Contributions à la création de corpus et de modèles d'apprentissage profond pour les données textuelles multilingues

Anna Pappa

LIASD

UNIVERSITÉ PARIS 8 VINCENNES-SAINT-DENIS

Université Paris 8
Laboratoire d'Intelligence Artificielle et Sémantique des Données
École doctorale nº 224 : Cognition, Langage, Interaction

Habilitation à Diriger des Recherches

Spécialité : Informatique

présentée par

Anna PAPPA

Contributions à la création de corpus et de modèles d'apprentissage profond pour les données textuelles multilingues

Soutenue le 22 avril 2024 devant le jury :

NICOLAS JOUANDEAU	Garant/Rapporteur	PR UNIVERSITÉ PARIS 8
CHRISTIAN BOITET	Rapporteur	PR ÉMERITE UNIVERSITÉ GRENOBLE ALPES
BENOÎT CRABBÉ	Rapporteur	PR UNIVERSITÉ PARIS CITÉ
MAX SILBERZTEIN	Examinateur	PR UNIVERSITÉ DE FRANCHE-COMTÉ
TITA KYRIACOPOULOU	Examinatrice	PR UNIVERSITÉ GUSTAVE EIFFEL
TRISTAN CAZENAVE	Examinateur	PR UNIVERSITÉ PARIS DAUPHINE
JEAN-JACQUES BOURDIN	Examinateur	PR UNIVERSITÉ PARIS 8

Cette HDR est dedicacée à la mémoire de JEAN MÉHAT

Remerciements

Je tiens à exprimer ma profonde gratitude envers Nicolas Jouandeau, pour avoir accepté d'être mon garant et rapporteur pour cette habilitation à diriger les recherches.
Je souhaite également exprimer ma gratitude envers Revekka Kyriakoglou et Alice Millour, nos nouvelles collègues qui ont accepté de fonder avec moi le groupe de recherche 'GLAÇON'. Leurs lectures et évaluations ont été précieuses, mais surtout, leur soutien et leurs encouragements m'ont aidé à avancer.
Je suis particulièrement reconnaissante envers mes rapporteurs Christian Boitet et Benoît Crabbé pour m'avoir fait l'honneur d'accepter de lire et évaluer mon travail.
Je remercie chaleureusement Tita Kyriakopoulou, Max Silberztein, Tristan Cazenave et Jean-Jacques Bourdin pour avoir accepté de faire partie du jury de mon habilitation à diriger des recherches. Leur présence est un honneur pour moi.
Mes collègues Françoise Balmas, Jacqueline Signorini et Isis Truck méritent toute ma gratitude pour m'avoir fait confiance d'encadrer ou co-encadrer les thèses de Maroua Boudabous, Lisa Medrouk, et Mohammed-Amine Abchir.
Je tiens à remercier tout particulièrement Farès Belhadj, dont le soutien infatigable a été essentiel pour lever mes doutes.
Un immense merci à l'ensemble de l'équipe PASTIS, anciens comme nouveaux, présents ou partis : Adrien Revault d'Allonnes, Pablo Rauzy, Jean-Noël Vittaut, Sylvia Chalençon et Vincent Boyer, ainsi qu'à Sven de Felice, Benjamin Dupont, Louis Falissard et Alexandros Singh. Je suis reconnaissante pour tous les moments partagés, bons et encore meilleurs.
Les travaux rassemblés dans cette habilitation à diriger des recherches n'auraient pas vu le jour sans le soutien constant et les riches échanges que j'ai eus avec plusieurs personnes tout au long de ce parcours. Un remerciement spécial va à mes doctorant-e-s et autres étudiant-e-s, qui sont une source constante de motivation pour moi et me poussent à sans cesse m'améliorer.
Un grand merci s'adresse à tou-te-s mes ami-e-s, trop nombreux-ses pour être tou-te-s cité-e-s ici, ainsi qu'à mes parents pour leur soutien indéfectible.
Enfin, je ne saurais assez remercier Pascal, Néphèli, Matthaios et Francis d'être là à chaque étape importante de ma vie.

Résumé

Cette habilitation à diriger des recherches se situe à la croisée de l'informatique et de la linguistique, condensant près d'une décennie de travaux qui explorent cette interface multidisciplinaire. Elle est articulée autour de trois axes majeurs.

Le premier axe présente la création de corpus multilingues et thématiques, spécifiquement dédiés à l'analyse des opinions et des aspects. Les méthodologies et les outils développés, qui visent à réduire le bruit et les irrégularités dans les données, servent de socle pour les expérimentations ultérieures.

Le deuxième axe explore l'analyse de sentiment par le biais de modèles hybrides. Ces modèles, qui combinent des Réseaux Neuronaux Convolutionnels (CNN) et Récurrents (RNN) tels que les LSTM, atteignent une précision qui oscille entre 90% et 100% sur divers corpus multilingues non annotés.

Le troisième axe aborde l'annotation d'aspects en utilisant une architecture combinée BiLSTM-CNN-CRF, enrichie par des techniques d'apprentissage profond comme l'apprentissage actif et l'apprentissage par transfert. Ces méthodes se sont avérées particulièrement efficaces pour améliorer les performances des modèles en contextes de rareté de données ou de langues peu représentées.

En synthèse, cette habilitation constitue une contribution pertinente à la fusion des méthodologies informatiques et linguistiques, exploitant des jeux de données sur mesure et des architectures de modèles hybrides pour résoudre des défis complexes en annotation sémantique et en analyse multilingue. Ce travail ouvre également la voie à des recherches futures pour affiner ces méthodologies dans des scénarios encore plus exigeants ou moins étudiés.

Table des matières

Table des figures

Liste des tableaux

Liste des algorithmes

Chapitre 1

Introduction

Dans le cadre de cette habilitation à diriger des recherches, nous exposons nos avancées et expérimentations centrées sur l'automatisation des processus de résolution de problèmes associés à diverses tâches en traitement du langage naturel. Notre recherche est motivée par la conjugaison de deux aspects fondamentaux de la résolution de problèmes en informatique appliquée à la linguistique : l'aspect cognitif-conceptuel et l'aspect pratique-technique. Le premier consiste en une réflexion théorique, où nous nous engageons dans des activités telles que l'analyse, la conception et la modélisation. Le deuxième consiste en une réflexion pratique de la programmation et de la mise en oeuvre de solutions concrètes. Ensemble, ces deux aspects forment un continuum dans la résolution de problèmes en informatique, en général, et pour notre recherche, en *informatique linguistique*, une approche du domaine du Traitement Automatique du Langage Naturel (TALN) dont la description est détaillé dans [164].
Nos travaux adressent des défis liés à l'analyse et au traitement de données textuelles. Ce processus intégré couvre depuis la collecte initiale de données jusqu'à leur structuration en ensembles de données exploitables, en passant par des analyses morphologiques et syntaxiques, pour finalement aboutir à des analyses sémantiques centrées sur les sentiment et les caractéristiques informatives inhérentes aux entités linguistiques. De plus, étant donné le caractère multilingue des données avec lesquelles nous travaillons, nous cherchons à élaborer des solutions qui soient transposables à divers contextes linguistiques.
Avant de détailler nos contributions dans les chapitres suivants, il est important de noter que cette habilitation présente une sélection de nos travaux réalisés entre 2014 et 2022. Ces travaux s'inscrivent dans un continuum de recherches initié avec ma thèse soutenue en 2003 à l'Université Paris 8. Nos approches récentes visent à couvrir l'ensemble des étapes du traitement du langage naturel, allant de la création de jeux de données jusqu'à leur analyse, leur balisage éventuel, et leur utilisation dans des modèles d'apprentissage automatique. Plusieurs travaux de recherche ayant abouti à des publications [130, 131, 132, 134, 133, 135, 137, 2, 3, 116], ne sont pas abordés dans cette habilitation.

Cependant, nous souhaitons exposer brièvement deux thématiques : la première concerne une approche computationnelle pour l'analyse morpho-syntaxique et la deuxième s'intéresse à la sémantique floue. Bien que cette dernière soit un peu éloignée de nos axes de recherche principaux, ces deux sujets ont été explorés dans nos travaux antérieurs.

1.0.1 Approche computationnelle

La première thématique porte sur le développement de méthodes et d'outils, à base de règles morpho-syntaxiques, pour la désambiguïsation, l'étiquetage grammatical et la génération de lexiques à partir des corpus non annotés. Ces travaux [1] sont axés sur l'analyse automatique de surface (*shallow parsing*) des corpus non annotés pour des langues peu dotées en ressources [132, 137]. Cette phase de recherche se situe dans l'approche de la *linguistique informatique* [2]. Dans cette optique, partant des postulats de Fodor [47] selon lesquels les caractéristiques grammaticales deviennent partie intégrante des représentations lexicales statiques, nous observons divers phénomènes linguistiques, tels que les changements morphologiques, l'homonymie, la distribution des unités et la structure grammaticale d'une phrase.
Dans [130, 134] nous abordons la problématique de résolution des ambiguïtés inhérentes à l'annotation morpho-syntaxique automatisée. Ces ambiguïtés sont particulièrement présentes dans les cas d'homonymes rencontrés dans plusieurs langues, comme c'est le cas des articles définis et des pronoms personnels en français et en grec.
Nous proposons une méthodologie computationnelle à trois étapes, combinant des règles morpho-syntaxiques et des statistiques, basée sur l'analyse distributionnelle proposée par Harris [57], pour lever les ambiguïtés et générer des lexiques. Chaque règle possède une "priorité" et une "valeur d'efficacité". Les valeurs de priorité déterminent dans quels cas la règle est la plus appropriée à utiliser, tandis que les valeurs d'efficacité établissent le degré de justesse de la règle appliquée. Les résultats de chaque étape du traitement sont filtrés en fonction de leur fiabilité. Cette première étape se poursuit par une deuxième étape de désambiguïsation suivie d'une dernière étape d'annotation. À l'issue de cette dernière étape, un lexique est généré à partir des cas jugés fiables. Ce lexique est composé uniquement de mots dont la catégorie grammaticale est bien établie, qu'il s'agisse de noms ou de verbes [133, 135].
L'une des contributions [3] de cette recherche est la robustesse de l'algorithme proposé, démontrée par son efficacité sur des langues structurellement distinctes telles que le français et le grec. Cette polyvalence linguistique ne se limite pas à la performance, mais

1. Il s'agit d'une poursuite des recherches initiées pendant ma thèse soutenue en 2003.

2. Selon [164] il y a deux approches majeures dans le domaine du Traitement Automatique des Langues, à savoir la *linguistique informatique* et l'*informatique linguistique*. La linguistique informatique formalise les descriptions linguistiques et utilise des algorithmes pour les modèles formels, tandis que l'informatique linguistique utilise des modèles stochastiques pour analyser de grands corpus textuels et en déduire des régularités.

3. START, le sujet de thèse sur le système de reconnaissance, est utilisé comme support pédagogique pour le cours d'Ingénierie des Langues en licence informatique.

aussi à la capacité de notre algorithme à désambiguïser des phénomènes linguistiques complexes qui sont spécifiques à chaque langue.

1.0.2 Sémantique floue

La deuxième thématique concerne les travaux menés dans le cadre du projet ANR SALTY [4]. Notre participation combine la sémantique avec la théorie des ensembles flous et le concept des variables linguistiques pour traiter des dialogues entre humains et dispositifs de communication. Nous visons à affiner l'interprétation des dialogues pour une réaction appropriée, comme le déclenchement d'alertes.

Les *fuzzy linguistic 2-tuples* (FL2T) représentant des couples (étiquette, valeur), sont employés pour une quantification précise des nuances sémantiques [60]. Cette approche avance au-delà de l'inférence de Zadeh [196], éliminant les biais dus aux approximations symboliques et fournissant une représentation adéquate des termes linguistiques, même lorsqu'ils présentent une distribution non uniforme.

En pratique, les FL2T permettent une représentation précise et nuancée de concepts tels que la distance, où des termes comme 'proche', 'très proche', et 'loin' sont contextualisés sur un axe sémantique. Cela est démontré dans la figure 1.1, illustrant la manière dont ces termes linguistiques sont appliqués dans le domaine de la géolocalisation pour évaluer la proximité entre, par exemple, un piéton et un véhicule.

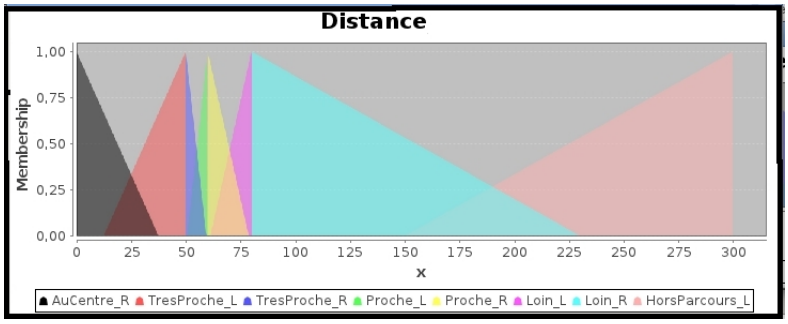


FIGURE 1.1 – Les nuances sémantiques

En amont, une partition floue avec notre modèle de couple (étiquette, valeur) a été réalisée par des experts pour chaque élément de données imprécis. Par exemple, l'expert attribue 5 termes : A, B, C, D, E qui sont par défaut uniformément répartis sur leur axe. Cependant, en recherchant des synonymes [5] et la distance entre les termes, notre algorithme permet d'obtenir 5 sacs de synonymes. Les taux de ressemblance entre les sacs sont ensuite

4. Projet ANR SALTY (Self-Adaptive very Large disTributed sYstems) (ANR-09-SEGI-012) https://anr.fr/Projet-ANR-09-SEGI-0012, https://hal.science/hal-00539093

5. Nous avons testé les synonymes proposés par le dictionnaire en ligne, CRISCO : https://crisco4.unicaen.fr/des/ (dictionnaire français)

calculés afin d'obtenir de nouvelles positions de A, B, C, D, E sur l'axe. En aval, les "jetons" sémantiques [6] (par exemple, "proche") sont exprimés à travers des couples (étiquette, valeur) et comparés à la partition floue. Les adverbes tels que "très" modifient leur couple associé à travers la valeur de $alpha$.
Notre approche :

- Prend en compte les étiquettes et leurs positions sur l'axe.
- Définit S = $\{(s_0, v_0), (s_1, v_1) \dots (s_n, v_n)\}$ où s_i est un label et v_i est sa position sur l'axe.
- Utilise le modèle de représentation en couples (étiquette, valeur) & les hiérarchies linguistiques.
- Inclut un partitionnement flou qui dépend :
  - des positions des étiquettes sur l'axe,
  - des distances entre deux étiquettes successives.

Les tests effectués démontrent la manière dont les partitions floues peuvent être adaptées en fonction des besoins contextuels. Cela met en valeur l'importance d'utiliser des méthodologies robustes et flexibles pour gérer les ambiguïtés et les imprécisions dans des applications en temps réel. Nos contributions [2, 3, 116], répondent à un double défi : premièrement, l'optimisation de l'interprétation sémantique dans des applications de TALN, et deuxièmement, la réduction des ressources computationnelles nécessaires pour réaliser cette interprétation dans le cadre de géolocalisation en temps réel.
De plus, notre recherche s'aligne avec le phénomène universel de la "fuzziness" (flou), tel que discuté dans les études linguistiques contemporaines. L'indétermination inhérente et l'imprécision dans le langage, analysées dans [105], résonnent avec les principes fondamentaux de la logique floue proposés par Zadeh. La nature relative et conditionnelle de la précision dans la traduction souligne la présence absolue et omniprésente du flou dans le langage — un concept qui est central dans notre approche utilisant les couples (*2-tuples*) linguistiques flous. Ces principes confirment la pertinence des ensembles flous et des FL2T pour gérer la nature nuancée et dynamique du langage, réaffirmant ainsi l'importance de la théorie de Zadeh dans les études linguistiques et de traduction modernes.

* * *

Après cette incursion dans nos travaux antérieurs, nous allons maintenant détailler les axes principaux de cette habilitation. Ils se concentrent essentiellement sur la création et l'annotation de corpus textuels multilingues, l'analyse de sentiment et l'annotation des aspects en utilisant des techniques d'apprentissage automatique et, plus particulièrement, d'apprentissage profond.

Le premier volet de ce manuscrit présenté dans le chapitre 2 aborde la question de la création de corpus textuels multilingues. Les corpus sont fondamentaux pour toute tâche d'apprentissage automatique. en TALN, car ils servent à la fois de terrain d'entraînement

6. Mots ayant un sens dans le texte.

et de banc d'essai pour les modèles. La performance du modèle est très dépendante de la qualité du corpus. Nous nous concentrons sur l'élimination du bruit, notamment les contenus inutiles et le contenu standardisé, lors de l'extraction des données textuelles du web.
Nous présentons une série d'outils et de méthodologies—REVSCRAP 2.2.1, DYCORC 2.2.2, et WEBT-IDC 2.2.3—qui, bien qu'ayant des approches et des spécificités distinctes, partagent l'objectif commun de créer des corpus de textes à partir du web.

REVSCRAP[115] est un outil spécialisé dans la construction de corpus d'avis d'utilisateurs relatifs à des produits spécifiques. Employant une interface interactive et une méthodologie séquentielle, ce système filtre les résultats issus des moteurs de recherche, élimine les publicités et les contenus non pertinents, et organise les données peu bruitées collectées en documents XML. Sa simplicité et son efficacité ont été démontrées par des résultats encourageants.

