Projet de Traitement Automatique du Langage Naturel (NLP)

Classification thématique automatique de documents

Sacha Chantoiseau Calvin Zegbeu

Master 1 Informatique
Année universitaire 2024-2025

Rapport de projet

Matière : Traitement Automatique du Langage (NLP)

Enseignant : *Elena Cabrio*

Université Côte d'Azur - UFR Sciences
Master 1 Informatique
Année universitaire 2024-2025

Table des matières

1 Introduction

Dans un contexte où la quantité d'informations textuelles disponibles en ligne et dans les bases de données ne cesse de croître de manière exponentielle, la capacité à organiser automatiquement les documents en fonction de leur contenu est devenue une nécessité incontournable. Les textes se multiplient dans des domaines variés tels que le droit, l'histoire, les sciences, la cuisine ou encore la communication personnelle. Face à cette masse hétérogène de contenus, il devient indispensable de développer des outils permettant de classer, trier et analyser ces documents de manière automatique, rapide et fiable.

Ce besoin s'inscrit pleinement dans le champ du Traitement Automatique du Langage Naturel (TAL ou NLP pour *Natural Language Processing*), une branche de l'intelligence artificielle qui se concentre sur l'interprétation et la génération automatique du langage humain par les machines. Parmi les nombreuses tâches que recouvre le NLP, la classification de texte est l'une des plus fondamentales. Elle consiste à attribuer une ou plusieurs étiquettes (ou catégories) à un document textuel en fonction de son contenu sémantique.

Tâches classiques du NLP

Le NLP couvre un ensemble de tâches variées, parmi lesquelles :

— **Classification de texte** : catégoriser un document selon un ensemble de thématiques.

— **Analyse de sentiment** : identifier la polarité *positive, neutre, négative* d'un texte.

— **Extraction d'information** : détecter des entités, relations, dates, etc.

— **Traduction automatique**, **question-réponse**, **résumé automatique**, etc.

Dans ce projet, nous nous sommes intéressés à la classification de documents selon leur type ou leur origine thématique. Plus précisément, l'objectif est de concevoir un modèle capable d'identifier automatiquement si un document appartient au domaine juridique, historique, scientifique, culinaire, encyclopédique (Wikipedia), à un courriel, ou encore à un texte lié à l'intelligence artificielle.

2 Données

2.1 Choix des catégories

Nous avons retenu sept types de documents distincts :

— **Juridique**, **Historique**, **Scientifique**, **Recettes**, **Wikipedia**, **Courriel**, **IA**

2.2 Sources et collecte

— **Juridique** : `harvard-lil/cold-french-law`, champ `article_contenu_text`

— **Historique** : `Nadav/historical_texts`, champ `text`. Ces textes en vieil anglais ont nécessité la création d'un script de traduction vers un anglais moderne, afin de garantir une compréhension optimale par les modèles de traitement automatique.

— **Scientifique** : scraping sciencesetavenir.fr

— **Recettes** : scraping marmiton.org/Recettes

— **Wikipedia** : API/dumps

— **Courriel** : Enron Email Dataset (traduits)

— **IA** : articles/blogs/spécifiques Wikipedia

Par ailleurs, afin de garantir une homogénéité linguistique (c'est-à-dire que tous les textes soient dans la même langue) et d'entraîner nos modèles sur des données exclusivement en français, nous avons conçu un ensemble de scripts de traduction automatique pour convertir l'ensemble du corpus vers le français.

Dans le cas des textes historiques, qui étaient rédigés en vieil anglais, nous avons d'abord dû créer des scripts spécifiques, faits main, pour les traduire en anglais moderne, car ils étaient difficilement exploitables tels quels. Ensuite, ces textes modernisés, comme l'ensemble des autres documents non francophones, ont été traduits automatiquement en français.

Nous avons fait ce choix pour que toutes les données soient sur un pied d'égalité : avoir une seule langue (le français) permet de garantir la cohérence des représentations et de ne pas biaiser l'entraînement du modèle en mélangeant des langues ou des structures syntaxiques différentes.

