notre approche.

3.2. Algorithmes de pointe

Les algorithmes de classification de texte sont largement utilisés dans les systèmes logiciels qui analysent de grandes quantités de données textuelles. Le choix entre les modèles de classification est déterminé par le type de données et le type de problème à résoudre. En fait, compte tenu de la précision de la prédiction et de l'efficacité du traitement, il est généralement utile d'évaluer plusieurs modèles de classification sur un ensemble de données donné.

3.2.1. Support Vector Machines (SVM)

Les machines à vecteurs de support (SVM) sont un ensemble de méthodes d'apprentissage supervisé utilisées pour la classification, la régression et la détection des valeurs aberrantes.

❖ Les avantages des machines à vecteurs de support sont :

- Efficace dans les espaces de grande dimension.

- Toujours efficace dans les cas où le nombre de dimensions est supérieur au nombre d'échantillons.

- Utilise un sous-ensemble de points d'entraînement dans la fonction de décision (appelés vecteurs de support), de sorte qu'il est également efficace en mémoire.

- Polyvalent : différentes fonctions du noyau peuvent être spécifiées pour la fonction de décision. Des noyaux communs sont fournis, mais il est également possible de spécifier des noyaux personnalisés.

❖ Les inconvénients des machines à vecteurs de support comprennent :

- Si le nombre de fonctionnalités est beaucoup plus grand que le nombre d'échantillons, évitez de trop ajuster le choix des fonctions du noyau et le terme de régularisation est crucial.

- Les SVM ne fournissent pas directement d'estimations de probabilité, celles-ci sont calculées à l'aide d'une validation croisée coûteuse quintuple (voir Scores et probabilités, ci-dessous).

Les machines à vecteurs de support dans scikit-learn prennent en charge à la fois des vecteurs d'échantillon denses (et convertibles en) et clairsemés (n'importe quel) comme entrée.

Cependant, pour utiliser une SVM pour faire des prédictions pour des données clairsemées, elle doit avoir été ajustée sur ces données. Pour des performances optimales, utilisez l'ordre C (dense) ou (clairsemé). [33]

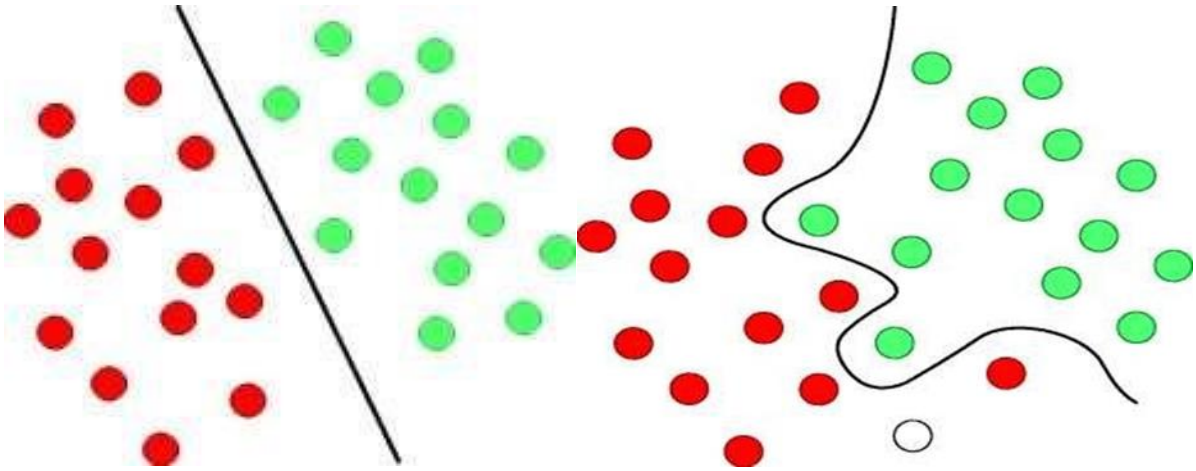


Figure 3.1 : Séparation linéaire et non linéaire dans l'espace de données d'entrée

Source : Google image-SVM-

3.2.2. Naïve bayes (NB)

Les méthodes de Bayes naïves sont un ensemble d'algorithmes d'apprentissage supervisé basés sur l'application du théorème de Bayes avec l'hypothèse « naïve » de l'indépendance conditionnelle entre chaque paire de caractéristiques compte tenu de la valeur de la variable de classe.