DYCORC [109] se démarque par son approche non supervisée et sa compétence à extraire des données sans bruit à partir de forums web. Il opère sans nécessiter de connaissance préalable ni de la langue en cours d'examen, ni des balises HTML. Bien qu'il soit plus chronophage que d'autres outils, comme BOOTCAT, il compense cette lacune par une précision et une F-mesure supérieures. DyCorC offre également la possibilité d'intégrer d'autres outils en vue d'améliorer sa performance globale.

WEBT-IDC [21] est un outil automatisé dédié à la collecte d'avis et de commentaires depuis des forums et des blogs. Sa particularité réside dans son indépendance vis-à-vis de l'architecture du site web d'où proviennent les données, à condition que le site dispose d'un élément de pagination, ainsi que dans sa capacité à assembler un corpus thématique multilingue. Pour optimiser ses performances, WEBT-IDC exploite des technologies *multithread* et accède directement aux sections pertinentes des sites web.

En ce qui concerne les contributions spécifiques, nos travaux se distinguent principalement par leur intérêt sur les langues sous-représentées en ligne comme le grec et l'arabe. De plus, nous avons accordé une attention particulière à la suppression du bruit associé aux données web, afin de générer des corpus qui ne requièrent pas de post-traitement pour l'apprentissage automatique par exemple.

Actuellement nous avons deux projets en cours chacun nécessitant un corpus spécifique pour les tâches que nous devons effectuer. Dans le cadre du projet MALANTIN (2019-2022 extended), nous collaborons avec des économistes pour créer un corpus web axé sur l'innovation dans 27 secteurs économiques spécifiques. Pour cibler les pages pertinentes, nous utilisons un ensemble de mots-clés. Un algorithme de filtrage des URL élimine les sources de bruit et catégorise les données selon le secteur et les mots-clés. Pour le projet CLEXIC (2023), nous constituons un corpus basé sur des projets de *crowdfunding*. Ce projet nécessitant une extraction à grande échelle, nous avons parallélisé la lecture des différentes arborescences HTML des projets. Ce corpus en construction servira de base pour les modèles d'apprentissage futurs. Le projet étant en cours, des ajustements sont prévus

pour améliorer la construction du corpus final. Le travail sur la création des corpus offre des outils et méthodes qui peuvent être réutilisés ou adaptés pour d'autres travaux scientifiques. Il témoigne d'une démarche de recherche engagée pour répondre aux besoins spécifiques de divers projets et pour étendre l'applicabilité de nos outils et méthodes à des contextes variés.

Après avoir exploré nos efforts continus dans la création et l'amélioration de corpus textuels, nous allons maintenant nous tourner vers un autre axe majeur de cette habilitation : l'analyse de sentiment à l'aide de techniques d'apprentissage profond.

Dans le chapitre 3 nous abordons l'analyse de sentiment et la classification thématique dans un contexte multilingue. Notre travail se distingue par son accent sur des données non pré-traitées, en particulier des avis utilisateurs provenant de divers secteurs. Nous utilisons des architectures avancées de réseaux neuronaux, telles que les CNN et les Recurrent Neural Network (RNN)—en particulier les LSTM—pour effectuer une analyse fine. Nous mettons en avant la complémentarité de ces différentes architectures, notamment dans leur capacité à identifier à la fois des corrélations locales et des dépendances contextuelles sur de longues distances. Les résultats qui s'appuient sur des corpus multilingues de critiques de restaurants et d'hôtels, attestent de l'efficacité de notre approche non supervisée.

Dans ce contexte, la problématique de notre recherche se divise en trois parties : la classification de la polarité des avis traités dans leur langue d'origine, la distinction entre différents types d'opinions dans des domaines sémantiquement proches (la restauration et l'hôtellerie) et l'utilisation de ces architectures pour extraire des caractéristiques pertinentes indépendamment de la langue. L'apprentissage profond s'avère être un outil particulièrement adapté à ces défis, grâce à sa flexibilité et sa robustesse.

Nos tests révèlent que nos modèles sont performants même sans le recours à des données pré-traitées [115, 113, 114]. Les architectures hybrides LSTM-ConvNet sont particulièrement efficaces, chaque type de réseau contribuant à l'extraction d'informations distinctes et complémentaires. Parmi les pistes pour des travaux futurs, nous envisageons notamment l'exploration de modèles plus sophistiqués capables de traiter différents niveaux de granularité dans l'analyse de sentiment, ainsi que l'intégration de tâches connexes telles que la détection de sarcasme et de l'ironie. Ce chapitre montre que la méthode non supervisée que nous proposons est à la fois robuste et adaptable, s'appliquant efficacement à des environnements linguistiques variés et complexes.

Après avoir exploré l'analyse de sentiment et la classification thématique dans un contexte multilingue, ce chapitre 4 se positionne comme une extension naturelle de ces travaux précédents. Alors que les chapitres précédents se sont concentrés sur des corpus multilingues non annotés et leur traitement à des fins de classification sentimentale, nous allons ici un peu plus loin, en introduisant une nouvelle dimension : l'annotation des aspects. Cette tâche vise à identifier et catégoriser des éléments plus granulaires comme les attributs ou

caractéristiques spécifiques d'un produit ou d'un service. Cette nouvelle dimension permet d'offrir une analyse plus riche et plus détaillée, ce qui est très intéressant dans le cadre des applications commerciales. Pour réaliser cette tâche, une méthode hybride combinant l'apprentissage par transfert et l'apprentissage actif est introduite [22]. Cette approche se distingue par son efficacité à annoter des corpus textuels multilingues sans nécessiter de prétraitement. Elle permet d'enrichir ces corpus, transformant des données textuelles brutes en ensembles structurés, particulièrement utiles pour les langues qui manquent de ressources annotées.

Le chapitre souligne l'importance croissante de l'annotation de granularité fine pour répondre aux besoins des algorithmes d'apprentissage automatique de plus en plus sophistiqués. Dans le contexte de l'annotation des aspects, deux types d'aspects sont particulièrement mis en évidence : les aspects explicites et les aspects implicites. Ce travail se concentre sur l'identification des aspects explicites, une tâche souvent abordée comme un problème d'étiquetage séquentiel supervisé. Ce chapitre contribue à une vision plus large qui vise à établir un cadre unifié pour l'exploitation approfondie des données textuelles multilingues. Il s'agit d'une démarche qui synthétise les avancées à la fois en matière d'analyse de sentiment et d'annotation des aspects, ce qui ouvre la voie à de nouvelles applications dans le domaine du traitement du langage naturel.

Organisation du manuscrit

Ce manuscrit se compose de cinq chapitres, introduction et conclusion incluses. Le chapitre 2 expose les méthodologies qui ont conduit à la mise au point d'outils capables d'extraire des données textuelles pertinentes à partir de pages web sans bruit. Les chapitres 3 et 4 mettent l'accent sur diverses tâches de TALN, en employant des modèles d'apprentissage profond pour leur entraînement et évaluation sur des corpus multilingues. Ces corpus sont générés grâce aux outils décrits dans le chapitre 2.

Le chapitre 3 se focalise sur l'analyse de sentiment et la classification thématique dans des contextes multilingues, avec une attention particulière portée à l'utilisation de données non-prétraitées. Cette orientation vers des données multilingues démontre la robustesse de nos méthodes pour extraire des caractéristiques pertinentes sans nécessiter une connaissance spécifique de la structure linguistique ou d'annotations.

Le chapitre 4 prolonge cette présentation de nos travaux de recherche en explorant une dimension complémentaire et essentielle : l'annotation d'aspects dans des corpus multilingues non-annotés.

Enfin, la conclusion 5 synthétise les contributions présentées dans les chapitres de ce manuscrit, tout en esquissant les travaux en cours et les futures directions de recherche que nos expérimentations pourraient inspirer.

Chapitre 2

Création de corpus

Sommaire

Dans ce chapitre nous décrivons nos travaux de recherche liés aux approches et outils permettant la collecte fiable (*sans bruit*) [1], des données textuelles issues du web. Nous présenterons trois outils que nous avons spécifiquement conçus pour répondre aux besoins de recherches bien définis et nous offrons une vue d'ensemble des différentes méthodologies et outils développés pour la création de corpus. Le chapitre s'ouvre avec une introduction 2.1, qui délimite le champ d'application des corpus et des *datasets* 2.1.1, et établit la problématique autour de la création de corpus 2.1.2.

1. Par 'bruit' nous désignons tous les éléments non pertinents ou indésirables qui peuvent être récupérés lors de la collecte de données tels que des éléments de navigation, des publicités et des modèles [85], ce qui rend le processus de collecte sur le web complexe. Nous approfondirons la notion du 'bruit' lorsque nous aborderons les corpus pour le *Machine Learning (ML).*

Le cœur du chapitre est consacré aux outils développés 2.2. Premièrement, le REVSCRAP 2.2.1 est exploré en détail, de son état de l'art 2.2.1 à la description du système incluant l'algorithme 2.2.1 et les filtres 2.2.1, jusqu'à la présentation des résultats 12. Deuxièmement, nous avons le DYCORC 2.2.2, qui suit la même structure narrative, ajoutant des expérimentations 2.2.2 et une évaluation comparative 2.2.2. Troisièmement, le WEBT-IDC 2.2.3 est présenté, abordant son architecture complexe à travers une série de composants et de modules 2.2.3–8, et concluant avec les expériences menées 8 et leur évaluation 8.

Enfin, la section sur les contributions et perspectives 2.3 résume les contributions apportées 2.3.1 et esquisse les travaux actuels et les orientations futures 2.3.2.

2.1 Introduction

Depuis les premières heures de la linguistique de corpus jusqu'à aujourd'hui, la manière dont nous percevons et utilisons les corpus a subi d'importantes mutations. Aux débuts de cette discipline, alors que l'ère informatique en était encore à ses balbutiements, des visionnaires comme John Sinclair pressentaient déjà l'avènement de l'interaction homme-machine[71]. Il prédisait un futur où l'homme pourrait communiquer librement avec sa machine. Cependant, il soulignait que cette réalité était encore éloignée, tout en admettant qu'il était possible de commencer à imaginer les voies pour y parvenir. [2]

Sinclair voyait au-delà de la technologie de son époque. Pour lui, le corpus n'était pas qu'une simple collection de textes, mais plutôt un reflet des complexités et nuances de la langue. Il soulignait combien un corpus conséquent est bénéfique pour l'analyse linguistique. En effet, une grande collection de textes met rapidement en lumière les imperfections et lacunes d'une analyse linguistique, exigeant ainsi une rigueur méthodologique accrue.

Aujourd'hui, la création et l'utilisation des corpus dans le domaine du TALN, ont pris une dimension encore plus cruciale. Ils sont la pierre angulaire de nombreux travaux, allant de la modélisation sémantique à la traduction automatique. Les corpus, qu'ils soient généralistes, thématiques, mono ou multilingues [3], servent non seulement de base d'entraînement pour les modèles, mais aussi de moyen d'évaluation pour mesurer leurs performances.

Cela dit, le défi de créer des corpus pertinents et fiables est toujours d'actualité. Ces collections textuelles sont une mine d'or pour la recherche linguistique, offrant un panorama sur les variations, structures, et évolutions de la langue au fil du temps.

Face à l'explosion des données disponibles en ligne, une technique s'est particulièrement

2. Il est intéressant de constater que les prédictions de Sinclair ont pris forme plus rapidement que prévu, vu que des interfaces comme ChatGPT[24] d'OpenAI sont la preuve vivante de cette vision.

3. Les corpus multilingues jouent un rôle crucial dans la mise en œuvre de modèles de TALN capables de gérer plusieurs langues. Des initiatives majeures telles que Common Crawl [41] et OPUS [74] ont pavé la voie à cet égard.

démarquée pour faciliter la création de corpus : Le terme 'web scraping' [4] désigne la technique de l'extraction automatisée de données depuis des sites web. Plus qu'une simple méthode, il s'agit d'un véritable outil d'ingénierie qui, en automatisant la collecte de données textuelles, ouvre la porte à la création de vastes corpus à partir des richesses du web.

2.1.1 Corpus et jeux de données

Le World Wide Web est devenu un élément incontournable du paysage numérique contemporain, hébergeant une croissance exponentielle de données variées, allant du texte aux images, en passant par les vidéos et les sons. Il est souvent perçu comme une mine d'informations inépuisable [111], offrant une multitude de données dans diverses langues et formats [15], [81]. En tant que tel, il s'est avéré être une source inestimable pour la création de corpus [6], [46], notamment en raison de l'interaction numérique croissante entre les utilisateurs à travers les blogs, les forums et les réseaux sociaux. Ces plateformes abritent une multitude de commentaires, d'opinions et de revues, faisant d'elles des terrains fertiles pour la collecte de données en vue de l'analyse de sentiment et de l'analyse d'opinion. Cette richesse constitue une mine d'informations pour la recherche linguistique, en particulier pour la constitution de corpus destinés à diverses analyses. [180] souligne l'efficacité de l'utilisation du web pour la construction de corpus dédiés à la traduction automatique de dialogues. Il démontre que ces corpus issus du web sont supérieurs aux corpus de parole transcrits et annotés traditionnellement utilisés. Cependant, la transformation du web en un vaste corpus a suscité de vifs débats parmi les chercheurs. Certains ont questionné sa représentativité, son équilibre et sa conception [5]. Malgré les critiques, l'attrait d'utiliser le web comme corpus est indéniable [110], en particulier pour sa capacité à fournir des données en temps réel, flexibles, multilingues et personnalisables.

Aussi, la collecte d'informations pertinentes à partir du web s'avère être une tâche complexe [56]. Les pages web sont souvent encombrées de matériel non pertinent, tels que les éléments de navigation, les publicités, les modèles, et bien plus encore. Cette complexité est exacerbée par la nature hétérogène des pages web, leur architecture changeante, et la présence abondante de bruit et de redondances.
Avant de plonger dans les nuances et les spécificités de cette exploration, il convient de définir deux termes centraux à notre discussion : *corpus* et *jeu de données* (ou *dataset* en anglais).

4. Bien que les termes 'moissonage' ou 'raclage' soient parfois suggérés en français [190], nous privilégierons l'expression 'extraction des données', fréquemment utilisée avec l'ajout 'à partir du web'. Il est important de noter que l'anglicisme 'web scraping' reste néanmoins le terme le plus répandu.

5. Bien que le web soit une source inépuisable d'informations linguistiques, il est aussi le reflet de la variabilité, des tendances éphémères et des répétitions inhérentes à la communication humaine. Dans [146] dédié aux corpus et à leur représentativité, de nombreux chercheurs ont soulevé des interrogations pertinentes à ce sujet.

Un *corpus* est traditionnellement défini comme un ensemble de données sélectionnées et rassemblées pour intéresser une même discipline [117]. Pour la linguistique computationnelle il s'agit d'une *collection structurée de textes* souvent représentative d'un type particulier de production linguistique [6] qui reflète les critères de conception de Sinclair[51].

Le *jeu de données* ou *dataset* en anglais, fait référence à une collection organisée de données textuelles, souvent accompagnée d'attributs spécifiques et de valeurs [11].

Les corpus et les jeux de données, sont le matériau principal de nombreux projets d'intelligence artificielle qui reposent sur des données, utilisant des techniques d'apprentissage automatique dans tous les aspects de l'enseignement aux machines pour qu'elles puissent effectuer des opérations complexes sur le langage naturel. Étant donné que les jeux de données sont au cœur de ces projets, nous avons constamment besoin de créer de vastes collections de corpus de haute qualité, composées de données multilingues rapides, flexibles, spécialisées et personnalisées provenant du web[78]. Un moteur de recherche [7] a même été créé pour aider à découvrir tous les jeux de données existants sur le web, sans préciser si l'utilisation des jeux de données est libre ou non.

Dans le contexte de l'apprentissage automatique, un dataset est spécialement conçu pour l'entraînement, le test et la validation de modèles, le rendant essentiel à la mise en œuvre et à l'évaluation des algorithmes. Dans ce contexte, les corpus web, qu'ils soient extraits de forums, de blogs ou d'autres plateformes, ont pris une importance primordiale, notamment pour l'opinion mining et l'analyse de sentiment. Cependant, leur utilité est souvent entravée par la présence de *bruit*, comme des éléments redondants, des publicités, et d'autres contenus non pertinents. De plus, la nature dynamique des sites web, avec leurs fréquentes mises à jour et modifications, rend l'extraction de contenu web complexe.

Face à ces défis, notre recherche vise à développer des méthodologies et des outils permettant de constituer des corpus thématiques pertinents et ciblés à partir du Web, tout en minimisant voire éliminant le bruit et en assurant une mise à jour régulière pour refléter les évolutions constantes du paysage numérique.

Nous présentons trois outils permettant la collecte automatique des données textuelles à partir des pages web, en commençant par le premier outil qui a été conçu pour créer un corpus multilingue des avis des utilisateurs pour les besoins de recherche sur l'analyse de sentiments.

2.1.2 Problématique

La création d'outils de collecte de données textuelles à partir du web reste d'actualité et constitue un domaine de recherche actif. La création d'un outil universel capable de 'racler'

6. Selon [71] *les ensembles de données* en langage naturel sont appelés *des corpus*, et *un seul ensemble* de données annoté avec la même spécification est appelé un *corpus annoté*. Les corpus annotés peuvent être utilisés pour entraîner des modèles de ML.