2.3 Organisation

Organisation par répertoires :

```
/Data
    /IA
    /Histoire
    /Juridique
    /Mail
    /Recettes
    /Sciences
    /Wikipedia
```

3 Pré-traitement

— Nettoyage HTML, ponctuation, doublons

— Conversion en minuscules, tokenisation

— Suppression des stopwords, lemmatisation

Analyse morpho-syntaxique (POS tagging)

— Extraction de la catégorie grammaticale de chaque mot : nom, verbe, etc.

— Utile pour enrichir les représentations ou exclure certains mots fonctionnels

4 Modélisation

Notre approche de modélisation (c'est-à-dire la manière dont nous construisons un système capable de reconnaître la catégorie d'un texte) a suivi deux axes principaux : l'utilisation de modèles dits classiques, fondés sur des vecteurs de type TF-IDF, et l'expérimentation avec un modèle plus avancé de type Transformer, nommé CamemBERT, qui est spécifiquement adapté à la langue française.

Un **modèle classique** désigne ici un algorithme d'apprentissage automatique relativement simple et peu coûteux en ressources de calcul, comme la régression logistique ou les machines à vecteurs de support (SVM, pour *Support Vector Machine*).

Le **TF-IDF** (*Term Frequency - Inverse Document Frequency*) est une méthode de représentation des textes qui attribue un poids à chaque mot en fonction de sa fréquence dans un document, pondérée par sa rareté dans l'ensemble du corpus. Cette approche permet de mettre en évidence les mots importants d'un document tout en atténuant l'importance des mots fréquents et peu informatifs.

À l'inverse, les modèles **Transformers** sont des architectures de *deep learning* (apprentissage profond), qui utilisent des mécanismes d'attention pour modéliser les relations entre les mots d'un texte, quelle que soit leur position. Ils permettent de mieux comprendre le contexte global du document.

Dans ce projet, nous avons utilisé **CamemBERT**, une variante francophone de BERT (*Bidirectional Encoder Representations from Transformers*), qui est un des modèles Transformers les plus connus. CamemBERT a été préentraîné sur un large corpus de textes français, ce qui le rend particulièrement performant pour les tâches en langue française.

Ce double choix – modèles classiques et modèles Transformers – repose sur une stratégie comparative :

— D'une part, tester des approches **rapides à entraîner**, peu gourmandes en mémoire et computation, qui peuvent fournir une bonne première estimation.

— D'autre part, valider l'intérêt de méthodes modernes issues du deep learning, plus **précises** mais nécessitant davantage de ressources.

L'objectif était à la fois de démontrer la portée de méthodes simples sur des données bien préparées, et d'évaluer dans quelles mesures les modèles les plus récents peuvent améliorer significativement les performances sur une tâche de classification thématique de documents en français.

4.1 Représentations textuelles

TF-IDF (classiques)

— Vectorisation avec n-grammes (1,2), vocabulaire limité à 10 000 termes

— Choix motivé par sa simplicité, son efficacité sur des modèles linéaires, et sa lisibilité pour l'analyse des features importantes. Cette représentation permet une première exploration des données, une compréhension des mots discriminants par classe, et un entraînement rapide même sur une machine standard.

CamemBERT (modèle BERT pour le français)

— Modèle pré-entraîné, sens contextuel des mots, fine-tuning possible

— CamemBERT repose sur une architecture de type Transformer qui capture le contexte global d'une phrase. Il est particulièrement adapté à la langue française, ce qui permet d'optimiser les performances sans avoir besoin d'entraîner un modèle from scratch. Il offre une compréhension fine de la sémantique, indispensable pour distinguer des classes stylistiquement proches.