Le théorème de Bayes énonce la relation suivante, étant donné la variable de classe et le vecteur caractéristique dépendant à travers : yx_1x_n

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

En utilisant l'hypothèse naïve de l'indépendance conditionnelle selon laquelle :

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

pour tous, cette relation est simplifiée à

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Puisque est constante compte tenu de l'entrée, nous pouvons utiliser la règle de classification suivante : P (x_1, …, x_n)

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Et nous pouvons utiliser l'estimation maximale a posteriori (MAP) pour estimer la première est alors, la fréquence relative des cours dans l'ensemble d'entraînement P(y)P(x_i | y)y.

Les différents classificateurs naïfs de Bayes diffèrent principalement par les hypothèses qu'ils font concernant la distribution de P (x_i| y).

Malgré leurs hypothèses apparemment trop simplifiées, les classificateurs naïfs de Bayes ont très bien fonctionné dans de nombreuses situations réelles, célèbres pour la classification des documents et le filtrage du spam. Ils nécessitent une petite quantité de données d'entraînement pour estimer les paramètres nécessaires. (Pour des raisons théoriques pour lesquelles Bayes naïf fonctionne bien, et sur quels types de données il fonctionne, voir les références ci-dessous.)

Les apprenants et les classificateurs naïfs de Bayes peuvent être extrêmement rapides par rapport à des méthodes plus sophistiquées. Le découplage des distributions d'entités conditionnelles de classe signifie que chaque distribution peut

être estimée indépendamment comme une distribution unidimensionnelle. Cela aide à son tour à atténuer les problèmes découlant de la malédiction de la dimensionnalité.

D'un autre côté, bien que Bayes naïf soit connu comme un classificateur décent, il est connu pour être un mauvais estimateur, de sorte que les résultats de probabilité ne doivent pas être pris trop au sérieux. [34]

3.2.3. Neural Network (NN)

Un réseau de neurones est un ensemble d'algorithmes inspirés par le cerveau humain. Le but de cette technologie est de simuler l'activité du cerveau humain, et plus spécifiquement la reconnaissance de motifs et la transmission d'informations entre les différentes couches de connexions neuronales.

Un Deep Neural Network, ou réseau de neurones profond, se distingue par une particularité : il est composé d'au moins deux couches. Ceci lui permet de traiter les données de manière complexe, en employant des modèles mathématiques avancés.

En général, un Deep Neural Network a une couche d'entrée, une couche de sortie et au moins une couche entre les deux. Plus le nombre de couches est élevé, plus un réseau est dit " profond ". Chacune de ces couches effectue différents types de tri et de catégorisation spécifique dans un processus nommé " hiérarchie de caractéristique ".

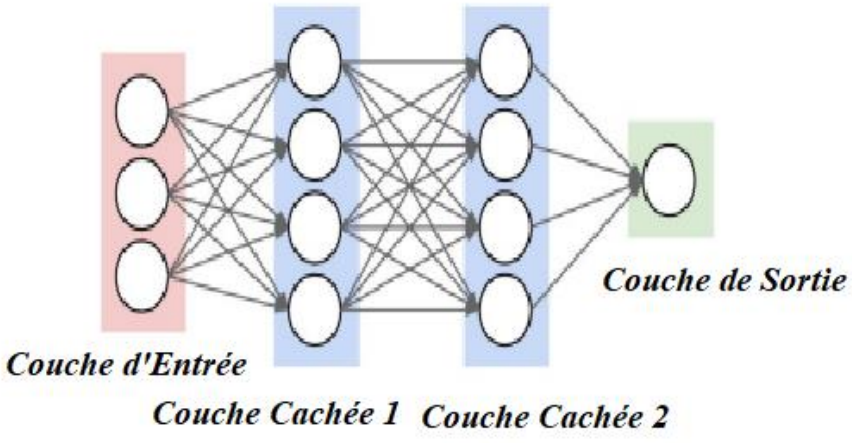


Figure 3.2 : Un réseau de neurones de base

Source : http://blog.christianperone.com

Pour mieux comprendre le fonctionnement d'un Deep Neural Network, il suffit d'observer le fonctionnement du cerveau humain. Plutôt que d'apprendre la structure du visage pour identifier les personnes, notre cerveau apprend à partir de la déviation d'un visage de base lui servant de modèle.

Lorsque nous voyons un visage, le cerveau cherche à déterminer comment celui-ci diffère de ce modèle de référence. Les caractéristiques telles que les yeux, les oreilles ou les sourcils sont ainsi passées en revue en une fraction de seconde.