7. (https://datasetsearch.research.google.com/) pour chercher le dataset qui nous intéresse.

toutes les pages web sans bruit est un défi difficile à relever. Il est cependant possible de développer des outils spécialisés pour des types de pages web spécifiques ou pour des domaines d'intérêt particuliers [8]. Ces outils peuvent être adaptés aux structures et formats de données couramment utilisés dans ces contextes spécifiques. La collecte de données à partir du web est complexe et comporte des défis significatifs. Nous présentons des outils que nous avons créé abordant des aspects comme :

- La structure et les formats de données variés : les pages web peuvent avoir des structures différentes, des formats de données variés et des langages de balisage tels que HTML, XML, JSON, etc. Créer un outil universel capable de 'racler' toutes les pages web sans bruit est difficile en raison de cette diversité.
- Le dynamisme des sites web : de nombreux sites web utilisent des technologies dynamiques comme JavaScript pour générer ou modifier le contenu à la volée. Cela peut rendre la collecte de données plus complexe, car il est nécessaire de traiter le rendu dynamique des pages pour récupérer l'ensemble des informations souhaitées.
- Le bruit et la qualité des données : lors de la collecte de données à partir du web, il est courant de rencontrer du bruit, des publicités, des liens indésirables ou des informations non pertinentes. Nettoyer et filtrer les données collectées pour garantir leur qualité nécessite des efforts supplémentaires.
- La politiques de sites web et respect de l'éthique : la collecte de données à partir du web nécessite de respecter les politiques des sites web, qui peuvent inclure des restrictions d'accès, des conditions d'utilisation, des limites de fréquence, etc. Il est essentiel de respecter ces politiques et de veiller à ne pas abuser ou perturber les sites web cibles.

La principale problématique consiste à déterminer le contenu pertinent dans les pages cibles, indépendamment de la langue et la structure de la page et l'extraire *sans bruit* afin de construire des corpus prêts à l'utilisation.

2.2 Les outils développés

2.2.1 REVSCRAP (*Review Scraper*)

L'objectif principal de cette recherche a été de créer automatiquement un corpus thématique à partir de données textuelles spécifiques pour le développement ultérieur d'outils d'analyse d'opinions. Dans ce cadre, nous avons conçu REVSCRAP, un outil destiné à la collecte de pages web de critiques pour la création de corpus thématiques.

8. L'outil WEBAFFIX proposé par [58] qui utilise également le web comme corpus, est un outil automatisé spécialisé dans l'extraction et l'analyse morphologique de lexèmes à partir de corpus web, servant des applications en linguistique et en traitement automatique du langage. Il fonctionne en deux étapes : premièrement il collecte de lexèmes candidats basée sur des suffixes spécifiques et ensuite il effectue un filtrage sophistiqué pour éliminer le bruit. L'outil présente néanmoins des limitations telles que le taux élevé de faux positifs et des défis dans la gestion de lexèmes complexes.

REVSCRAP [115] est conceptualisé comme une application web autonome, pouvant être déployée sur n'importe quel serveur web Java. Avant de détailler notre approche, il est pertinent de définir le problème et de clarifier certains termes et suppositions sémantiques, largement utilisés dans le domaine du TALN, pour poser le cadre théorique de nos recherches :

— Définition 1 (corpus thématique) : Pour notre système, un corpus thématique est un ensemble de données textuelles brutes recueillies à partir de pages web en relation avec une requête de mots-clés spécifiques.

— Définition 2 (mots-clés) : Le mot-clé utilisé pour le moteur de recherche est envisagé comme un 'couple' (*tuple*) composé d'un synonyme de termes : critiques, commentaires, opinions suivi d'un terme se référant à un produit, un article, une personne, un lieu, etc. pour lequel nous souhaitons recueillir des informations en vue d'une analyse ultérieure. La forme du mot-clé reste la même quelle que soit la langue requise.

— Définition 3 (aspect-résultat) : Le résultat présenté dans le document XML final est une instance d'aspect, également envisagée comme un 'couple' (*tuple*) composé de l'URL et du mot-clé utilisé dans la requête.

L'information fournie par le web est complexe à extraire en raison de la complexité croissante des mises en page avec des menus, des formulaires, des *sidebars*, des publicités et tout autre matériel non pertinent présent sur les pages. Le système REVSCRAP est capable de saisir spécifiquement des données textuelles des pages de sites web renvoyées par un index web suite à des requêtes utilisateur. La méthode appliquée est conçue comme une série d'étapes séquentielles, chacune étant accessible à l'utilisateur.

1. La première étape permet à l'utilisateur de définir les mots-clés pour la requête de recherche.
2. La seconde étape collecte les URL, filtrées pour exclure le bruit comme les publicités et les documents formatés (pdf ou ppt).
3. La troisième étape est le cœur du processus : elle emploie un algorithme spécifique de l'arbre DOM, nommé ScrapRev [9], pour détecter et identifier les nœuds de commentaires, indépendamment de la structure hétérogène de la page, tout en évitant les textes redondants et le bruit.

En dépit de l'hétérogénéité et de la complexité des pages web, les données que nous ciblons sont généralement toujours situées dans les mêmes nœuds répétitifs, indépendamment de la langue. Cela nous a permis d'automatiser cette distinction et de fournir un corpus thématique prêt à l'emploi.

9. La technique ScrapRev est basée sur l'algorithme d'analyse de l'arbre de haut en bas, ce qui facilite sa mise en œuvre par programmation manuelle. Les résultats sont collectés dans des documents XML bien structurés portant le nom des termes de la requête. La taille des documents XML varie en fonction de la quantité de commentaires présents dans la page Web récupérée.

Contexte de travaux connexes

Cette section examine les travaux antérieurs afin de situer notre recherche et d'en souligner l'originalité. Résumant l'essentiel, cela nous permet d'identifier les défis rencontrés et les solutions déjà proposées. Des outils comme BootCat [14], WebCorp [77] ou encore le *Linguist's Search Engine* (Linguist's Search Engine (LSE)) [140] ont été consacrés à la construction des corpus et utilisent des requêtes automatisées ainsi que d'autres nouvelles méthodologies. Toutefois, ces outils avaient leurs limitations : ils se basaient principalement sur des corpus majoritairement réalisés à la main, ne fournissaient pas de résultats facilement téléchargeables adaptés à une utilisation ultérieure, ou étaient trop coûteux pour de nombreux budgets académiques.

Les méthodes d'extraction d'information Information Extraction (IE) appliquées aux services d'information web offrent des technologies efficaces pour découvrir des connaissances précieuses et pertinentes dans un ensemble défini de concepts exprimés par un corpus de textes, généralement assemblés manuellement pour former un domaine d'information spécifié [141]. La construction de corpus web spécifiques à l'aide de requêtes automatisées sur des moteurs de recherche a débuté en 2001 avec CorpusBuilder de [52], un outil pour générer automatiquement des requêtes de recherche web afin de construire des corpus pour des langues minoritaires. [15] a proposé une méthode pour construire des corpus web spécifiques au domaine appelés BootCaT, utilisant un petit ensemble caractéristique de graines pour des requêtes automatisées. Les résultats sont ensuite utilisés pour étendre le corpus et ainsi de suite. [82] ont utilisé la même méthode pour construire une usine de corpus des principales langues du monde, en rassemblant une liste de mots graines de cent mots et en sélectionnant aléatoirement trois de ces mots pour créer une requête de recherche. Cette opération est répétée plusieurs milliers de fois pour obtenir un grand corpus. Les pages correspondantes sont ensuite nettoyées, segmentées, lemmatisées, étiquetées grammaticalement avant d'être chargées en tant que corpus résultant. [160] a également utilisé cette approche avec une liste de 500 mots graines.

À la première étape, nous définissons un ensemble de mots-clés : un ensemble de termes sémantiquement proches de l'opinion (dans notre cas), qui constituent l'entrée de la requête du moteur de recherche.

Par exemple, en français, un mot-clé pourrait être *avis utilisateurs + produit*. Pour nos tests, nous avons défini des mots-clés équivalents dans trois langues : anglais, français et grec, et nous avons utilisé le moteur de recherche Google pour sauvegarder toutes les URL correspondantes.

Cette approche lexicale, en utilisant des synonymes dans la requête, évite d'obtenir uniquement des URL avec une description du produit, sans avis d'utilisateurs. Un premier filtrage est effectué une fois que la requête est traitée et que le moteur de recherche renvoie les URL correspondantes. Un filtre est appliqué pour supprimer les publicités ou les pages de formatage (par exemple pdf) afin d'éviter les liens inutiles ou non pertinents. Chaque page retournée est traitée par l'algorithme de parcours d'arbre *top-down RevScrap DOM*.

Par "traitée", nous entendons que la page est divisée en plusieurs blocs de contenu selon l'architecture DOM du W3C, qui identifie la structure logique des documents et la manière dont elle est gérée en termes d'accès et de manipulation. Dans un DOM, une page web peut être analysée et représentée sous forme d'une structure arborescente qui modélise les relations parent-enfant entre les balises HTML, et dans laquelle les nœuds feuilles contiennent du contenu ou des textes d'ancrage. La figure 2.1 montre un exemple de représentation graphique d'un arbre DOM.

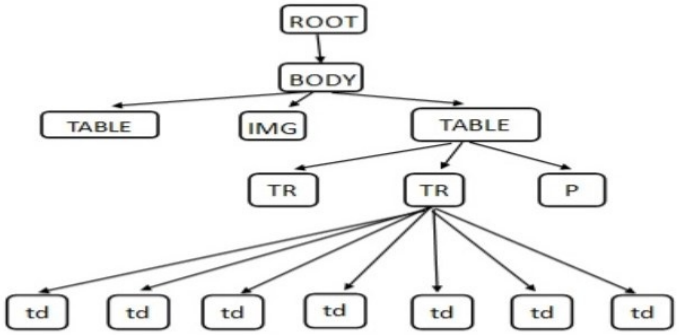


FIGURE 2.1 – Exemple de représentation graphique d'un arbre DOM

Notre approche est basée sur l'hypothèse que, dans une page HTML donnée divisée en blocs, les commentaires sont situés dans les nœuds répétés d'un sous-bloc. Ainsi, nous parcourons la page à la recherche des nœuds répétés. Les nœuds répétés plus de n fois [10], avec un contenu moyen supérieur à 250 caractères, est considéré comme du contenu de commentaire potentiel à extraire. Un nombre de caractères moins important pourrait entraîner l'extraction de contenu non pertinent, tel que des liens, des images, etc. Tout le contenu des nœuds feuilles répétés est marqué comme des commentaires potentiels et sauvegardé pour un traitement ultérieur après la réduction des nœuds appropriés potentiels. Nous présenterons le module qui traite le problème du filtrage des pages web et résume les nœuds récursifs dans la section suivante.

Filtres

Ce module se charge de réduire le bruit des nœuds qui potentiellement contiennent des informations pertinentes pour nos requêtes. Les avis sur le web sont généralement associés à des scores, à des évaluations par étoiles et toujours à des dates. Les premiers tests effectués sur les filtres ont tendance à montrer que le nœud comportant la notion de *date* est le nœud le plus couramment associé aux avis.

10. On cherche le maximum des avis, les tests ont été effectués avec $n > 10$

...

...

Exemple : noeuds comportant la notion *date*

Par conséquent, nous définissons le nœud de date (notion élargie) comme premier filtre associé aux commentaires correspondants récupérés, et l'algorithme remonte jusqu'au troisième ancêtre à la recherche d'un nœud de date.

Un deuxième filtre est associé aux commentaires qui ont une URL qui ne correspond pas entièrement au mot-clé [11] initial de recherche. À cette fin, nous procédons à une analyse *pattern matching* (correspondance de motifs) et à une correspondance de requête pour tous les commentaires récupérés avec une URL non correspondante. Par exemple, pour une requête du type "avis samsung galaxy S4", il est possible de récupérer des commentaires associés à une URL d'un galaxy S5, comme le montre la figure 2.2. Ainsi, ce deuxième filtre de correspondance de motifs est mis en place pour limiter voire éviter ces cas de non-correspondance.

<idCommentaire>9ed38bfe-d858-45fe-ad49-
5d0bf831e4f0</idCommentaire>
<contenu>bien date :31 mars 2014 vous aimez : qualité photo,
qualité écran tres beau téléphone + 15points 15of 15voted this as
helpful. faîtes connaître produit : merci! vous avez réussi à
soumettre un commentaire pour cet avis.</contenu>
<score>0.0</score>
<domaine>avis.orange.fr</domaine>
<url>http://avis.orange.fr/6044-fr_fr/3561292201638/samsung-
samsung-galaxy-s5-noir-reviews/reviews.htm</url>

FIGURE 2.2 – Une non-correspondance d'URL avec un mot-clé

Sur la figure 2.3, nous avons utilisé la même requête de mots-clés en grec, et certaines URL correspondent et d'autres ne correspondent pas. Nous avons appliqué le filtre afin d'éviter l'activation de pages web non pertinentes.

En plus des filtres précédemment mentionnés, le système identifie les URL des

11. Concernant la requête avec le "mot-clé", il faut préciser la création d'un petit lexique de synonymes (par exemple : avis, opinions, notes, etc.) qui à son tour est utilisé dans la requête du moteur de recherche afin de couvrir largement le champ sémantique de la requête

FIGURE 2.3 – Exemple de requête en grec et quelques URL retournées

avis qui comportent plusieurs pages d'avis et les explore pour récupérer tous les commentaires associés à ces pages.

L'algorithme dans son ensemble

Les parties qui composent notre algorithme sont décrites ci-dessous :

```
1  fonction extract_theme ( request ) :
2      𝒮 ← getResults ( request ) ;
3      𝒰 ← ∅ ;
4      for s ∈ 𝒮 do
5          𝒰 ← 𝒰 ∪ crawl ( s ) ;
6      ℒ ← ∅ ;
7      for u ∈ 𝒰 do
8          𝒯 ← getTreeDOM ( u ) ;
9          ℒ ← ℒ ∪ getComments ( 𝒯 ) ;
10     return ℒ ;
```

Alg. 1: Exploration des avis pour l'extraction des thèmes.

```
1 fonction crawl ( url ) :
2     r ← numberOfRotate ( ) ;
3     𝒰 ← ∅ ;
4     for i ∈ [0, r − 1] do
5         abs ← getLinksAbsolutes ( url ) ;
6         rel ← getLinksRelatives ( url ) ;
7         link ← combine ( abs , rel ) ;
8         𝒰 ← 𝒰 ∪ noiseFilter ( link ) ;
9     return 𝒰 ;
```

Alg. 2: Filtrage d'une *url*.

```
1 fonction getTreeDOM ( url ) :
2     ℰ ← getChildren ( url ) ;
3     𝒯 ← ∅ ;
4     level ← getLevel ( ) ;
5     for e ∈ ℰ do
6         𝒯 ← 𝒯 ∪ addDOM ( e , level) ;
7     return 𝒯 ;
```

Alg. 3: Construction du modèle DOM d'une *url*.

```
1  fonction getComments ( dom ) :
2      ℰ ← getChildren ( dom ) ;
3      ℒ ← getLevel ( dom ) ;
4      ℛ ← ∅ ;
5      for i ∈ ℒ do
6          ℰ′ ← getNodesAtLevel ( ℰ , i) ;
7          if size ( ℰ′ ) > 5 then
8              𝒯 ← getTextOfEachNode ( ℰ′ ) ;
9              for t ∈ 𝒯 do
10                 if size ( t ) > 500 then
11                     ℛ ← ℛ ∪ t ;
12     return ℛ ;
```

Alg. 4: Recherche de commentaires dans une *url*.