4.2 Modèles testés

— **Classiques** : Naive Bayes, SVM, Regression Logistique, Random Forest

  - Ces modèles permettent une première analyse rapide, sont facilement interprétables et peu coûteux computationnellement. Ils servent de référence (baseline) pour comparer les améliorations obtenues avec des modèles plus complexes.
- **Deep learning** : fine-tuning de `camembert-base`
  - Le fine-tuning de CamemBERT nous a permis d'adapter un modèle pré-entraîné à notre tâche spécifique de classification. Nous avons opté pour cette approche car elle permet de tirer parti de représentations puissantes tout en adaptant les couches finales à nos classes. D'autres modèles plus lourds comme GPT ou des architectures RNN n'ont pas été retenus pour des raisons de complexité et d'adaptabilité moindres au français dans notre contexte.

4.3 Évaluation

Pour comparer objectivement les modèles testés, nous avons utilisé plusieurs métriques standard en apprentissage automatique supervisé pour la classification.

- **Accuracy** : il s'agit du pourcentage de documents correctement classés parmi l'ensemble du corpus test. C'est une métrique simple à interpréter mais qui peut être trompeuse en cas de classes déséquilibrées.
- **F1-score macro** : cette métrique combine précision (exactitude des prédictions positives) et rappel (capacité à identifier tous les exemples pertinents), puis calcule une moyenne égale entre les classes. Elle est particulièrement utile pour évaluer les performances globales dans un contexte multi-classes avec un léger déséquilibre.
- **Matrice de confusion** : tableau croisant les prédictions et les réalités, utile pour visualiser quelles catégories sont souvent confondues. Cela permet d'identifier des confusions systématiques entre classes (par exemple entre Wikipedia et Sciences).

Ces métriques nous permettent non seulement d'avoir une vue d'ensemble des performances, mais aussi d'identifier les points faibles précis de chaque modèle afin d'envisager des améliorations ciblées.

5 Résultats

Cette section présente les résultats obtenus par nos modèles de classification sur le corpus test, en s'appuyant sur les métriques d'évaluation définies précédemment (accuracy, F1-score macro, matrice de confusion). Elle permet d'observer les performances globales, mais aussi d'analyser les comportements du modèle sur des exemples concrets issus de notre corpus hétérogène.

De manière générale :

- Les modèles classiques (comme la régression logistique ou le SVM) ont obtenu de bonnes performances sur les catégories lexicalement bien définies, comme les recettes (`Marmiton`) ou les textes juridiques formels.
- Le modèle CamemBERT, grâce à sa compréhension contextuelle du langage, a surpassé les modèles classiques sur des catégories plus ambiguës comme `Wikipedia`, `IA` ou `Sciences`.
- La matrice de confusion met en évidence des confusions récurrentes entre certaines catégories proches : `Wikipedia` est souvent prédit pour des textes historiques, scientifiques ou juridiques, notamment lorsque ces derniers adoptent un ton neutre et informatif.

Nous avons ainsi observé que la classe `Wikipedia` agit fréquemment comme une catégorie "refuge", absorbant les textes présentant une structure descriptive ou un style encyclopédique, même lorsqu'ils relèvent clairement d'un autre domaine. Ce phénomène, que nous qualifions de *biais stylistique*, est notamment visible dans les documents issus des catégories `Histoire`, `Sciences` ou `Juridique`, mal reconnus lorsqu'ils sont rédigés sous forme d'archives ou de résumés formels.

En revanche, les catégories aux structures lexicales marquées, comme `Marmiton` (recettes) ou `Juridique` (textes de loi), sont très bien détectées avec des scores supérieurs à 98 %. Cela montre que le modèle s'appuie fortement sur des motifs linguistiques récurrents pour établir sa prédiction.

La classe `IA` reste difficile à modéliser, notamment en raison du chevauchement lexical avec `Sciences` ou `Wikipedia`, et du manque de termes spécialisés dans certains textes. Quant à la classe `Mail`, elle demeure la plus difficile à détecter, probablement en raison de la forte hétérogénéité des styles et des contenus dans les courriels.

Ces observations confirment que le modèle est particulièrement efficace pour distinguer des catégories à forte identité textuelle, mais qu'il rencontre encore des difficultés pour distinguer des domaines plus proches ou lorsque le style du texte brouille les repères thématiques.