Les différences entre le visage perçu et le modèle de visage " de base " sont quantifiés par un signal électrique dont la puissance varie. Toutes les déviations sont combinées pour produire un résultat.

Les différents nœuds du système sont similaires aux neurones du cerveau humain. Chaque couche est composée de plusieurs nœuds. Dès lors qu'ils sont touchés par stimuli, un processus se déclenche.

Le réseau de neurones interprète les données collectées par des capteurs ou directement injectées par un programmeur. Ces données peuvent être des images, des textes ou même des sons qui seront ensuite convertis sous forme de valeurs numériques.

Les différentes données entre la couche d'entrée et la couche de sortie doivent être traitées progressivement pour résoudre une tâche ou établir une prédiction. La première couche du réseau reçoit les données et exécute un calcul de fonction d'activation pour produire un résultat. Il peut s'agir par exemple d'une prédiction de probabilité.

Ce résultat est transmis à la couche suivante de neurones. La connexion entre deux couches successives est associée à un " poids ". Ce poids définit l'influence des données sur le résultat produit par la couche suivante et éventuellement pour le résultat final. [35]

3.2.4. Language Identification (LangId.py)

L'identification de la langue (LangID) est la tâche de déterminer la langue naturelle dans laquelle un document est écrit. C'est une étape clé dans le traitement automatique des données du monde réel, où une multitude de langues peuvent être présentes. Les techniques de traitement du langage naturel présupposent généralement que tous les documents en cours de traitement sont écrits dans une langue donnée (par exemple l'anglais), mais à mesure que l'accent est mis sur le

traitement de documents provenant de sources Internet telles que les services de microblogging, cela devient de plus en plus difficile à garantir.

L'identification de la langue est également un élément clé de nombreux services Web. Par exemple, la langue dans laquelle une page Web est écrite est une considération importante pour déterminer si elle est susceptible d'intéresser un utilisateur particulier d'un moteur de recherche, et l'identification automatique est une étape essentielle dans la création de corpus linguistiques à partir du Web. Cela a des implications pratiques pour les réseaux sociaux et les médias sociaux, où il peut être souhaitable d'organiser les commentaires et autres contenus générés par les utilisateurs par langue. Il a également des implications pour l'accessibilité, car il permet la détermination automatique de la langue cible à des fins de traduction automatique.

De nombreuses applications pourraient potentiellement bénéficier de l'identification automatique de la langue, mais la construction d'une solution personnalisée par application est d'un coût prohibitif, en particulier si une annotation humaine est requise pour produire un corpus de documents de formation étiquetés par langue à partir du domaine d'application. Ce qu'il faut, c'est donc un outil générique d'identification de la langue, utilisable sur étagère, c'est-à-dire sans formation de l'utilisateur final et avec une configuration minimale. Dans cet article, nous présentons langid.py, un outil LangID avec les caractéristiques suivantes : (1) rapide, (2) utilisable dans le commerce, (3) non affecté par les fonctionnalités spécifiques au domaine (par exemple HTML, XML, markdown), (4) fichier unique avec des dépendances minimales et (5) interface flexible. [36]

3.2.5. Language Detecting (LangDetect)

Dans le traitement du langage naturel, l’identification de la langue ou la devinette de la langue est le problème de déterminer dans quelle langue naturelle se trouve le contenu donné. Les approches computationnelles de ce problème le considèrent comme un cas particulier de catégorisation de texte résolu avec diverses méthodes statistiques.

Il existe plusieurs approches statistiques de l’identification de la langue utilisant différentes techniques pour classer les données. Une technique consiste à comparer la compressibilité du texte à la compressibilité des textes dans un ensemble de langues connues. Cette approche est connue sous le nom de mesure de distance basée sur l’information mutuelle. La même technique peut également être utilisée pour construire empiriquement des arbres généalogiques de langues qui correspondent étroitement aux arbres construits à l’aide de méthodes historiques. [Citation

nécessaire] La mesure de distance basée sur l'information mutuelle est essentiellement équivalente à des méthodes plus conventionnelles basées sur des modèles et n'est généralement pas considérée comme nouvelle ou meilleure que des techniques plus simples.

Une autre technique, telle que décrite par Cavnar et Trenkle (1994) et Dunning (1994) consiste à créer un modèle de langage n-gram à partir d'un « texte d'entraînement » pour chacune des langues.