Résultats

Le résultat obtenu est une collection de fichiers XML avec les balises suivantes : idcommentaire, contenu, score, domaine, path et URL, comme illustré ci-dessous :

Le premier lot de test a été défini pour comparer REVSCRAP à un corpus fait à la main, tous deux d'après la même recherche de requête afin d'évaluer la capacité de REVSCRAP à ne récupérer que les commentaires pertinents. L'idée était de requêter les mêmes avis de produit et de comparer les résultats ; les requêtes sélectionnées sont "avis Samsung S4" et "avis iPhone 6". Pour les besoins des tests, nous avons fixé un seuil de test aux 100 premiers URL renvoyés. Les tableaux suivants montrent les résultats obtenus :

Samsung S4	**URL pertinents**	**URL non pertinents**	**Erreurs**	**Commentaires extraits**	**Commentaires pertinents**
Corpus de référence	57	40	3	885	418
REVSCRAP	72	23	5	736	666

iphone 6	**URL pertinents**	**URL non pertinents**	**Erreurs**	**Commentaires extraits**	**Commentaires pertinents**
Corpus de référence	49	48	3	593	335
REVSCRAP	75	20	5	717	671

TABLE 2.1 – REVSCRAP versus corpus de référence

```
<commentaire>
<idCommentaire>ae8a1dd8-dfc1-4265-a214-7855ad8f08a0</idCommentaire>
<contenu>maximei membre junior inscrit: 5 septembre 2014 messages:
34 j'aime reçus: 1 en ayant l'iphone 6 depuis sa sortie, je peux
t'assurer que l'autonomie a vraiment été améliorée. je tiens
facilement 1 jours et demi entre deux recharges avec plus de 15h en
veille et 8h en utilisation. après ça dépendra aussi de ton
utilisation et de tes réglages (4g activée, actualisation de la
météo en arrière plan et wifi allumé la moitié du temps pour moi).
#2 maximei, 20 octobre 2014</contenu>
<score>0.0</score>
<domaine>forums.macg.co</domaine>
<url>http://forums.macg.co/threads/autonomie-iphone-6.1253720/</url>
<path>div.uix_message </path>
<nbrMotsListeMotsUnGram>0</nbrMotsListeMotsUnGram>
</commentaire>
```


FIGURE 2.4 – Fichier résultat en format xml

Conclusion

L'objectif de ce travail était de créer un système capable de construire des corpus d'avis d'utilisateurs, dans différentes langues, selon la requête sur un produit spécifique de l'utilisateur. Le système offre une interface interactive qui fonctionne séquentiellement sur la base d'une méthode étape par étape. Tout d'abord, nous définissons les termes représentant les mots-clés pour l'analyse d'opinions. Un petit lexique de synonymes est ensuite créé, qui à son tour est utilisé dans la requête du moteur de recherche afin de couvrir largement le champ sémantique de la requête. Le moteur de recherche renvoie de nombreux URL et un premier filtre est appliqué pour éliminer les publicités et les types de pages web comme les pdf. Les liens vers les pages sont activés après avoir analysé et filtré les nœuds html pertinents. Les nœuds des données d'opinion sont détectés et analysés. Les données sont ensuite regroupées dans des documents XML contenant peu ou pas du tout de bruit. Selon ces expériences, nous observons que REVSCRAP a obtenu des résultats encourageants en termes de nombre de commentaires pertinents récupérés à partir des URLs pertinentes. Nous pouvons obtenir une couverture de liens élevée sans utiliser un algorithme chronophage. Le REVSCRAP est un outil simple et efficace pour créer automatiquement des corpus thématiques à l'aide d'une interface web interactive. Il est capable de récupérer et nettoyer les pages d'avis, fournissant ainsi un corpus prêt à l'emploi. Bien que nos résultats actuels indiquent de bonnes performances, plusieurs aspects peuvent être améliorés :

— Le filtrage sur le nœud de la date était le choix le plus judicieux, mais pour une meilleure précision, il serait approprié d'ajouter d'autres filtres de

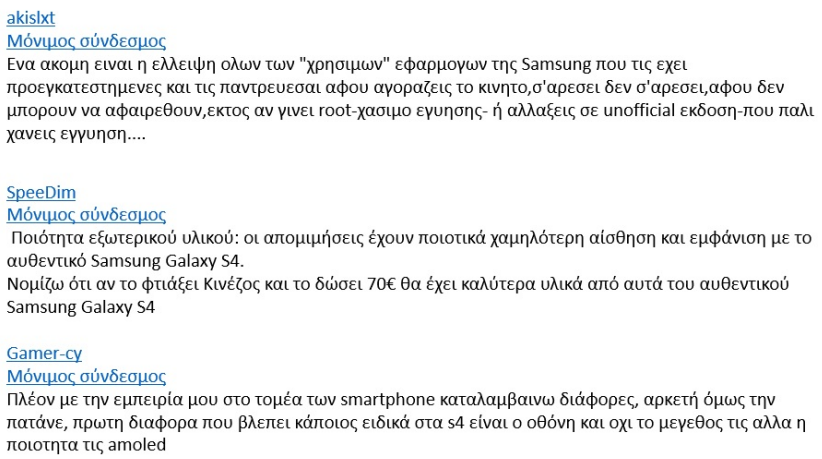


FIGURE 2.5 – Résultats en grec

nœuds tels que le score par exemple [12].

— Tester l'extension du système pour prendre en charge davantage de langues.

Le système a déjà été mis à jour pour prendre en charge l'anglais et le grec. Les travaux futurs incluent la comparaison de l'exactitude et de l'efficacité informatique de la méthode d'extraction et de regroupement des aspects [13] avec d'autres approches. Nous testons une version de traitement parallèle avec plusieurs threads. De plus, une évaluation quantitative à plus grande échelle de nos corpus de données d'opinions sera réalisée.

12. Pour une étude qui nous intéresse sur l'association du score à l'emploie de la négation et de la notion de l'ironie

13. Dans une version améliorée nous prévoyons l'ajout des caractéristiques des *produits* pour lesquels nous récupérons les avis.

2.2.2 DYCORC (*Dynamic Corpus Constructor*)

L'essor des interactions numériques sur les réseaux sociaux et autres plateformes en ligne a engendré un besoin pressant d'extraire et d'analyser les commentaires, avis et interactions des utilisateurs. L'un des défis majeurs est le filtrage du bruit omniprésent tels que les publicités, les duplicatas, et autres contenus non pertinents. Dans la plupart des forums en ligne, comme illustré dans la figure 2.6, le bruit peut parfois surpasser le contenu pertinent, rendant le corpus brut quasiment inutilisable. Les approches manuelles ou semi-supervisées actuelles de nettoyage [45] ont plusieurs inconvénients. Elles exigent souvent que l'utilisateur définisse manuellement des conteneurs ou des enveloppes, parfois *via* des expressions régulières ou des programmes spécifiques. Ces méthodes sont non seulement laborieuses, mais également spécifiques au forum analysé, rendant la standardisation difficile. De plus, le paysage varié des systèmes de gestion des forums entraîne une diversité de structures de page, rendant le nettoyage manuel encore plus ardu [20].

DYCORC est introduit comme une solution pour le problème du bruit. Il s'agit d'un extracteur de contenu non supervisé. Il se distingue par son analyse de la structure DOM des pages, utilisant diverses distances de chaîne pour isoler les éléments pertinents. Suite à une évaluation rigoureuse de sept distances de chaîne, seules les plus performantes ont été retenues.

L'absence notable de corpus multilingues de référence accessibles amplifie ce défi. Pour pallier ce manque, nous avons créé notre propre corpus de référence.

Contexte de travaux connexes

Au cours des dernières années, un certain nombre de systèmes de collecte et d'extraction ont été développés. Un aperçu de ces techniques est détaillé dans [172], qui recommande Focus [70], un crawler supervisé de forums web qui utilise l'alignement partiel des arbres et, selon ses auteurs, surpasse les autres méthodes. À titre d'autre exemple, [112] crée des corpus à grande échelle à partir de Twitter pour l'évaluation de la détection d'événements, tandis que [38] proposent un modèle évolutif pour les messages courts (tweets) dont la technique principale repose sur un agent évaluant l'intérêt d'un message en fonction d'une liste prédéfinie d'interactions, en utilisant des bases de données comme Wikipedia ou DBPedia pour associer le contenu identifié à des données pertinentes. Les méthodes d'extraction de contenu sans bruit à partir de forums peuvent être classées selon :

- les données sur lesquelles elles travaillent : le texte, la structure HTML ou les deux ;
- le nombre de pages en entrée : page par page ou ensemble de pages ;
- les connaissances utilisées : sur les balises HTML (par exemple, les balises

FIGURE 2.6 – Contenu pertinent et *bruit* d'un forum

les plus souvent utilisées pour le contenu standardisé, densité des balises, structure des URL), sur le texte (densité des mots, mots vides, syntaxe superficielle) ;

— si elles examinent l'architecture de la page ou l'arbre DOM pour détecter l'emplacement le plus probable du contenu pertinent.

L'un des premiers *crawlers* de forums web non supervisés et sans bruit est Roadrunner [32]. Il travaille sur l'arborescence HTML (équivalent de l'arbre DOM), prend un ensemble de pages en entrée et compare leurs arbres à l'aide d'un algorithme de correspondance d'arbres simple. Le contenu pertinent est déterminé en comparant les codes HTML de plusieurs pages et en déduisant l'enveloppe. Les systèmes non supervisés les plus connus, facilement accessibles aux chercheurs car ils sont open source et maintenus par de grandes équipes, sont Apache Nutch [80] et BootCat [14] .

Apache Nutch, implémenté en Java, est divisé en quatre composants : le *crawler*, l'indexeur, le stockage de la base de données, le récupérateur. Le serveur SolR est utilisé pour lire les données indexées. Pour le contenu standardisé, il utilise la bibliothèque Boilerpipe [85], conçue pour extraire le contenu des sites web d'ac-

tualités. Cette bibliothèque fonctionne à la fois sur le texte et les balises HTML, page par page, en utilisant des connaissances telles que : les lettres majuscules, la longueur des mots, le début et la longueur des phrases (ce qui implique une connaissance de la structure des phrases), les mots-clés (pour le contenu standardisé), une liste de balises de contenu standardisé probables, le nombre de liens dans les balises. La version incluse dans Nutch n'utilise que les densités des mots et des balises trouvées dans les branches de l'arborescence HTML. Le contenu pertinent est là où la densité des mots est plus élevée et la densité des balises est plus faible.

BootCat est basé sur les résultats des moteurs de recherche et les techniques d'expressions régulières ; un inconvénient majeur est que son *crawler* ne vérifie pas si une URL a déjà été parcourue. Pour le contenu standardisé, il travaille à la fois sur le texte et la structure HTML, page par page, en utilisant des connaissances linguistiques (une liste de mots vides) et un modèle de langage n-gramme. Il compte le nombre de balises dans les nœuds et le compare au nombre de n-grammes. Suivant quelque peu le même principe que Nutch, le contenu pertinent est là où le nombre de n-grammes est plus élevé et celui des balises est plus faible. Il est une boîte à outils open-source qui facilite la construction de corpus thématiques à partir des résultats des moteurs de recherche et d'une liste d'entrée de termes de base décrivant les besoins des utilisateurs. Cette étape de configuration n'est pas triviale et prend du temps, car l'utilisateur doit respecter un ensemble de règles prédéfinies.

Parmi les modèles récents utilisant des techniques non supervisées pour l'extraction de contenu de forums, on trouve le système REVSCRAP [115] qui travaille sur l'arbre DOM, page par page, en utilisant des connaissances sur le format de la date dans les expressions régulières. Il recherche la branche contenant la date. Cette heuristique très simple donne des résultats intéressants, mais aucune information sur le bruit n'est disponible, et cette méthode ne fonctionne pas pour les formats de date en langue arabe ou chinoise. Plus en rapport avec notre objectif et facilement accessible [14], JusText [142] est une bibliothèque Python qui fonctionne à la fois sur du contenu textuel et HTML. JusText travaille page par page, en utilisant des connaissances sur la langue utilisée (une liste de mots grammaticaux) et sur les balises HTML. Il compte le nombre de mots et le nombre de mots grammaticaux dans les feuilles, et utilise une liste de balises probables de contenu standardisé.

Description du système

DYCORC parcourt les forums en ligne et extrait le code source des pages web correspondantes. Il n'a pas de connaissance préalable de la langue ni des balises

14. `http://corpus.tools/wiki/Justext`

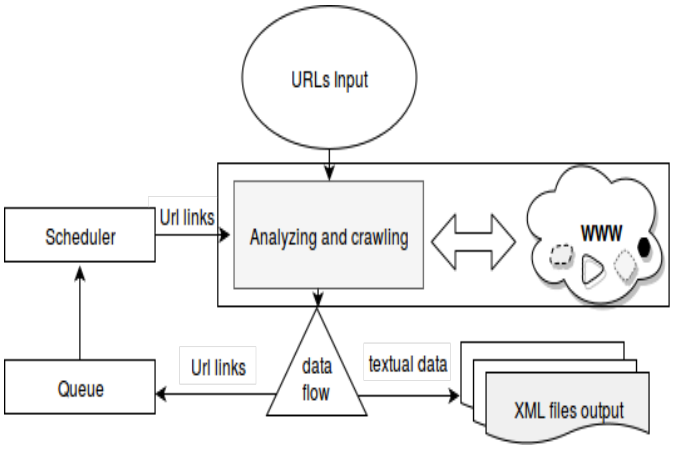


FIGURE 2.7 – Aperçu du *crawler*

utilisées. L'analyse est effectuée page par page. Nous travaillons à la fois sur le texte et la structure HTML, et nous utilisons l'arbre DOM en comparant ses branches afin de localiser l'emplacement le plus probable du contenu pertinent.

La structure HTML, une fois nettoyée des erreurs et incohérences à l'aide de Tidy Html 3 du W3C [15], est convertie en un arbre DOM [197] et enregistrée au format XML. L'extraction du contenu pertinent peut être effectuée immédiatement sur l'arbre DOM ou ultérieurement en lisant le fichier XML, qui est lui-même enregistré dans un fichier XML. DYCORC est écrit en C++ à l'aide de la bibliothèque Qt5.

La Figure 2.7 présente un aperçu des étapes de parcours de DYCORC.

Le schéma de la figure 2.8 représente les étapes d'analyse dynamique suivies pour extraire le contenu pertinent avec DYCORC.

L'étape de parcours ne présente pas de caractéristique particulière, nous utilisons le multithreading et détectons les pages en double à l'aide d'une somme de contrôle. Nous collectons également des données à partir d'une liste de liens (URL), web ou locaux.

Notre principale contribution réside dans l'algorithme d'extraction de contenu, détaillé ci-dessous :

Nous utilisons l'arbre DOM en tant qu'entrée, que nous divisons en régions de données, à savoir le contenu pertinent et le bruit. Le principe général de notre

15. `http://api.html-tidy.org/tidy/tidylib_api_next/language__fr_8h_source.html`

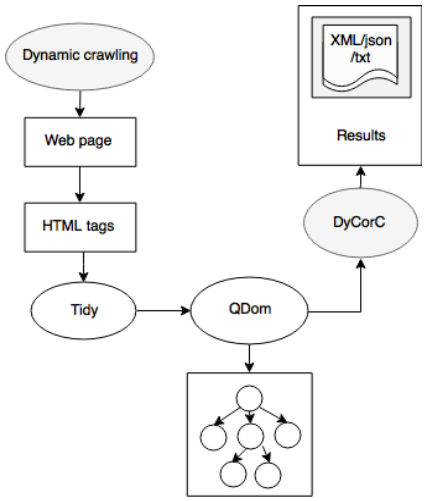


FIGURE 2.8 – Analyse dynamique

méthode est que les structures les plus récurrentes contenant des textes les plus divers sont plus susceptibles de contenir du contenu pertinent.

Dans une première étape, nous détectons les nœuds similaires dans l'arbre DOM et les regroupons par catégories, sans tenir compte du texte ni des attributs.

Chaque nœud est représenté par son chemin depuis la racine, par exemple $\{< body >< h1 >< p >< span >\}$, converti en une chaîne de caractères, comme "bodyh1pspan". En réalité, nous omettons la balise *body*, car elle fait partie de tous les chemins.

La similarité est détectée de la manière suivante :

Pour chaque paire de nœuds d'un niveau donné dans l'arbre DOM, nous comparons leurs chaînes s_i et s_j à l'aide d'une distance de chaîne. Si cette distance est inférieure à la longueur moyenne de leurs chaînes (inégalité 2.1), ils sont dans la même catégorie. Nous répétons cette opération pour chaque niveau. Les catégories résultantes contiennent des blocs similaires, comme illustré dans la figure 2.9.

$$Dist(s_i, s_j) < \frac{length(s_i) + length(s_j)}{2} \tag{2.1}$$

Ensuite, nous calculons la moyenne de la distance textuelle par catégorie, ce qui nous donne une mesure de la diversité du texte dans les blocs récurrents : dans chaque catégorie c_k, nous calculons la distance entre chaque paire de contenus textuels des branches, nous additionnons toutes les distances, puis nous divisons

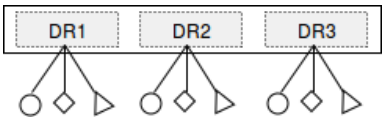


FIGURE 2.9 – Registre de données de la même catégorie

par le nombre de combinaisons par paires 2.2.

$$MeanDist_T(c_k) = \frac{\sum_i \sum_{j>i}^{\forall b_i, b_j \in c_k} Dist_T(b_i, b_j)}{C_2^{|c_k|}} \tag{2.2}$$

Dans cette équation, $MeanDist_T(c_k)$ représente la moyenne des distances textuelles dans la catégorie c_k, b_i et b_j sont les branches de la catégorie c_k, et $C_2^{|c_k|}$ est le nombre de combinaisons de 2 parmi $|c_k|$. Voici la traduction de la description de l'équation :

La première étape crée souvent des singletons, que nous essayons d'incorporer aux catégories existantes en utilisant le processus suivant.

Pour chaque catégorie c_k, nous sélectionnons le nœud n_i dont la distance textuelle moyenne par rapport aux autres est la plus proche de la distance textuelle moyenne de c_k (équation 2.3). Ce nœud sera la représentation de c_k.

$$repr(c_k) = \min_{i=0}^{|c_k|-1} \left\{ MeanDist_T(c_k) - \frac{\sum_{j\neq i}^{j<|c_k|} Dist_T(n_i, n_j)}{|c_k| - 1} \right\} \tag{2.3}$$

Dans cette équation, $repr(c_k)$ représente la représentation de la catégorie c_k, $|c_k|$ est le nombre de branches dans la catégorie c_k, et la formule calcule la différence entre la distance textuelle moyenne de la catégorie c_k et la moyenne des distances textuelles par rapport à tous les autres nœuds dans la catégorie c_k.