6 Conclusion

Ce projet nous a permis de concevoir un système de classification automatique capable de distinguer plusieurs types de documents textuels en langue française, en appliquant des techniques de traitement automatique du langage naturel (NLP). À travers la construction d'un corpus multi-thématique, le nettoyage des données, l'entraînement de plusieurs modèles — classiques et profonds — ainsi que l'analyse des performances, nous avons exploré les différentes étapes clés d'un pipeline complet de classification de texte.

Nous avons pu observer que les modèles classiques, tels que la régression logistique ou les SVM, offrent une première approche robuste et rapide, notamment lorsqu'ils sont couplés à une vectorisation de type TF-IDF. Cependant, ces modèles atteignent rapidement leurs limites lorsqu'il s'agit de capturer des nuances contextuelles dans le langage. C'est dans ce contexte que l'intégration de modèles modernes, comme CamemBERT, a montré toute sa pertinence. Ce modèle, basé sur l'architecture Transformer et préentraîné sur des corpus français, a permis d'améliorer significativement les résultats, en particulier sur des catégories plus subtiles ou à la frontière entre plusieurs thématiques.

L'analyse qualitative des résultats a mis en évidence certains biais : le modèle tend à privilégier la classe Wikipedia dans les cas ambigus ou lorsqu'un texte est rédigé de manière neutre et descriptive. Cela révèle l'importance du style rédactionnel dans la prise de décision du classifieur. À l'inverse, les textes à forte identité lexicale (comme les recettes ou les documents juridiques) sont très bien identifiés, même par les modèles classiques. Des catégories comme IA ou Mail se sont révélées plus complexes à modéliser, en raison soit d'un vocabulaire trop générique ou partagé, soit d'une grande hétérogénéité des contenus.

En somme, ce projet nous a permis de :

— mettre en œuvre un pipeline complet de classification thématique en NLP,

— comparer différentes représentations et architectures de modèles,

— construire un corpus multicatégoriel et multiformat,

— réfléchir aux biais, aux limites et aux besoins d'une analyse qualitative fine.

Ce que nous aurions pu améliorer :

— Raffiner davantage la sélection et la préparation des données, notamment en équilibrant mieux les styles rédactionnels dans chaque classe.

— Multiplier les exemples pour les classes les plus faibles ou ambiguës (IA, Mail).

— Mieux gérer la traduction automatique en s'assurant que la conservation du sens est cohérente.

— Automatiser les analyses d'erreurs et la validation croisée, pour un meilleur suivi des performances.

— Intégrer une étape de détection du style rédactionnel pour désambiguïser les textes encyclopédiques.

Perspectives futures :

— Mettre en place un système de classification *multi-label*, capable d'attribuer plusieurs étiquettes à un même texte — utile pour les documents hybrides (ex. : un article scientifique d'opinion).

— Étendre le système à un corpus multilingue ou entraîner un modèle sans traduction préalable en s'appuyant sur des architectures multilingues comme XLM-R.

— Enrichir les représentations des textes avec des métadonnées (ex. : source, date, auteur) ou des éléments structurels (titres, paragraphes, balises).

— Déployer le modèle sous forme de prototype ou d'API REST accessible via une interface utilisateur web, pour faciliter l'intégration dans des outils de veille ou de tri documentaire.

— Tester d'autres modèles performants comme DistilBERT (plus léger) ou XGBoost (pour une approche tabulaire hybride).

Ce projet nous a donc offert un cadre riche pour mettre en pratique les notions théoriques du NLP tout en affrontant des problématiques concrètes de terrain. Les résultats obtenus sont prometteurs, mais révèlent aussi la complexité de la classification thématique dans un contexte réel, avec ses zones floues, ses biais sémantiques et ses défis d'interprétation. C'est cette complexité qui rend le domaine si passionnant à explorer, et qui nous encourage à aller plus loin.