Ces modèles peuvent être basés sur des caractères (Cavnar et Trenkle) ou des octets codés (Dunning) ; dans ce dernier cas, l'identification de la langue et la détection du codage des caractères sont intégrées. Ensuite, pour tout morceau de texte devant être identifié, un modèle similaire est créé, et ce modèle est comparé à chaque modèle de langage stocké. Le langage le plus probable est celui dont le modèle est le plus similaire au modèle du texte à identifier. Cette approche peut être problématique lorsque le texte d'entrée est dans une langue pour laquelle il n'existe pas de modèle. Dans ce cas, la méthode peut renvoyer un autre langage « le plus similaire » comme résultat. Les éléments de texte d'entrée composés de plusieurs langues, comme c'est le cas sur le Web, sont également problématiques pour toute approche.

Pour une méthode plus récente, voir Řehůřek et Kolkus (2009). Cette méthode peut détecter plusieurs langues dans un morceau de texte non structuré et fonctionne de manière robuste sur des textes courts de seulement quelques mots : quelque chose avec lequel les approches n-gram ont du mal. Une méthode statistique plus ancienne de Grefenstette était basée sur la prévalence de certains mots fonctionnels (par exemple, « the » en anglais). [37]

3.3. Conclusion

Dans ce chapitre, on a présenté le schéma général de quelques outils de classification de machine learning SVM, NN, NB, LangId, LangDetect. Ces algorithmes sont couramment utilisés dans différentes tâches de classification de textes et ont produit des résultats prometteurs. Le chapitre suivant rédige les résultats expérimentaux de notre tâche.

Chapitre 4 :

Résultats et Expérimentations

CHAPITRE 4 : RESULTATS ET EXPERIMENTATION

4.1. Introduction :

Dans ce chapitre, nous présentons les détails de notre travail et les problèmes obtenu ainsi que la collection du corpus ainsi que quelques statistiques. De plus, nous traçons et discutons des différents résultats obtenus par quelques classificateurs adoptés, où nous présentons les résultats de cette classification.

4.2. Description du corpus :

Les dialectes et sous-dialectes algériens ont suscité un intérêt croissant au cours de la dernière décennie. D'autre part, il existe également des travaux liés à la reconnaissance linguistique des textes multilingues écrits en arabe algérien.

La création du nouveau corpus appelé DZDC12, qui est collecté du réseau social Facebook., les textes collectés sont des commentaires d'utilisateurs écrits en arabizi-français. Dans l'ensemble, le corpus DZDC12 couvre 12 villes différentes, d'où on trouve quatre villes situées à l'est (Skikda, Annaba, Guelma et Constantine), quatre villes du centre (Blida, Alger, Tipaza et Medea) et quatre villes de l'ouest (Mostaghanem, Oran, Sidi-BelAbbes, Tlemcen et Sidi-BelAbbes). Ainsi, les textes du corpus DZDC12 ont été collectés par deux citoyens algériens de l'est du pays. Seuls les commentaires des utilisateurs rédigés en arabizi-français sont pris en compte, tandis que les commentaires en français et en anglais sont ignorés.

Enfin, afin de préserver l'anonymat des utilisateurs, les tags des utilisateurs sont remplacés par "######". Au total, il y a 200 textes pour chaque ville (100 textes écrits par des hommes et 100 textes écrits par des femmes), et Le corpus contient donc 2400 textes d'une longueur moyenne de 17 mots.

4.3. Installation et configuration :

Pour nous effectuons la tâche qu'on veut réaliser, on a tout d'abord commencé par l'installation du langage Python (3.9) et avoir une idée à propos de ce langage, puis on a lancé quelques configurations avant l'expérimentation avec les classificateurs de détection et d'identification de la langue aussi que les classificateurs machine learning.

Nous avons utilisé quelque bibliothèques open source pour implémenter les classificateurs qu'on a besoin comme le « Scikit-learn ».

Certaines bibliothèques Python supplémentaires ont été utilisées pour lire et préparer des données et pour la classification automatique.

4.4. Résultats expérimentaux

Tout d'abord le problème dans la présentation est structuré autour des orientations suivantes :

- Unicité du fichier annoté (fichier unique).
- L'ossature restante est identifiée comme un texte pur.
- Le format des fichiers (format texte) est inadapté au besoin des outils qui sont exprimés en valeurs numériques.

De ce fait, les résultats demandés ne peuvent être satisfaites avec ces conditions initiales. Alors, il devient judicieux de formater et convertir les données en fichier CSV avec lesquels le traitement devient possible.