Ensuite, nous prenons en compte les attributs Att_S du singleton S et de toutes les représentations, et calculons pour chacun l'indice de Jaccard des attributs. Le singleton sera placé dans la catégorie ayant l'indice de Jaccard le plus élevé, s'il est supérieur au seuil fixé par l'utilisateur θ (inégalité 2.4), où N est le nombre total de catégories.

$$\left(\max_{k=0}^{N-1} \frac{|Att_S \cap Att_{repr(c_k)}|}{|Att_S \cup Att_{repr(c_k)}|} \right) > \theta \tag{2.4}$$

Dans cette équation, Att_S représente les attributs du singleton S, $Att_{repr(c_k)}$ représente les attributs de la représentation de la catégorie c_k, et la formule calcule l'indice de Jaccard entre ces ensembles d'attributs. Si l'indice de Jaccard est supérieur au seuil θ, le singleton est assigné à la catégorie correspondante. Après l'intégration d'un singleton, nous recalculons la distance textuelle moyenne de la catégorie. Ce processus est appliqué de manière récursive jusqu'à ce qu'aucune intégration supplémentaire ne puisse avoir lieu.

Enfin, nous déterminons quelle catégorie contient le contenu textuel le plus riche, c'est-à-dire celle avec la plus grande distance textuelle moyenne (équation 2.5). Les feuilles de cette catégorie représentent les contenus des messages.

$$\max_{k=0}^{N-1} MeanDist_T(c_k) \tag{2.5}$$

Dans une étape de post-traitement, nous extrayons les dates et les identités des auteurs à partir du contenu des messages en utilisant des expressions régulières.

Les distances et similarités de chaînes que nous avons implémentées sont les suivantes :

- Levenshtein [94], ou distance d'édition, basée sur le nombre d'opérations d'insertion, de suppression et de substitution de caractères nécessaires pour convertir une chaîne en une autre ;
- Damerau-Levenshtein [16], qui ajoute les transpositions à Levenshtein ;
- Jaro [69], basée sur le nombre de caractères identiques, ceux dont les index sont plus proches de la moitié de la taille de la chaîne la plus longue, et sur le nombre de transpositions ;
- Jaro-Winkler [17], qui accorde plus de poids au début des chaînes qu'à la fin ;
- *distance de caractéristiques*, basée sur les n-grammes communs ;
- Plus longue sous-séquence commune ou LCS, dont le nom est transparent ;
- Distance de Jaccard [18], basée sur le rapport entre les cardinaux de l'intersection et de l'union de deux chaînes.

Expérimentations

Nous avons réalisé deux séries d'évaluations basées sur deux critères. Les critères sont les suivants : (a) temps de calcul et (b) qualité de l'extraction (ou nettoyage du bruit). La qualité de l'extraction est mesurée par la précision, le rappel et

16. https://fr.wikipedia.org/wiki/Distance_de_Damerau-Levenshtein
17. https://fr.wikipedia.org/wiki/Distance_de_Jaro-Winkler
18. https://fr.wikipedia.org/wiki/Indice_et_distance_de_Jaccard

la F-score. En prenant W_f comme le nombre de mots trouvés et W_g comme le nombre de mots de référence (vérité terrain), la *précision* est définie par $\frac{W_f \cap W_g}{W_f}$, le *rappel* est défini par $\frac{W_f \cap W_g}{W_g}$ et la *F-score* est la moyenne harmonique des deux. La première série d'évaluations a été réalisée entre les différentes distances de chaînes, afin de choisir la meilleure en fonction des deux critères. La deuxième série d'évaluations a comparé notre système, DYCORC, aux systèmes de pointe tels que Apache Nutch, BootCat et JusText. Pour ces tests, nous avons élaboré un corpus de référence spécifique pour les raisons exposées dans la section 2.2.2. Ce corpus sera présenté dans la sous-section suivante, où nous aborderons également les expériences relatives aux distances, au temps de calcul et à la qualité de l'extraction. Enfin, nous effectuerons une comparaison des systèmes mentionnés ci-dessus.

Corpus de Référence

Le corpus de référence utilisé dans cet article est décrit dans le tableau 2.2. Nous avons choisi d'avoir approximativement le même nombre de mots dans les quatre langues (entre 716 000 et 779 000).

TABLE 2.2 – Corpus de référence

	domaine	**pages**	**mots**	**mots pertinents**
Français	developpez.com	33	226 327	33 489
	ubuntu-fr.org	47	300 364	201 126
	etudes-litteraires.com	113	226 595	79 184
	Total Français	**193**	**753 286**	**313 799**
Grec	ubuntu-gr.org	85	366 284	122 843
	dotnetzone.gr	57	195 153	41 805
	fe-mail.gr	5	203 136	139 252
	Total Grec	**147**	**764 573**	**303 900**
Anglais	englishforums.com	366	716 276	186 886
Arabe	forum.ency-education.com	238	715 572	71 192
Total		**944**	**2 949 707**	**875 777**

Les expériences présentées ici sont exécutées avec 8 *threads* d'exécution sur un PC. Le débit moyen de réception internet lors des expériences était d'environ 3 Mbps.

Comparaison des distances de chaînes

Le tableau 2.3 donne les valeurs moyennes de nos critères pour chaque distance. Le temps, exprimé en secondes par page, est le temps moyen d'extraction (sans le crawling) réalisé avec 8 threads, calculé uniquement pour les pages où un contenu pertinent a été trouvé. Comparaison des valeurs moyennes pour les distances :

TABLE 2.3 – Comparaison valeurs moyennes de distances

distance	recall	precision	F-measure	time
Distance de caractéristiques	91.76%	93.75%	**91.25%**	0.026
Levenshtein	91.76%	93.70%	91.22%	0.742
Damerau-Levenshtein	91.76%	93.70%	91.21%	1.1
Jaro	56.38%	57.29%	56.54%	0.043
Jaro-Winkler	56.38%	57.29%	56.54%	0.037
LCS	56.38%	56.90%	56.63%	0.03
Jaccard	49.46%	56.22%	48.51%	0.06

Derrière ces valeurs moyennes, cependant, différentes situations ont été observées. Par exemple, sur certains forums, certaines distances produisent des résultats rapidement mais avec des erreurs. En général, l'architecture des pages joue un rôle important dans le calcul et ses résultats. Ici, nous examinerons trois cas types. Le premier cas (le plus fréquent) est illustré dans le tableau 2.4 : la meilleure distance est la *distance de caractéristiques*, suivie de près par Levenshtein et Damerau-Levenshtein. LCS est plus rapide, mais aucun contenu pertinent n'a été trouvé.

TABLE 2.4 – Temps moyen : meilleur *distance de caractéristiques* suivi de Levenshtein

Domaine	Distance	F-measure	Temps (s/p)
developpez.com	Levenshtein	95.94%	0.82
	Damerau-Levenshtein	95.94%	1.13
	Feature-distance	**96.17%**	**0.18**
	Jaro	-	1.5
	Jaro Winkler	-	1.25
	Jaccard	-	1.42
	LCS	-	0.003

Dans l'ensemble, Levenshtein et Damerau-Levenshtein donnent des résultats presque aussi bons que la *distance de caractéristiques*, mais prennent plus de temps. Jaro-Winkler donne les mêmes résultats que Jaro et prend plus ou moins le même temps. Jaccard n'est jamais le meilleur, et LCS est inférieur ou égal à Jaro en termes de qualité et était meilleur en termes de temps de calcul.

TABLE 2.5 – Temps moyen de toutes les distances pour un site

Domaine	**Distance**	**F-score**	**Temps (s/p)**
etudes-litteraires.com	Levenshtein	94.20%	0.011
	Damerau-Levenshtein	94.20%	0.014
	distance de caractéristiques	94.20%	0.01
	Jaro	94.20%	**0.002**
	Jaro Winkler	94.20%	0.002
	Jaccard	94.20%	0.02
	LCS	94.20%	0.01

Les distances les plus intéressantes, selon les tests, sont les suivantes :

— *distance de caractéristiques*, car elle est la meilleure en termes de qualité et elle présente le temps de calcul moyen le plus court, et

— *la distance de Jaro*, car dans certains cas elle divise par 5 le temps d'extraction sans aucune perte d'information, et présente même de meilleurs résultats en termes de qualité.

Leurs résultats détaillés sont présentés dans les tableaux 2.6 et 2.7.

TABLE 2.6 – Résultats pour les différents sites par *Distance de caractéristiques*

Domaine	**lang.**	**Rappel**	**Précision**	**F-measure**	**Temps**
developpez.com	french	100%	92,61%	96,17%	0.18
ubuntu-fr.org	french	100%	92,28%	95,99%	0.008
etudes-litteraires.com	french	100%	89,04%	94,20%	0.01
ubuntu-gr.org	greek	95,03%	100%	97,45%	0.004
dotnetzone.gr	greek	100%	95,22%	97,55%	0.02
fe-mail.gr	greek	100%	95,40%	**97,65%**	**0.0002**
englishforums.com	english	43,68%	100%	60,80%	0.03
forum.ency-education.com	arabic	95,44%	85,50%	90,20%	0.02

Le seul forum où la *distance de caractéristiques* présente un mauvais rappel est

englishforums.com, et c'est le seul où la distance Jaro est la meilleure en termes de qualité. Aucune connaissance sur la langue n'est intégrée dans notre algorithme, la seule influence possible est l'architecture du forum.

TABLE 2.7 – Distance de Jaro résultats

Domaine	**lang.**	**Rappel**	**Précision**	**F-measure**	Temps
developpez.com	french	0%	0%	-	1.5
ubuntu-fr.org	french	0%	0%	-	0.002
etudes-litteraires.com	french	100%	89,04%	94,20%	0.002
ubuntu-gr.org	greek	95,03%	100%	97,45%	0.012
dotnetzone.gr	greek	100%	95,22%	97,55%	0.0025
fe-mail.gr	greek	100%	95,40%	**97,65%**	**0.0004**
englishforums.com	english	56,07%	78,72%	65,49%	0.07
forum.ency-education.com	arabic	0%	0%	-	0.07

Nous constatons que la distance de Jaro fonctionne soit aussi bien (voire mieux) que les autres distances, soit pas du tout. Le facteur semble être la longueur des chemins dans l'arbre DOM. Il en va de même pour les distances de Jaro Winkler et de LCS, et partiellement pour la distance de Jaccard. Sur les 944 pages de notre corpus de référence, correctement analysées par la *distance de caractéristiques*, Levenshtein et Damerau-Levenshtein, Jaro et Jaro Winkler n'ont traité que 626 pages, LCS 668 pages et Jaccard 260 pages.

Notre conclusion jusqu'à présent a été d'intégrer la *distance de caractéristiques* dans notre algorithme, car elle présente une qualité stable et le temps moyen de calcul le plus court.

Évaluation comparative

Nous avons sélectionné Apache Nutch [80] (version 1.12) avec le serveur SolR en version 4.10), BootCat [14] (version stable 2014) et JusText comme modèles de référence. Nous comparons leurs valeurs moyennes [19] (non pondérées) à celles de DYCORC (avec *Distance de caractéristiques*) dans le tableau 2.8.

Le tableau 2.9 présente les performances globales pour le traitement de l'ensemble du corpus de référence pour chaque modèle. Le temps de traitement (exprimé en

19. Puisque JusText ne parcourt pas les données et BootCat et Nutch ne séparent pas le temps de parcours du temps d'extraction, nous avons adapté JusText à notre *crawler* (sur 8 *threads* (fils) d'exécution) afin de comparer des processus similaires.

TABLE 2.8 – Valeurs moyennes des modèles

Modèle	**Précision**	**Rappel**	**F-score**	**Temps (s/p)**
Nutch	61,53%	75,03%	52,5%	1.61
BootCat	36,73%	96,91%	49,19%	0.85
JusText	43,76%	**100%**	58.29%	1.26
DYCORC	**93,75%**	91,76%	**91,25%**	**0.77**

secondes) et le nombre de mots récupérés [20] sont indiqués.

TABLE 2.9 – Résultats : temps et nombre de mots

	Nutch	**BootCat**	**JusText**	**DYCORC**
Temps (secs.)	1581.9	**390.44**	911.07	468.15
Mots récupérés	480 257	266 760	361 163	**800 932**

Ces résultats mettent en évidence les performances de chaque modèle ; ainsi, globalement, DYCORC est le modèle qui offre la meilleure qualité d'extraction ; il retrouve plus de 800 000 mots réels dans le corpus de référence (sur un total de 875 777).

En ce qui concerne le temps, DYCORC affiche le meilleur temps moyen, mais comme les valeurs moyennes ne sont pas pondérées par le nombre de pages de chaque domaine, cela signifie simplement que DYCORC est moins influencé par la structure du forum, tandis que BootCat (tableau 2.11) a été entravé par *fe-mail.gr*, un domaine très petit.

En examinant de plus près les résultats présentés dans les tableaux 2.10, 2.11, 2.12, nous observons également que c'est le forum en arabe [21] qui cause de 'mauvais' résultats à Nutch, qui est sinonle plus rapide (suivi de BootCat). Il est difficile de dire si cela est dû à la structure, à la langue ou aux deux.

Conclusion

Nous avons présenté DYCORC, un système non supervisé pour extraire un contenu sans bruit des forums web, avec un algorithme qui n'utilise aucune connaissance sur la langue ou les balises. Il surpasse les trois modèles de pointe que nous avons

20. On calcule le texte extrait par nombre de mots.
21. `https://ency-group2.ahlamontada.com/`

TABLE 2.10 – Nutch

Domaine	**Lang.**	**Précision**	**Rappel**	**F-score**	**Temps**
developpez.com	fr	49,53%	100%	66,25%	1.23
ubuntu-fr.org	fr	20,63%	89,79%	33,56%	0.33
etudes-litteraires.com	fr	95,75%	84,31%	89,67%	0.1
ubuntu-gr.org	gr	85,81 %	62,18%	72,11%	0.39
dotnetzone.gr	gr	72,98%	90,72%	80,89%	0.37
fe-mail.gr	gr	1,80%	96,28%	3,54%	3.2
englishforums.com	en	71,31%	76,99%	74,04%	0.1
forum.ency-education.com	ar	94,46%	46,41%	62,24%	5.57
Valeurs moyennes		61,53%	75,03%	52,50%	1.61

TABLE 2.11 – BootCat

Domaine	**Lang.**	**Précision**	**Rappel**	**F-score**	**Temps**
developpez.com	fr	36,15%	100%	53,10%	1.30
ubuntu-fr.org	fr	6,25%	99,75%	11,76%	0.6
etudes-litteraires.com	fr	81,85%	98,83%	89,54%	0.23
ubuntu-gr.org	gr	24,15%	98,02%	38,75%	1
dotnetzone.gr	gr	40,44%	96,61%	57,01%	0.55
fe-mail.gr	gr	16,13%	100%	27,77%	2.6
englishforums.com	en	23,99%	83,24%	37,25%	0.24
forum.ency-education.com	ar	64,91%	98,86%	78,37%	0.32
Valeurs moyennes		36,73%	96,91%	49,19%	0.85

testés sur notre corpus de référence dans quatre langues, en termes de précision et de F-mesure, bien qu'il soit surpassé par BootCat en temps de calcul, prenant 20 % de temps en plus, tout en restant dans une plage raisonnable.

L'intégration de DYCORC avec BootCat pourrait combiner le talent de DYCORC en tant que détecteur de modèles avec les performances de BootCat en tant que *crawler*. DYCORC pourrait être davantage amélioré afin de détecter la meilleure distance à utiliser, alternant entre la *distance des caractéristiques* et la distance Jaro, en faisant une estimation de la longueur des arbres dans un forum.

Le problème des forums arabes qui ralentissent Nutch devrait être soigneusement

TABLE 2.12 – JusText

Domaine	**Lang.**	**Précision**	**Rappel**	**F-score**	**Temps**
developpez.com	fr	34%	100%	50,75%	0.15
ubuntu-fr.org	fr	11,53%	100%	20,68%	0.1
etudes-litteraires.com	fr	82,23%	100%	90,25%	0.1
ubuntu-gr.org	gr	31,21%	100%	47,57%	0.15
dotnetzone.gr	gr	36,69%	100%	53,68%	0.11
fe-mail.gr	gr	55,11%	100%	71,06%	2.1
englishforums.com	en	44,41%	100%	61,50%	0.4
forum.ency-education.com	ar	54,91%	100%	70,89%	1.11
Valeurs moyennes		43,76%	100%	58,29%	0.52

évalué, car il reste l'un des *crawlers* automatiques les plus intéressants (et, sauf pour ce forum, le plus rapide). Si cela est dû à BoilerPipe, la possibilité de le remplacer par une bibliothèque extraite de DYCORC pourrait conduire à des résultats plus intéressants en l'intégrant avec BootCat.

2.2.3 WEBT-IDC (*WebTool for Intelligent Dataset Creation*)

WEBT-IDC [22] est un outil de raclage des pages web [23] pour les forums et blogs, et permet la création de jeux de données (*datasets*) [24]. Il construit des corpus de retours d'opinions sur différents produits quelle que soit la langue, et sans bruit. La méthode repose sur un modèle d'extraction unique basé sur le schéma d'élément de pagination, indépendant de la structure DOM.

WEBT-IDC couvre toutes les étapes, depuis la requête de l'utilisateur, l'analyse et le raclage des pages web, jusqu'à l'extraction des données et la construction de corpus. Sa pertinence et sa fiabilité ont été démontrées en intégrant le dataset créé lors des tests, dans un modèle d'architecture Transformer[179], soulignant ainsi son potentiel pour une utilisation immédiate dans des tâches de ML.

Malgré le grand nombre de jeux de données en ligne, peu de langues disposent de corpus constitués à partir de ressources telles que les forums et les blogs. Encore moins nombreux sont les corpus constitués uniquement de critiques et d'opinions rédigées dans n'importe quelle langue sur un produit ou un service spécifique disponible n'importe où dans le monde.