Si les données de base seraient bien formulées, la phase programmation et l'application des outils seront effectuées sans entrave et les résultats seront interprétables.

Par conséquent l'analyse faite par deux (02) outils SVM et NB d'ensemble des outils demandés (05) conduit aux constations suivantes.

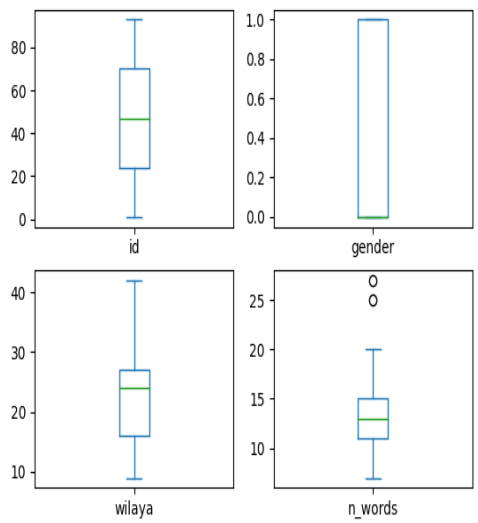
On a essayé de convertir 93 commentaires obtenus de douze wilayas en format CSV en gardant les informations essentielles du commentaire selon le fichier annoté ("id", "gender", "wilaya", "n_words", "w1", "w2", "w3", "w4", "w5" …), 'n_words' permis d'afficher le nombre de mots dans chaqu'uns des commentaires utilisés. Cette phase affiche tout le contenu du corpus utilisé aussi que le nombre lignes et colonnes

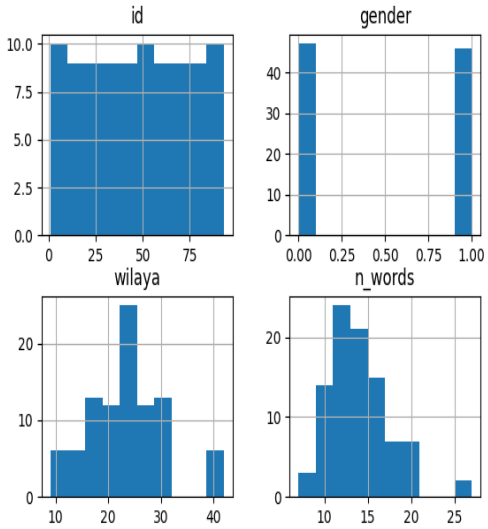
La figure suivante nous permettons de voir l'affichage obtenu :

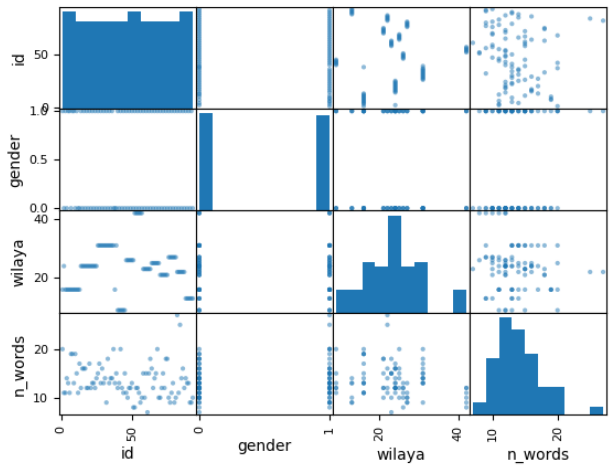
```
>>> %Run lng3.py
    id  gender  wilaya  n_words  ...      w12      w13    w14      w15
0    1       1      16       20  ...  nchalh       ya  rabi      NaN
1    2       0      24       11  ...      NaN      NaN   NaN      NaN
2    3       1      16       11  ...      NaN      NaN   NaN      NaN
3    4       0      16       14  ...       je   quitte    le   groupe
4    5       1      16       13  ...       ma   djiche.   NaN      NaN
..  ..     ...     ...      ...  ...      ...      ...   ...      ...
88  89       0      13       13  ...     yben     kter   NaN      NaN
89  90       1      13       13  ...        f   porsey   NaN      NaN
90  91       0      13        9  ...      NaN      NaN   NaN      NaN
91  92       1      13        9  ...      NaN      NaN   NaN      NaN
92  93       0      13       10  ...      NaN      NaN   NaN      NaN

[93 rows x 19 columns]
(93, 19)
               id     gender     wilaya    n_words
count   93.000000  93.000000  93.000000  93.000000
mean    47.000000   0.494624  23.344086  13.548387
std     26.990739   0.502681   7.813721   3.561619
min      1.000000   0.000000   9.000000   7.000000
25%     24.000000   0.000000  16.000000  11.000000
50%     47.000000   0.000000  24.000000  13.000000
75%     70.000000   1.000000  27.000000  15.000000
max     93.000000   1.000000  42.000000  27.000000
```

Figure 4.1 : Données affichées du corpus utilisé

Voici les graphes et les histogrammes figurantes de tache précédente :







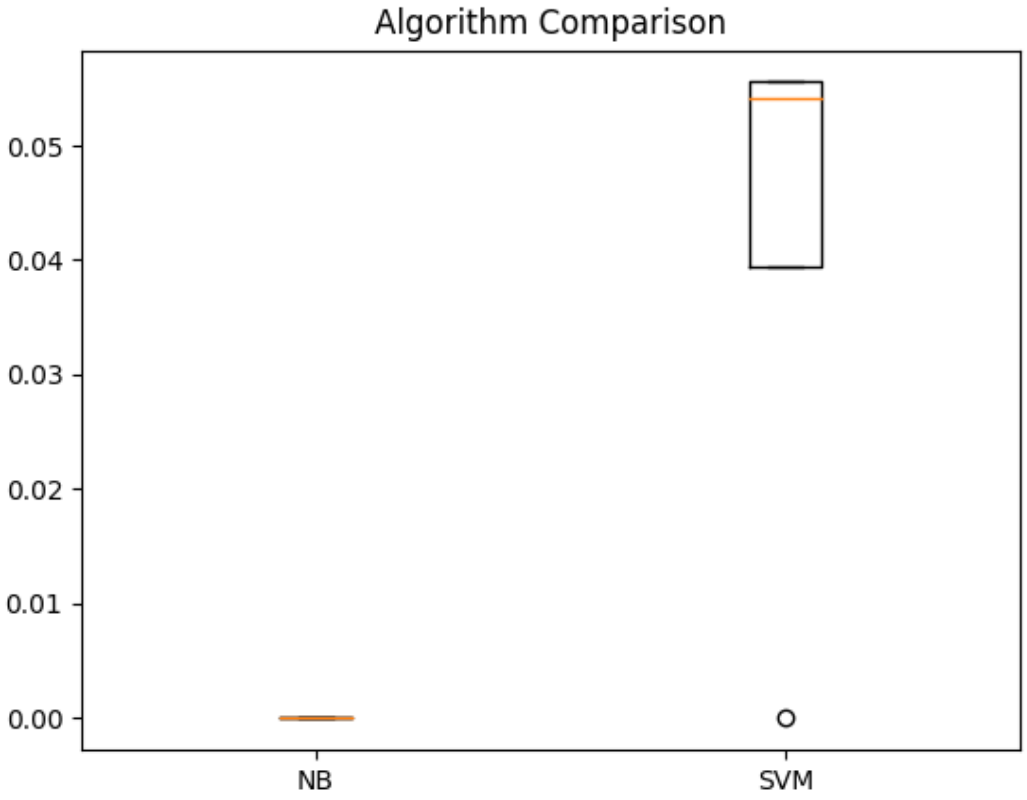
Graphique 4.2 : Graphes et Histogramme des données

De plus les outils lancer dans cette expérimentation le NB et le SVM ont interprétés de faible résultat car le corpus n'est pas assez grand pour effectuer de bons résultats. Lors d'utilisation de seulement 93 commentaires le pourcentage d'outil Naïve Bayes (NB) est nul par contre le Support Vectorial Machine (SVM) a donné un pourcentage de 41% selon la matrice numérique :

```
SVM: 0.040936 (0.023664)
0.0
[[0 0 0 0 0 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 1 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]
 [1 0 0 0 0 0 0 0 0 0 0 0]]
```

Ce qu'on a constaté que lorsque le nombre de commentaires est grand l'application et les résultats des outils seront plus élevées, nous allons illustré en-dessous un modèle confirmant à cette constatation.

❖ Comparaison entre les deux outils appliqués :



Graphique 4.3 : Comparaison entre le NB et le SVM

4.5. Conclusion

Les résultats expérimentaux réalisés sur DZDC12 ont montré de faibles performances, ce qui signifie que le problème abordé est assez difficile et nécessite une étude approfondie pour sélectionner le meilleur classificateur avec les meilleures caractéristiques. Comme perspectives, nous aimerions étudier des algorithmes efficaces pour augmenter les performances d'identification, ainsi que pour augmenter la taille du corpus.