La possibilité de traiter ce type de contexte de manière précise offre une vision réaliste du marché, notamment en ce qui concerne les produits et les services. Cependant, face à la rareté des données sur des sujets particuliers et les défis associés au pré-traitement de données (filtrage des contenus standardisés et répétitifs, analyse, classification, étiquetage), la création et la maintenance des jeux de données restent souvent un travail manuel et laborieux.

Le système a été évalué non seulement sur sa capacité à filtrer le bruit et sa rapidité de traitement, mais aussi sur des critères tels que la précision et le rappel.

Travaux liés à l'extraction de contenu web

L'extraction du contenu du web (Web Content Extraction), faisant partie du *Web Content Mining*, est étroitement liée au *web mining* (fouille du web) et peut souvent se chevaucher avec ce dernier dans ses applications. Selon [86], le web mining se divise en trois sous-catégories : l'extraction de contenu web [72], le *mining* de

22. Outil développé pour les besoins de travaux de thèse de Maroua Boudabous.
23. Extraction de données web.
24. Le terme "dataset" qui est utilisé souvent de manière interchangeable avec "corpus" dans le domaine de TALN, est un terme plus général et s'utilise dans de nombreux domaines, y compris mais sans s'y limiter à la science des données, la statistique, et le *machine learning*. Un dataset est un ensemble structuré de données qui peut contenir des textes, des nombres, des images, etc. Les datasets sont souvent utilisés pour l'entraînement et l'évaluation des modèles algorithmiques, y compris ceux utilisés pour l'analyse de texte [73].

structure web, et le *mining* d'utilisation web [192]. Le *mining* d'utilisation web, ou Web Log Mining, détecte les comportements des utilisateurs en se concentrant sur les fichiers journaux des serveurs web, tandis que le *mining* de structure web analyse la structure du réseau web. L'extraction de contenu web se concentre sur le contenu des pages web et facilite l'extraction de données utiles dans un format prêt à l'emploi.

La création de corpus web se situe dans la catégorie de l'extraction de contenu web et repose sur la technique du "raclage du web" [108]. Plusieurs cadres de pointe sont utilisés pour cette tâche, y compris Scrapy [123], un outil open-source écrit en Python.

Cependant, les outils existants ne garantissent pas toujours une extraction sans bruit ni un processus d'extraction entièrement automatique. Plusieurs outils ont été développés pour faciliter la collecte de données, tels que Screen-scraper [25], Mozenda [26], Web Info Extractor, et Web Content Extractor [27].

L'élimination du bruit est une étape essentielle dans l'implémentation d'un outil Web Content Mining (WCM). Des approches basées sur différentes mesures de distance de chaînes et des techniques comme les champs conditionnels aléatoires Conditional Random Field (CRF) et les réseaux neuronaux artificiels Artificial Neural Network (ANN) ont été utilisées pour classifier les parties d'une page web en sections bruyantes et pertinentes [109], [168], [65].

Contrairement à WEBT-IDC, ces outils ne garantissent pas un processus complet et reposent sur une extraction semi-automatique nécessitant souvent l'intervention humaine. Ils peuvent devenir chronophages lors de la création de grands ensembles de données.

Dans la section suivante (Déscription du système), nous décrivons les deux principaux composants mis en œuvre par WEBT-IDC : le composant de filtrage et d'exploration web et le composant de création de corpus et de "raclage du web".

Description du système

Pour répondre à l'urgence d'avoir de grands corpus multilingues prêts à l'emploi pour les tâches d'apprentissage automatique, nous avons développé un outil web capable de construire un corpus textuel sans bruit composé d'avis et d'opinions sur les produits et services postés sur des forums et des blogs.

25. `https://www.screenscraper.fr/`
26. `https://www.mozenda.com/master-the-matrix/`
27. Plusieurs outils de ces types sont proposés sur le web.

Le WEBT-IDC commence par la requête de l'utilisateur, sans aucune connaissance de l'architecture de la page web ou de la langue utilisée pour exprimer les avis. Il est composé de deux modules assurant respectivement l'exploration web avec le filtrage et le "raclage" avec la composante de création de corpus. Le premier recherche sur le web mondial les pages qui correspondent le mieux à la requête de l'utilisateur *via* les moteurs de recherche et filtre les contenus redondants et répétitifs. Le second extrait les textes d'avis des utilisateurs ainsi que des informations complémentaires telles que la date de publication, la localisation, le taux de classement et les sauvegarde dans des fichiers de sortie semi-structurés prêts à être utilisés pour les tâches d'apprentissage automatique. La figure 6 détaille l'algorithme de WEBT-IDC.

Le composant d'exploration et de filtrage

Comme illustré par la figure 5, le composant d'exploration et de filtrage prend en entrée la requête utilisateur q et un sujet t. Il commence par parcourir le web pour collecter des pages pertinentes suivant la requête de l'utilisateur grâce à la méthode *crawl_for_query()* qui utilise l'outil GoogleScraper [175], un outil open-source extrayant tous les liens trouvés. Ensuite, il valide la liste des URL collectées en utilisant un composant de filtrage qui supprime les liens HTTP en double (*validate_remove_duplicates()*), évitant ainsi d'extraire les mêmes données plusieurs fois dans le jeu de données final. À la fin de cette étape, nous obtenons une liste propre d'URL initiales prêtes à être traitées par le composant d'extraction et de création de corpus décrit dans la section suivante.

```
1 fonction crawling ( q , t ):
2     candidate_url ← craw_for_query ( q , t );
3     L ← validate_remove_duplicate ( candidate_url );
4     for e ∈ L do
5         scrapURL ( e , t );
```

Alg. 5: `WebT-IDC` *crawling*.

Le composant d'extraction et de création de corpus

Afin de fournir un traitement d'extraction robuste et adaptable (pour différents types de pages web), nous avons d'abord effectué une analyse des structures DOM des forums. Nous avons obtenu des informations approfondies sur la manière dont les données sont organisées. L'élément commun qui permet une robustesse est la présence de l'*élément de pagination*. Un élément de pagination est un contrôle

d'interface qui permet aux utilisateurs de naviguer à travers différentes pages d'un ensemble de contenus. C'est souvent utilisé sur les sites web qui ont de grandes quantités de contenu, comme par exemple les produits dans une boutique en ligne, ou les commentaires et avis des utilisateurs. Si nous consultons une liste d'avis utilisateurs sur un site de commerce électronique, et qu'il y a des centaines d'avis pour un produit, ces avis pourraient être divisés sur plusieurs pages web. Au bas de la page, il pourrait y avoir une barre de pagination qui permet aux utilisateurs de passer à la page suivante, à la précédente, de sauter à la première ou à la dernière page, ou de sélectionner un numéro de page spécifique.

Ensuite, nous avons implémenté le composant d'extraction des données, en commençant par cibler les régions de données pertinentes au sein de la page web jusqu'à la génération du jeu de données. La section suivante détaille ce processus.

```
1 fonction scraping( t , f ):
2     xq ← make_xpath_pagination_bar();
3     𝒰 ← direct_access_pagination_urls( t );
4     for u ∈ 𝒰 do
5         r ← target_relevant( xq );
6         r′ ← extra_info_tags_recognition( r );
7         xq′ ← prepare_xpath( r′ );
8         extract_data_into_file( xq′ , f);
```

Alg. 6: `WebT-IDC` *scraping*.

Analyse de l'architecture des pages web

Cette étape d'inspection vise à déterminer comment les pages web avec des avis et des commentaires sont présentées et donc à obtenir des informations générales afin de créer un composant d'extraction des données adaptable. Par la suite, nous pouvons séparer structurellement le contenu d'un site web donné en sections de données pertinentes et bruyantes. Nous classons comme "contenu bruyant" les publicités et les balises de formatage HTML. Cependant, nous ne classons comme section pertinente que les parties de la page qui correspondent aux attentes et besoins de l'utilisateur. La section pertinente dans les forums et les blogs est celle qui rassemble les avis et les réponses des utilisateurs. La figure 2.10 délimite la section de page pertinente que nous souhaitons exporter à un exemple de site web pour les produits de beauté.

L'étape d'inspection est principalement basée sur l'arbre DOM (Document Ob-

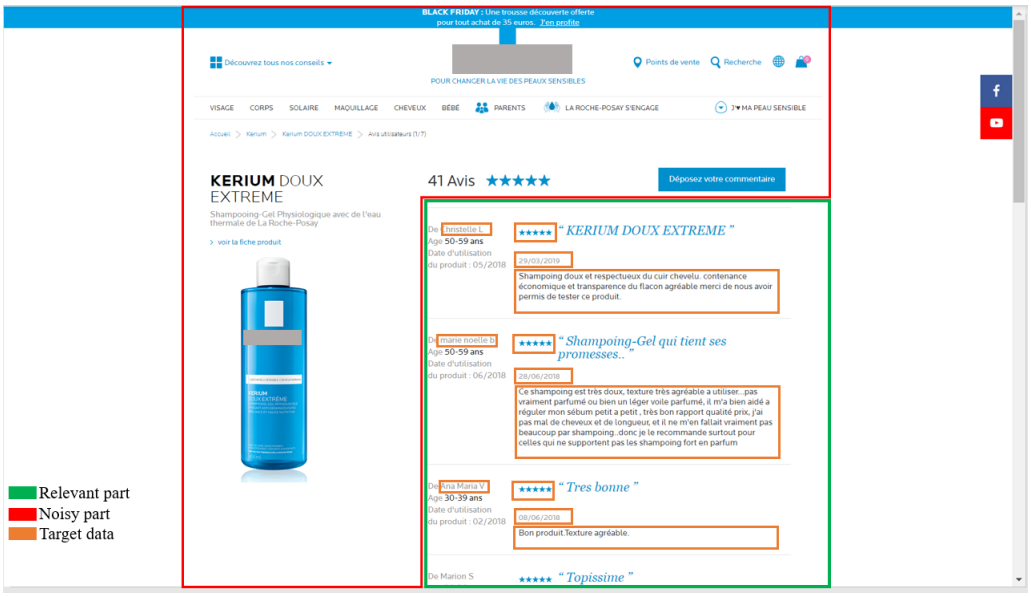


FIGURE 2.10 – Section pertinente vs Section avec bruit

ject Model) décrivant n'importe quelle page web. Sur le Web, chaque URL est considérée comme un document décrit par un modèle objet en forme d'arbre hiérarchique représentant comment l'information de la page interne est organisée en plusieurs composants élémentaires avec leurs contraintes de positionnement. Chaque composant est associé à une classe de feuille de style (CSS) appliquant des caractéristiques visuelles et de rendu.

Cette étape d'analyse a révélé que l'élément HTML de pagination est le composant central utilisé dans les avis de site web pour rendre l'expérience des utilisateurs plus fluide et conviviale. L'élément de pagination permet l'organisation du contenu principal, c'est-à-dire les avis et les réponses, en plusieurs sous-groupes de n éléments répartis sur x liens de pagination. L'élément *pagination* est le composant principal de notre algorithme d'extraction des données du web' car il assure un accès direct aux nœuds les plus probablement pertinents de l'arbre DOM sans autres algorithmes d'analyse séquentielle consommateurs de temps.

Notons D la profondeur du nœud *barre de pagination* qui signifie la distance entre le nœud et l'élément racine de l'arbre DOM. La profondeur des données pertinentes est très probablement soit au même niveau de profondeur que le nœud *élément de pagination*, soit la profondeur précédente où $0 < n < 2$.

L'accès direct à la section cible pertinente réduit considérablement le temps d'exécution d'extraction car il évite de parcourir tous les éléments de l'arbre DOM.

Le processus d'extraction récupère de la section de page pertinente le texte de l'avis et les informations relatives à l'ID de l'évaluateur, au score d'évaluation, à la date de publication et à la localisation lorsqu'ils sont fournis. Ces caractéristiques supplémentaires sont utiles pour l'annotation des jeux de données d'entraînement dans les tâches d'apprentissage automatique, telles que la prédiction des tendances temporelles et géographiques, des scores positifs et négatifs. Dans notre algorithme, la présence de ces caractéristiques est détectée en utilisant des éléments de correspondance de motifs structurels, détaillés dans la section suivante.

Définir la logique d'extraction des données à partir du web

Le traitement d'extraction des données à partir du web est principalement basé sur les éléments de l'arbre DOM (tags) et les classes de feuilles de style. Il est écrit en langage de programmation Python en utilisant en partie le framework Scrapy précédemment introduit [123] et le langage de requête XPath [28] pour sélectionner les nœuds dans des fichiers semi-structurés. XPath simplifie l'accès direct aux nœuds cibles en introduisant trois concepts : l'axe de navigation, le test de nœud et les prédicats. Il convient de noter que Scrapy ne représente qu'une partie du framework de développement tandis que la logique et les règles d'extraction sont implémentées par notre algorithme.

Accès direct à la partie la plus pertinente d'une page web donnée

Dans le cadre de cette étude, l'élément de pagination fait référence à un élément HTML impliqué dans la navigation entre différentes parties de contenu. Il peut être défini par un nœud HTML tel que , <nav/>, <a/> ou <button/>, ayant au moins une classe de feuille de style qui correspond entièrement ou partiellement au lemme des mots 'page' et 'plus'. Cette spécification découle de notre analyse préliminaire et est utilisée pour identifier les éléments de navigation spécifiques sur une page web. Nous notons que les pages web sans élément de pagination ne sont pas prises en charge par notre outil.

En effet, cette fonction est complexe et joue un rôle crucial dans le processus d'extraction car elle aide à ne considérer que la partie de la page pertinente à chaque cas d'utilisation. Elle utilise la requête de navigation créée pour parcourir l'architecture DOM de la page. La fonction pointe directement vers un sous-arbre spécifique du DOM, qui correspond très probablement à une liste de blocs d'avis. Cette liste contient toutes sortes d'informations nécessaires pour le cas d'utilisation en question.

Reconnaissance et extraction des données complémentaires

Nous mettons en œuvre une fonction de correspondance de motifs pour rendre l'extraction d'informations complémentaires plus adaptable et générique. La fonction *extra_info_tags_recognition()* fonctionne comme suit. D'abord, nous considérons comme entrée la partie HTML précédemment ciblée. Ensuite, nous la convertissons en format chaîne en concaténant ses balises composantes avec leur classe de feuille de style correspondante, en omettant la balise racine. Ensuite, nous recherchons des motifs de chemin récurrents pour localiser le bloc d'avis unitaire (le texte de l'avis, la date, l'évaluation et la localisation). Chaque élément des informations complémentaires extraites est distingué par une liste prédéfinie de classes de feuilles de style fréquemment utilisées, déduites de l'étape d'inspection précédente. Étant donné que les sites web commerciaux, les forums et les blogs présentent des avis *via* de multiples liens hypertextes, WEBT-IDC gère automatiquement les URLs de pagination et applique l'algorithme défini ci-dessus sur chacun d'entre eux. La figure 2.11 illustre le processus de correspondance de motifs (*pattern recognition*) appliqué à une page web ayant l'architecture HTML présentée précédemment dans la figure 2.10.

FIGURE 2.11 – Reconnaissance des motifs dans des informations complémentaires

Création corpus thématique et multilingue

Pour rendre les avis collectés exploitables par d'autres tâches de *machine learning* (ou d'autres tâches de TALN), nous les stockons dans des fichiers CSV semi-structurés séparés, regroupés par thème et par langue. Ainsi, chaque fichier représente un sujet concernant soit un produit, soit un service, exprimé dans une langue différente. Chaque ligne présente le retour d'information de l'utilisateur décrit par le texte du commentaire, le nom de l'utilisateur, la date de publication et la localisation (lorsqu'elles sont fournies), ainsi que des annotations supplémentaires comme le thème correspondant et la langue dans laquelle le commentaire est exprimé.

Afin de conserver l'efficacité de notre outil pour récupérer les données du web en termes de temps d'exécution, nous avons appliqué le paradigme du multi-threading à deux niveaux :

— *Extraction des URLs sources* : Le composant d'extraction des données à partir du web prend en entrée un fichier texte simple, qui contient une liste d'URLs pertinentes. Ces URLs sont sélectionnées en fonction de la requête de l'utilisateur et proviennent de l'étape de recherche et de filtrage précédente. Ces pages web sont analysées en parallèle à l'aide de plusieurs fils d'exécution (*threads*) en Python, respectant un degré de parallélisme prédéfini *n*. Par défaut, le paramètre *n* est fixé au nombre d'URLs en entrée. Toutefois, il peut être ajusté librement pour s'adapter aux ressources matérielles disponibles.

— *URLs de pagination* : Pour les pages web de forums et de blogs utilisant le mécanisme de pagination HTML, nous avons défini un degré de parallélisme *p* afin de traiter simultanément plusieurs URLs de pagination.