Conclusion Générale

Conclusion générale

Dans ce manuscrit, nous avons présenté notre approche pour traiter de la détection des textes arabizi qui ont une forme d'écriture irrégulière, le phénomène de changement de code, et les sous-dialectes dérivés du même dialecte et partagent plusieurs caractéristiques linguistiques. Nous avons étudié les principes fondamentaux de la linguistique computationnelle et des méthodes de catégorisation des textes, et nous avons mis en évidence différents types de langues et pour distinguer le dialecte algérien de l'arabe moderne. Par la suite, nous avons défini le langage offensant avec ses catégories, ses risques et ses résultats. Nous avons brièvement discuté des différents algorithmes d'apprentissage automatique et d'apprentissage profond couramment utilisés dans la catégorisation de texte.

Une création d'un nouveau corpus appelé DZDC12 (fait référence au corpus DZiri Dialectes de 12 villes), pour lequel les textes sont collectés sur le réseau social Facebook. Plus précisément, les textes sont des commentaires d'utilisateurs rédigés en arabe-français par commutation de code liés à trois différents pays du Nord. En raison des fausses informations utilisées par de nombreux utilisateurs, nous ne recommandons pas de méthodes automatiques de collecte de données lors du traitement de tels textes, car cela peut entraîner des résultats erronés dans la tâche de catégorisation. Ainsi, les textes du corpus DZDC12 ont été explorés manuellement par deux citoyens algériens de l'Est, en tout le corpus est rassemblé de 200 textes pour chaque ville (textes écrits par des hommes et des femmes), puis le corpus contient 2400 textes d'une longueur moyenne de 18 mots.

Nous avons utilisé plusieurs algorithmes de classification pour entraîner nos modèles avec le corpus créé. Nous avons évalué des modèles de classification automatique sur notre corpus pour évaluer les performances des algorithmes de pointe sur ces données textuelles. Les résultats expérimentaux réalisés sur DZDC12 ont montré de faibles performances car les résultats demandés ne peuvent être satisfaites avec ces conditions initiales, alors il devient judicieux de formater et convertir les données avec lesquels le traitement devient possible. Si les données de base seraient bien formulées, la phase programmation et l'application des outils seront effectuées, ce qui signifie que le problème abordé est assez difficile et nécessite une étude approfondie pour sélectionner le meilleur classificateur avec les meilleures caractéristiques.

Dans les travaux futurs, notre corpus doit être élargi pour couvrir un contenu plus satisfaisant, et notre algorithme basé sur des règles doit être étudié pour résoudre les problèmes de complexité linguistique et obtenir de meilleurs résultats de classification.

Bibliographie

[1] Site internet : ” https://monkeylearn.com/what-is-text-classification/”

[2] H. MATALLAH ” Classification Automatique de Textes Approche Orientée Agent”, Mémoire de magister en informatique, Université Abou Bekr Belkaid-Tlemcen, 2011

[3] Site internet : ” https://La notion de classe pour les systèmes de classification – Apprendre en ligne (clicours.com) ”

[4] Site internet : ”https://cours en ligne-dspace.univtlemcen.dz sur la classification automatique de textes”

[5] C. OUALI ” Classification automatique de textes ”, Mémoire de Master en informatique, Université Mohamed Boudiaf- M’Sila, 2014

[6] Site internet : ”https://moncoachdata.com/blog/algorithme-des-k-plus-proches-voisins/”

[7] N. RIMOUCHE, H. HACHEMI, ” Amélioration du produit scalaire via les mesures de similarités sémantiques dans le cadre de la catégorisation des textes ”, Mémoire de Master en Informatique, Université Abou Bekr Belkaid- Tlemcen, 2015

[8] Site internet : ” https://Microsoft-Word-White_Paper_Categorisation_fr.doc (digimind.fr) ”

[9] R. Lefébure, G.Venturi, « Le Data Mining » Edition EYROLLES, deuxième tirage 1998.

[10] S.Raheel, « L’Apprentissage Artificiel pour la Fouille de Données Multilingues: Application à la Classification Automatique des Documents Arabes », Thèse de doctorat en Sciences de l’Information et de la Communication, Université Lumière Lyon 2, 2010.

[11] T.DERDRA Amel, F.BENSFIA, « La Représentation Conceptuelle pour la Catégorisation des Textes Multilingue », Mémoire de Master, Université Abou Bakr Belkaid– Tlemcen, 2011-2012.