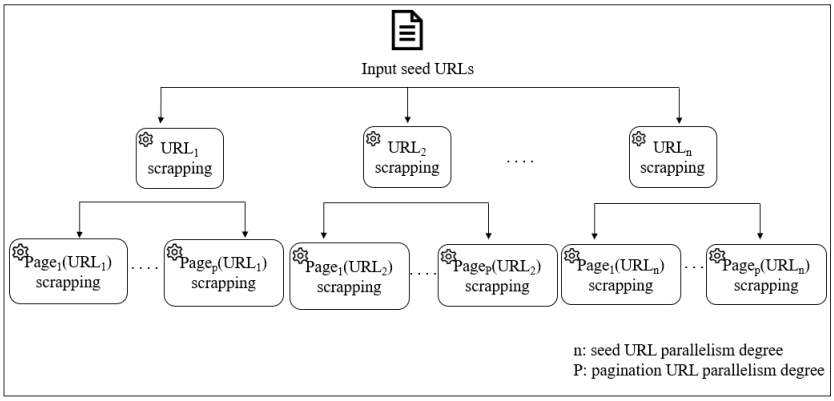


FIGURE 2.12 – WEBT-IDC parallélisation du processus

La figure 2.12 montre le traitement parallèle appliqué pour optimiser le temps

d'exécution du processus d'extraction des données. Les deux degrés de parallélisme dépendent des performances des dispositifs électroniques, tels que le disque dur et la mémoire. Pour faciliter l'ajustement des degrés de parallélisme, nous avons décidé de fixer p en référence au nombre de paginations pour une URL donnée.

Expériences et résultats

Afin d'évaluer la performance de l'algorithme en termes d'adaptabilité, de temps d'exécution et de précision d'extraction, nous avons effectué plusieurs cycles d'expériences, chaque fois avec un thème spécifique différent. Ci-après, nous désignons par thème soit un produit (smartphone, crème de beauté) soit un service (services aériens, services de livraison). En tant que données d'entrée, nous devons uniquement fournir la requête de recherche de l'utilisateur, incluant l'information sur le sujet. Pour les tests, nous avons configuré le composant de crawling (cette fonction est adaptable) pour collecter les dix premières URLs (différentes) renvoyées par le moteur de recherche, considérant qu'elles incluent les pages web les plus pertinentes correspondant à la requête donnée.

Les URLs web explorées étaient structurées en 25 architectures HTML différentes et exprimées dans différentes langues (français, anglais, espagnol, italien, allemand, russe et grec).

Toutes les expériences ont été réalisées sur un ordinateur portable dual-core avec 8 Go de RAM. Nous avons ajusté le degré de parallélisme de l'URL source pour qu'il soit égal à la taille de la liste de sortie crawlée pour chaque exécution d'expérience. Cependant, pour une page web donnée, le degré de parallélisme, noté p, est dynamique et s'adapte au nombre d'URLs de pagination.

Comme mentionné précédemment, notre outil est conçu pour être multilingue et adaptable à n'importe quelle architecture de page web, indépendamment de la langue et de la dimension de la structure. Il est automatique dans toutes les étapes de traitement, il n'est donc pas nécessaire de définir un paramètre spécifique pour déterminer la langue utilisée dans une page URL d'entrée. Elle est automatiquement déduite des paramètres et de la structure de l'URL.

Évaluation des résultats

Nous avons évalué les performances de WEBT-IDC en utilisant les métriques de précision et de rappel, ainsi que le temps d'exécution.

Le tableau 2.13 présente la comparaison entre WEBT-IDC et trois outils de 'raclage' qui ne fournissent pas la fonctionnalité de *crawling*. Tandis que le tableau 2.14

compare l'outil présenté à des outils existants construits sur des cadres permettant à la fois le 'parcours' et le 'raclage'.

Tous ces outils testés sont implémentés avec une approche d'extraction semi-automatisée qui nécessite une expertise humaine pour identifier les nœuds de données pertinents.

Ce processus de sélection doit être ajusté chaque fois en fonction des changements dans l'architecture inhérente de la page web.

Comme ces outils ne fournissent pas le mécanisme de *crawling*, nous avons utilisé notre modèle pour fixer les URLs sources d'entrée à 'racler'. Les résultats montrent que WEBT-IDC peut atteindre 100% de précision avec un processus d'extraction entièrement automatique. La performance en temps d'exécution est également compétitive.

Caractéristiques				Performances	
Scrapping Tool	Crawling	Scraping	Adaptability	Precision (%)	Temps (ms)
WebExtractor	No	semi-automatic	No	100	135
Mozenda	No	semi-automatic	No	100	72
Webscraper.io	No	semi-automatic	No	100	**62**
WEBT-IDC	**Yes**	**automatic**	**Yes**	**100**	**67**

TABLE 2.13 – WEBT-IDC résultats comparatifs.

Dans le tableau 2.14, nous avons évalué la qualité des données extraites (suite au 'raclage du web') en comparaison à l'expérience présentée dans [109], car l'outil n'est plus disponible pour de nouveaux tests. Ainsi, nous avons répété la même expérience en testant vingt pages web du domaine "ubuntu-fr.org". WEBT-IDC surpasse tous les outils considérés. Il fournit un ensemble de données sans bruit avec 100% de précision et de rappel et obtient le meilleur temps d'exécution. Pour ce test, le mécanisme de *crawling* a été désactivé pour WEBT-IDC puisque les URLs sources d'entrée étaient fixes.

	Nutch	Heritrix	BootCaT	DYCORC	**WEBT-IDC**
Precision	62,68%	**100,00%**	33,46%	**100,00%**	**100,00%**
Recall	89,79%	56,58%	98,65%	96,97%	**100%**
Scraping Time (ms)				107	**25**

TABLE 2.14 – WEBT-IDC évaluation avec les mêmes tests que DYCORC

Dans un autre exemple de performance de WEBT-IDC, sur un sujet tel que "services aériens", nous avons obtenu un nombre global de 44 348 avis répartis comme suit :

— 17 083 commentaires sont exprimés en français,
— 16 668 en anglais,
— 2 240 en espagnol,
— 3 300 en italien et,
— 5 057 en allemand

La variance observée entre les langues pourrait être partiellement expliquée par le fait que le nombre d'URL par langue n'est pas uniformément distribué sur le World Wide Web, en particulier pour les sites de critiques qui dépendent fortement des différentes cultures.

Enfin, pour évaluer la capacité de WEBT-IDC à éliminer les données bruitées lors du processus d'extraction, nous utilisons une mesure de pureté p qui calcule le nombre d'avis contenant des données bruitées proportionnellement au nombre total d'avis extraits comme exprimé dans l'équation 2.6.

$$p = \frac{Number \quad of \quad reviews \quad containing \quad noise}{Total \quad number \quad of \quad extracted \quad reviews} * 100 \tag{2.6}$$

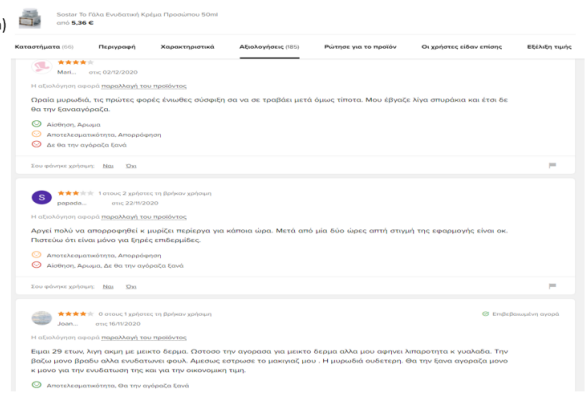
Pour l'évaluation du filtrage du bruit, nous comparons le corpus automatiquement construit d'avis sur les services aériens produit par WEBT-IDC, à un corpus de référence créé manuellement à partir d'un sous-ensemble des URLS d'output parcourues. Nous obtenons une mesure p égale à zéro, ce qui signifie que notre outil de 'raclage' correspond exactement aux données cibles et génère des données complètement exemptes de bruit.

De plus, nous avons utilisé WEBT-IDC pour extraire des avis sur des produits de beauté. Dans ce contexte, nous avons considéré des pages web dans différentes langues (français, anglais, grec, russe, italien, ...). La Figure 2.13 montre un exemple des informations extraites respectivement des sites web grec et russe.

En plus des informations extraites par notre outil (le texte de l'avis, le score de classement, la date et le lieu de publication), nous avons également pré-étiqueté chaque avis avec les principaux sujets déduits automatiquement des requêtes thématiques données au composant de *crawling* ainsi qu'avec l'étiquette de langue. Ainsi, chaque entrée du jeu de données est associée à une catégorie de produit, à une étiquette de polarité déduite de l'attribut du score de classement (lorsque le score de classement est supérieur à 3, nous estimons que l'avis est positif, négatif sinon) et à une étiquette définissant la langue utilisée pour exprimer l'avis.

Pour garantir que le jeu de données de sortie convient bien aux tâches d'apprentissage automatique, nous avons préparé un modèle prêt à effectuer une tâche de classification multi-étiquettes et multi-classes. Il repose sur le modèle pré-entraîné

(a)



Review	User	Score	Date
Ωραία μυρωδιά, τις πρώτες φορές ένιωθες σύσφιξη σα να σε τραβάει μετά όμως τίποτα. Μου έβγαζε λίγα σπυράκια και έτσι δε θα την ξαναγόραζα.	Maria.Pa.	4	02/12/2020
Αργεί πολύ να απορροφηθεί κ μυρίζει περίεργα για κάποια ώρα. Μετά από μία δύο ώρες απτή στιγμή της εφαρμογής είναι οκ. Πιστεύω ότι είναι μόνο για ξηρές επιδερμίδες.	Αναστασίαραπ adakistylania	3	22/11/2020
Ειμαι 29 ετων, λιγη ακμη με μεικτο δερμα. Ωστοσο την αγορασα για μεικτο δερμα αλλα μου αφηνει λιπαροτητα κ γυαλαδα. Την βαζω μονο βραδυ αλλα ενυδατωνει φουλ. Αμεσως εστρωσε το μακιγιαζ μου . Η μυρωδιά ουδετερη. Θα την ξανα αγοραζα μονο κ μονο για την ενυδατωση της και για την οικονομικη τιμη.	Joanna_J0	4	16/11/2020

(b)

Review	User	Score	Date
Дневной крем для лица Биокон "Натуральный уход" 25+ мне подошел идеально. Начнем с упаковки. Это, конечно, не очень удобная баночка, так как, когда откручиваете крышку, то большая поверхность крема контактирует с воздухом. В общем, не...	LyubaGrom	5	23/10/2018
Новая серия ТМ Биокон "Натуральный уход" привлекла меня прежде всего своим названием- при выборе косметики я стараюсь отдавать предпочтение натуральным средствам. И вот перед Новым годом сделала я себе небольшой подарок- купила три средства из...	NIKITA83	5	07/03/2014

FIGURE 2.13 – Exemples d'extraction d'avis depuis (a) une page web en grec (b) une page web en russe. La partie supérieure montre la page web et la partie inférieure représente le résultat de l'extraction

"bert-base-multilingual-uncased" BERT [35] (Bidirectional Encoder Representations from Transformers) capable de gérer l'encodage multilingue empilé à trois couches linéaires qui codent respectivement trois classes différentes : la langue de l'avis, la catégorie du produit et le score donné (positive ou négative).

Nous avons utilisé un jeu de données d'entrée contenant des avis liés à la beauté pour alimenter le modèle d'apprentissage automatique proposé. Il s'agit d'un jeu de données multilingue composé de 45 000 avis en anglais, 36 000 avis en français et 24 000 avis en italien. Le jeu de données d'entrée a été divisé en un ensemble d'entraînement et un ensemble de validation, avec un ratio de 8 :2.

Dans cet exemple, Bert donne des résultats de classification impeccables pour la langue du commentaire et les tâches de classification des avis des utilisateurs avec une précision de 100 %. Pour la classification des catégories de produits, une précision de 89 % est obtenue. Cette tâche est plus difficile par rapport aux précédentes, car des catégories de produits similaires existent et des mots-clés similaires sont employés pour décrire les caractéristiques des produits évalués, créant ainsi des ambiguïtés sémantiques comme les produits de soin de la peau et produits de protection solaire.

Conclusion

Nous avons présenté WEBT-IDC, un outil automatisé, qui facilite la collecte d'avis et de commentaires depuis des forums et des blogs. Il fournit un corpus filtré et structuré, prêt à être utilisé directement dans des tâches d'apprentissage automatique Nous avons montré que notre méthode, combinant à la fois des composants de *crawling* et de 'raclage', se défend bien face aux méthodes de pointe. WEBT-IDC est indépendant de l'architecture intrinsèque de la page web et donc adaptable à différentes architectures de forums et blogs qui contiennent l'élément de pagination, comme décrit dans la section 8. WEBT-IDC génère un corpus thématique multilingue sans bruit, composé des données cibles pertinentes. L'utilisation du traitement *multithread* et de l'accès direct à la partie pertinente de la page web est un facteur fondamental qui renforce les performances globales.

Les travaux futurs envisagent l'intégration de tâches d'apprentissage automatique afin d'améliorer l'identification des régions de données pertinentes et de générer automatiquement des jeux de données d'entraînement adaptés à un domaine sélectionné. De plus, nous proposons d'ajouter une autre couche fonctionnelle supplémentaire pour l'annotation sémantique superficielle. L'objectif est d'affiner les caractéristiques de l'ensemble de données extrait, pour optimiser la qualité de l'analyse.

2.3 Contributions et perspectives

Avant de faire un bilan sur les contributions apportées par toutes ces approches et techniques pleines de réflexions et idées véhiculées dans les outils présentés, nous nous devons répondre à la question sur *la nécessité de créer de nouveaux corpus alors qu'il existe déjà de nombreux corpus et datasets disponibles*, que nous considérons pertinente dans le contexte de l'essor phénoménal de notre ère numérique.

Nous répondons sans hésitation par la positive et nous exposons brièvement les raisons qui motivent notre réponse :

- La spécificité des besoins : même s'il existe une multitude de corpus disponibles, chaque projet de recherche ou d'application peut avoir des besoins très spécifiques en termes de contenu, de langue, de format, ou de contexte. Un corpus générique ne pourrait pas nécessairement répondre à des besoins pointus.
- L'évolution du contenu des pages web : puisque le web évolue constamment les contenus changent. Il y a des ajouts des modifications des suppressions.

Il faut actualiser les corpus sur les opinions si l'on veut refléter les tendances récentes par exemple.

— Biais et représentativité : les datasets existants peuvent présenter des biais, soit parce qu'ils ont été créés dans un contexte particulier, soit parce qu'ils ne représentent pas une grande partie de la population ou des langues. Créer un nouveau corpus permet d'adresser ces biais et travailler avec des datasets plus représentatifs.

— Adaptabilité aux nouveaux modèles : avec l'évolution rapide des modèles de *machine learning*, en particulier dans le domaine du TALN, les besoins en termes de composition de données peuvent changer ; souvent il faut enrichir les datasets pour ces nouveaux modèles.

— Confidentialité et droits d'utilisation : certains corpus disponibles peuvent avoir des restrictions d'utilisation, ce qui pourrait limiter leur utilisation dans différents projets. Créer son propre corpus offre une maîtrise complète des données et de leurs droits d'utilisation.

— Les projets de recherche : la création de nouveaux corpus est essentielle pour la recherche. Tester de nouvelles hypothèses ou développer de nouvelles méthodes nécessite souvent des données spécifiques.

Donc en réponse à cette question nous affirmons que la création de nouveaux corpus reste essentielle pour répondre aux besoins spécifiques des projets de recherche et d'application malgré la disponibilité de nombreux corpus et datasets de et pour la communauté ML.

2.3.1 Contributions

Nous présentons les contributions des approches et outils décrits dans les sections précédentes, qui ont permis d'avancer dans différents projets de recherche.

Ce qui distingue nos approches sont l'intérêt que nous portons aux langues sous-dotées [28], nos outils ont été conçus et testés pour les langues peu représentées comme le grec et l'arabe. Le web est largement dominé par quelques langues majeures comme l'anglais, le chinois, le russe, l'espagnol, etc. Cela pose des défis pour la diversité linguistique et culturelle, car de nombreuses langues et leurs cultures associées sont sous-représentées en ligne. Nous faisons l'effort de promouvoir la diversité linguistique sur le web, en développant des technologies pour soutenir ces langues.

Un autre point qui distingue nos approches est notre effort constant pour réduire

28. Langues sous-dotées désignent les langues avec une faible représentation sur le web par rapport à leur nombre de locuteurs. Pour certaines, elles peuvent avoir des millions de locuteurs, mais une présence digitale réduite. Cette sous-représentation peut être due à divers facteurs, comme un manque de contenu numérique produit dans ces langues.

voire éliminer le *bruit* des données extraites, afin de produire un résultat sans post-traitement de nettoyage, prêt à être utilisé comme corpus de référence ou corpus d'entrainement et de test pour différentes tâches du *machine learning*.

Nous décrivons, brièvement, les caractéristiques qui ont contribué à avancer dans la conception des fonctionnalités des outils pour qui permettent la création des corpus des données extraites du web.

— **REVSCRAP** : notre premier outil qui intégrait un *crawler* et un *scraper* (pour les pages choisies) a démontré sa capacité à construire des corpus thématiques et multilingues. Il a permis une extraction efficace et ciblée d'avis pertinents à partir de sites variés. Les améliorations potentielles de REVSCRAP comprennent une meilleure précision dans le filtrage des données et une extension pour couvrir davantage de langues. Cet outil a créé les corpus multilingues, qui ont été utilisés pour les besoins de recherches sur l'analyse d'opinion, sous ma direction, de la doctorante Mme Lisa Medrouk (thèse soutenue le 6 décembre 2018).

— **DYCORC** : cet outil, spécialisé dans l'extraction de contenu de forums web, s'est avéré supérieur en termes de précision par rapport aux modèles existants. L'association possible de DYCORC avec d'autres outils, comme BootCat, pourrait améliorer encore les performances en termes de vitesse d'extraction. L'approche choisie, basée sur la distance de Levenshtein, visait à réduire voire éliminer le bruit des données extraites. Elle a fait l'objet des recherches du doctorant M. Otman Manad (thèse soutenue le 6 mars 2018).