[12] Simon RÉHEL, « Catégorisation automatique de textes et Cooccurrence de mots provenant de documents non étiquetés », Mémoire, Université Laval Québec, Canada, Janvier 2005.

[13] Site internet : ” https://www.espacefrancais.com/la-langue-le-langage/ [14] Site internet : L'origine des langues (ulaval.ca) ”

[15] Site internet : ” elearning.univ-djelfa.dz/course/view.php?id=79 ”

[16] Site internet : ” annexe_doc_31.pdf (irdp.ch) ”

[17] Site internet : ” https://atlasocio.com/ ”

[18] Site internet : ” https://blog.assimil.com/langue-dialecte-argot-patois-quelles-differences/ ”

[19] Site internet : ” https://geoconfluences.ens-lyon.fr ”

[20] Site internet : ” https://www.laroutedeslangues.com/blog/code-switching-une-menace-ou-un-developpement-enrichissant/ ”

[21] Site internet : ” https://stringfixer.com/fr/Romanisation”

[22] Site internet : ” https://www.espacefrancais.com/largot/ ”

[23]Site internet:”https://www.alloprof.qc.ca/fr/eleves/bv/francais/les-abreviations-f1013”

[24] Zipf, George K., Human ” Behavior and the Principle of Least Effort, an Introduction to Human Ecology ”, Addison-Wesley, Read- ing, Mass., 1949

[25] William B. Cavnar et John M. Trenkle -Institut de recherche environnementale du Michigan

[26] A. Xafopoulos, C. Kotropoulos∗, G. Almpanidis et I. Pitas ” Language identifcation in web documents using discrete HMMs ”

[27] Marco Lui and Timothy Baldwin NICTA VRL, ” Accurate Language Identification of Twitter Messages ”, Department of Computing and Information Systems University of Melbourne

[28] Marco Lui and Timothy Baldwin NICTA VRL, ” Cross-domain Feature Selection for Language Identification ”, Department of Computer Science and Software Engineering University of Melbourne

[29] Shervin Malmasi, ” German Dialect Identification in Interview Transcriptions ”, Harvard Medical School, USA, Macquarie University, Australia Marcos Zampieri University of Cologne

[30] Radu Tudor Ionescu and Andrei M. Butnaru, ” Learning to Identify Arabic and German Dialects using Multiple Kernels ”, University of Bucharest, Department of Computer Science

[31] Radu Tudor Ionescu and Andrei M. Butnaru, ” Learning to Identify Arabic and German Dialects using Multiple Kernels ”, University of Bucharest, Department of Computer Science

[32] Kheireddine Abainia-Université 8 mai 1945 – Guelma, ” Detecting Algerian Sub-Dialects of On-Line Commentators in Social Media ”, Networks Conference Paper October 2018

[33] Site internet : ” https://scikit-learn.org/stable/modules/svm.html#scores-probabilities ”

[34] Site internet : ” https://scikit-learn.org/stable/modules/naive_bayes.html ”

[35] Site internet : ” https://datascientest.com/neural network ”

[36] Marco Lui and Timothy Baldwin NICTA, ” langid.py: An Off-the-shelf Language Identification Tool ”, VRL Department of Computing and Information Systems University of Melbourne

[37] Site internet : ” https://wikipfr.icu/wiki/Language_identification ”

[38] Site internet : ” https://www.mcours.net/cours/pdf/leilclic3/leilclic929.pdf ”

[39] JC.RISCH ” Enrichissement des Modèles de Classification de Textes Représentés par des Concepts”, Mémoire de doctorat en informatique, Université DE REIMS CHAMPAGNE-ARDENNE-France, 2017

[40] David M. Eberhard, Gary F. Simons and Charles D. Fennig (eds.). 2021. Ethnologue: Languages of the World. Twenty-fourth edition. Dallas, Texas: SIL International.

[41] Site internet : ” https://lingvo.info/fr/babylon/language_families ”

[42] Site internet : ” https://www.lalanguefrancaise.com/dictionnaire/definition/dialecte#0”

[43] J. CLAIRE ” LE CODE-SWITCHING : UN MOYEN DE FACILITATION POUR LE BILINGUE APHASIQUE ? ”, Mémoire pour l’obtention du certificat d’Orthophoniste, Université Nice Sophia Antipolis-Faculté de Médecine-Ecole d’Orthophonie-Nice, 2015