— **WEBT-IDC** : la dernière version de notre outil de *parcours et extraction des données à partir du web* performant pour la collecte d'avis depuis forums et blogs. Il se distingue par sa capacité à s'adapter aux diverses architectures web, et le rend particulièrement précieux pour la création de corpus thématiques multilingues. Avec l'intégration de techniques d'apprentissage automatique, son efficacité pourrait être accrue, offrant une valeur ajoutée pour les utilisateurs finaux. Cet outil a permis de créer les corpus [29] qui sont utilisés pour les besoins de recherche, sous ma direction, de la doctorante Mme Maroua Boudabous (thèse en cours).

29. Un corpus par langue, par produit, utilisés séparément et ensemble

2.3.2 Travaux actuels et perspectives

Nous avons deux projets de recherche en cours le projet MALANTIN [30] et le projet CLEXIC, nécessitant des corpus spécialisés. Les deux projets regroupent les recherches du domaine d'informatique du point de vue technologique et du domaine de sciences humaines et plus particulièrement d'économie, de point de vue lexique spécialisé. Les deux portent sur *la création automatique des lexiques spécialisés autour du concept de l'innovation technologique et non technologique*. Pour illustrer l'évolution de nos outils et comment ils s'adaptent aux besoins particuliers de divers projets, nous décrivons brièvement une caractéristique qui s'ajoute pour mieux permettre la création des corpus spécialisés.

Projet MALANTIN (2019-2022) (MAchine Learning for Analysing Non Technological INnovation) : Dans le cadre de ce projet, nous avons créé un corpus composé de données textuelles issus du web, suivant un mode opératoire nouveau afin de satisfaire les critères de sélection que nous ont été fixés par les collègues économistes avec lesquels nous collaborons pour le thème du corpus. Ici une des caractéristiques particulières est l'ajout d'un filtre au niveau url afin de délimiter le nombre des pages candidates dont le contenu sera extrait.

Le corpus doit comporter, pour l'intérêt du projet, des données issues (des pages web) des entreprises faisant partie de 27 catégories spécifiques du domaine économique (pharmaceutique, petrolière, chimie, finances, etc.).

Afin de cibler les pages web pertinentes nous avons utilisé des mots-clés pour cibler et filtrer à la fois les urls candidates à nos requêtes. Ceci nous a permis de mettre en évidence les mots-clés et les contextes sémantiques pertinents pour identifier l'innovation dans divers secteurs.

Pour mieux illustrer le filtrage des urls, nous donnons un exemple extrait de nos résultats. Suivant les retours à nos requêtes, une des entreprises dans la catégorie finance est la *COOPERATIEVE RABOBANK U.A.*. Plusieurs pages font partie de cette entreprise. Pour mieux cibler les pages comportant le contenu qui nous intéresse pour notre corpus, nous avons recherché celles comportant un des 4 mots-clés (que nous avons désignés [31] pour notre thème : "innovation", "research", "development" et "design") et voici un extrait des liens :

1. https ://www.rabobank.com/en/raboworld/articles/the-good-fashion-fund-is-**redesigning**-a-dirty-industry.html

30. le projet MALANTIN est officiellement fini au niveau financement en 2022, mais nous poursuivons nos travaux avec les autres participant-e-s pour perfectionner le lexique spécialisé créé.

31. le choix de ces mots-clés a été effectué après étude sur les urls de plusieurs entreprises en cherchant des motifs qui pourraient contenir des textes ayant un rapport avec l'innovation

2. https ://www.rabobank.com/en/**research/**index.html
3. https ://research.rabobank.com/markets/en/home/index.html
4. https ://www.rabobank.com/en/about-raboban**/innovation/**design-en-innovation/how-we-innovate.html
5. https ://www.rabobank.com/en/about-rabobank**innovation/**food-and-innovation/index.html

Certains urls comportent des mots-clés et d'autres non, ici la ligne 3 montre l'absence de mot-clé. Donc le scrapeur n'ouvrira pas cette page pour récupérer le contenu.

Cette étape constitue un premier filtre pour éliminer le bruit (à cette phase du processus le *bruit* sont les urls non pertinentes pour notre thème), ainsi qu'une catégorisation par secteur et mot-clé.

Projet CLEXIC (2023) (Création LEXique Innovation Crowdfunding) : Dans le cadre de ce projet, nous avons créé un corpus composé de données textuelles décrivant des projets présents sur les sites de crowdfunding. Chaque plateforme crowdfunding comporte des liens vers des centaines voire des miliers des projets participatifs.

Pour ce projet, dont la particularité consiste à la présence de milliers de liens à partir d'une url, nous avons cherché à optimiser le processus en parallélisant la lecture des arborescentes html différentes pour chaque site de chaque projet. Le projet étant en cours, il y aura éventuellement d'autres ajouts ou modifications pour mieux construire le corpus final qui servira par la suite de corpus d'entrainement et de test pour le modèle d'apprentissage.

Nous cherchons à développer des outils de 'raclage du web' qui sont spécifiquement adaptés à nos besoins [32]. Cela inclut la capacité à naviguer avec intelligence en tenant compte des mécanismes anti-robot, tels que les fichiers robots.txt, pour accéder au contenu souhaité. De plus, nous ciblons uniquement le texte pertinent, en éliminant le bruit, et nous pouvons récupérer certaines balises pour annoter partiellement les éléments du corpus créé. Ces fonctionnalités sont construites

32. En développant ces outils et méthodologies, nous nous engageons à respecter pleinement les aspects éthiques et légaux associés au *web scraping*, en particulier dans notre travail avec les forums, blogs et sites commerciaux contenant des avis sur les produits ou services. Nous nous conformons aux directives des fichiers robots.txt et nous assurons que l'extraction de données est effectuée en respectant les lois sur la propriété intellectuelle, la confidentialité et les conditions d'utilisation des sites web ciblés. De plus, il est important de noter que les corpus que nous créons ne sont utilisés que pour notre recherche académique et ne sont pas divulgués à des fins commerciales ou autres. Notre objectif est de promouvoir une pratique responsable et éthique de l'extraction de contenu web, qui respecte les droits des propriétaires de sites web et les intérêts des utilisateurs.

en adaptant et en étendant les bibliothèques existantes de Scrapy, en ajoutant des règles et des fonctions spécifiques pour réaliser ces tâches.

Chapitre 3

Apprentissage profond pour l'analyse de sentiment et classification thématique sur corpus multilingues

Sommaire

Dans ce chapitre, nous explorons l'analyse de sentiment [1] et la classification thématique en contexte multilingue, en nous concentrant particulièrement sur l'exploitation de données sans pré-traitement. Ces données se composent principalement d'avis d'utilisateurs [2] issus de divers domaines [3]. Notre étude s'appuie sur les CNN [91] et les RNN [153], [42], en particulier les LSTM [63], pour extraire des caractéristiques sans dépendance linguistique, sans nécessiter de pré-traitement ni d'annotation.
Les mots sont représentés par des vecteurs à hautes dimensions et les modèles utilisés sont formés pour en extraire les caractéristiques utiles.
Notre étude met également en évidence la complémentarité des CNN et des RNN dans la détection des corrélations de voisinage et des informations contextuelles à longue distance. Ils fournissent des informations complémentaires en classification de texte[184]. Les résultats, basés sur des corpus multilingues de critiques de restaurants et d'hôtels, démontrent l'efficacité et la précision de notre approche non supervisée.
Le chapitre se lit comme suit : nous introduisons le cadre théorique et technologique de nos recherches (3.1), nous dressons la problématique (3.2), en faisant référence au contexte des travaux connexes (3.3) et le corpus utilisé 3.4. Nous continuons avec la description des modèles proposés (3.5), des expérimentations menées et les résultats obtenus (3.6), et nous terminons en soulignant notre contribution (3.7).

1. Dans le domaine du TALN, il existe plusieurs termes qui se chevauchent pour décrire l'analyse computationnelle des opinions, des sentiments et de la subjectivité dans le texte. Ces termes incluent "fouille d'opinions", "analyse de sentiment", et "analyse de subjectivité" (en anglais *opinion mining*, *sentiment analysis*, *subjectivity analysis* selon [127]. Ils sont souvent utilisés de manière interchangeable et abordent des problématiques similaires, tout en mettant l'accent sur des aspects spécifiques de l'analyse. Nous nous concentrons sur la classification du sentiment exprimé dans un texte comme étant positif, négatif ou neutre, l'objectif étant de comprendre l'émotion ou le ton global.

2. Dans nos travaux le terme "avis" se trouve souvent interchangeable avec les synonymes comme "critique", "opinion", "commentaire", "revue" sans porter une spécificité sémantique particulière. Le terme *multilingual reviews corpus* [79] qui désigne un corpus d'avis multilingues se trouve, dans la littérature, est traduit de différentes manières : "commentaires des utilisateurs" : "évaluations des utilisateurs", "témoignages des utilisateurs", "appréciations des clients", "retour ou feedback des clients", "observations des clients" et plein d'autres synonymes qui désignent 'l'avis' qu'un internaute laisse sur un produit ou un service, sur le web.

3. Ces travaux font également partie de la recherche doctorale, sous ma direction, de Mme Lisa Medrouk. Thèse soutenu le 6-12-2018 [113, 114].

3.1 Introduction

L'avènement de l'apprentissage profond a transformé de nombreux domaines de la recherche, y compris l'analyse de sentiment. Dans l'univers du web, les avis utilisateurs dépassent les frontières linguistiques. Cette richesse multilingue offre une grande complexité, rendant nécessaire l'emploi de méthodes à la fois robustes et évolutives.
Bien que les programmes informatiques soient capables de comprendre nos énoncés de base, ils rencontrent des difficultés mesurables pour saisir les nuances du langage. Des aspects comme le sarcasme ou l'humour ainsi que d'autres nuances culturelles dans différentes langues augmentent la complexité de l'approche et la pertinence de ce type d'analyse.
Le défi de l'analyse des opinions multilingues demeure donc pertinent [4].
Le choix de l'apprentissage profond dans cette étude s'explique par sa capacité à apprendre des représentations complexes[34] sans l'intervention manuelle, ce qui est particulièrement utile pour aborder le paysage multilingue et complexe de la classification des sentiment et des thèmes. Ainsi, nous avons exploré le potentiel des différentes architectures de réseaux neuronaux profonds, notamment les CNN, les RNN et les LSTM, pour la classification de sentiment et de thèmes dans un contexte multilingue complexe.

3.2 Problématique

Face à l'abondance d'avis multilingues générés chaque jour sur diverses plateformes, comment pouvons-nous créer un modèle efficace d'apprentissage profond pour :

- Analyser et classifier efficacement la polarité des avis dans leur langue d'origine sans recourir à la traduction, préservant ainsi la richesse des avis originaux ?
- Distinguer avec précision entre différents types d'opinions, même lorsque les domaines sont sémantiquement similaires, tout en utilisant un corpus multilingue ?
- Montrer que les architectures de réseaux profonds, par leur conception intrinsèque, peuvent extraire des caractéristiques pertinentes indépendamment de la langue ou du contexte du texte ?

L'apprentissage profond, offre une approche prometteuse pour relever ces défis, notamment par sa flexibilité dans le choix des architectures et des modèles. De plus, cette technique pourrait bien être la clé pour surmonter les limitations actuelles

4. Nous présentons des statistiques qui soulèvent des questions importantes sur l'accessibilité et la diversité langagière dans le monde du web dans la section 3.3

liées à la rareté des ressources dans les langues moins dotées [5], dont le cadre de leur informatisation est décrit en profondeur dans [17] et démontrer la robustesse des réseaux profonds dans des contextes multilingues et multidomaines sans aucun pré-traitement.

3.3 Contexte de travaux connexes

Nous présentons l'état de l'art [6] en analyse de sentiment et analyse d'opinion en général, et en rapport avec l'apprentissage profond en particulier, tout en mettant l'accent sur l'aspect multilingue.
Pour souligner cet aspect et appuyer notre motivation, nous présentons divers statistiques sur les langues et leur présence sur le web.

Statistiques sur les langues et leur présence sur le web
Nous nous concentrons sur les travaux réalisés dans des langues autres que l'anglais [7], largement couvert dans la littérature consacrée à la fouille d'opinions [128] [102]. Le choix de références dans des langues dites moins dotées est pertinent pour montrer les efforts de la recherche pour la création de ces ressources.
Pour illustrer statistiquement le contraste entre les langues parlées par des locuteurs natifs, les langues les plus parlées dans le monde et les langues les plus présentes sur le web, nous exposons les tableaux suivants [8] :

- La proportionnalité des langues parlées par nombre de locuteurs natifs est donnée dans le tableau 3.1.
- Le tableau 3.2 montre les 10 langues les plus parlées dans le monde.
- Le tableau 3.3 montre les langues les plus fréquemment utilisées en fonction du contenu web.
- Le tableau 3.4 montre la proportion de la population mondiale par régions et la proportion de cette population qui utilise internet.

Il faut préciser que pour le tableau 3.3 nous avons retenu les langues ayant plus de 1% de présence sur les sites web [9].

5. Nous donnons une définition pour les langues dites "moins dotées" ou "sous-dotées" à 2.3.1

6. L'état de l'art présenté ici représente (pas de manière exhaustive) les travaux de la période de notre recherche dans ce domaine (2014-2018). Néanmoins, les tableaux qui donnent des valeurs statistiques sur les langues sont actualisées afin de faciliter la lecture et leur vérification.

7. Informations sur le nombre de langues officielles par l'ONU, voir https://fr.wikipedia.org/wiki/Liste_des_langues_officielles#Langues_officielles_de_l'ONU.

8. Dernière date de sources consultées le 06/05/2023 : `https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world`.

9. Nous présentons les données dont la dernière mise à jour est effectuée en janvier 2023, source :
https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/.

Langue	**Nombre de locuteurs natifs**	**% de la population mondiale**
Chinois	**1 300 million**	**16,25**
Espagnol	485 million	6,06
Anglais	373 million	4,66
Arabe	362 million	4,53
Hindi	344 million	4,30
Bengali	234 million	2,93
Portugais	232 million	2,90
Russe	154 million	1,93
Japonais	125 million	1,56
Lahnda (Pendjabi occidental)	101 million	1,26
Total : **10 langues**	**3,7 milliard**	**66,27%**

TABLE 3.1 – Top 10 des langues par nombre de locuteurs natifs

Nous voulons souligner que plus que la moitié de la planète, 66,27% de la population mondiale, est locuteur natif ou locuteur de une de 10 langues les plus parlées au monde, comme le montrent les tableaux 3.1 et 3.2. Une remarque au fait que deux langues l'anglais et le mandarin sont parlées à elles deux par plus d'un tiers de la population de la planète, mais l'anglais couvre plus de moitié, exactement 58,8 % des pages web, incontestablement dominante, loin devant des autres langues.
Cependant, cette réalité n'est pas reflétée dans le contenu web, par exemple la langue ourdou est quasi inexistante : moins de 0,1 % de tous les sites web dont la langue du contenu est connue [10]. Nous observons que le tableau 3.3 ne contient pas des langues comme l'Hindi ou l'Arabe [11], pourtant très parlées dans le monde. Par contre, nous trouvons 7 langues (marquées en *italic*) sur un total de 14 ayant une présence dans le contenu web alors qu'elles sont parlées chacune par moins de 150 millions de locuteurs de par le monde [12]

Et si nous voulons mettre en évidence le nombre de personnes qui utilisent internet dans différentes régions du monde [13], nous constatons que la distribution d'utilisateurs d'Internet est disproportionnée par rapport aux langues parlées dans ces

10. Voir plus d'informations sur la répartition des langues sur le web ici : `https://w3techs.com/technologies/details/cl-ur-`.

11. L'arabe représente 0,9% du contenu web, selon la même source.

12. Pour le persan, les chiffres varient selon la source. Selon `https://maisonpersane.fr/langue-persane/`, il y a 120 millions de locuteurs dans le monde, contrairement à `https://www.donneesmondiales.com/langues/persan.php` qui mentionne 60,5 millions de locuteurs de la langue persane.

13. Source : `https://internetworldstats.com/stats.htm`

Langue	Nbr de locuteurs au monde	%(8 milliard)
Anglais	**1,4 milliard**	**18,15**
Mandarin Chinois	**1,1 milliard**	**13,99**
Hindi	602 million	7,53
Espagnol	559 million	6,99
Arabe Standard	274 million	3,43
Français	274 million	3,43
Bengali	273 million	3,41
Russe	258 million	3,23
Portugais	258 million	3,23
Ourdou	231 million	2,88
Total : **10 langues**	**5,2 milliard**	**66,27**

TABLE 3.2 – Top 10 des langues les plus parlées au monde

régions. Le tableau ci-dessous offre un aperçu de cette disproportion.

Nous avons présenté différentes statistiques sur les langues parlées par les locuteurs natifs, les langues parlées par tous (en deuxième langue par exemple), leur présence sur le web et, la proportion des utilisateurs internet par régions du monde. Nous pouvons observer comment est repartie l'accessibilité et la diversité linguistique sur Internet et souligner ainsi notre motivation pour des modèles multilingues.

3.3.1 Analyse de sentiment et réseaux profonds

Nous présentons les travaux connexes en analyse de sentiment et fouille d'opinions [14], en appuyant particulièrement l'aspect multilingue.

Avec l'essor des NTIC (Nouvelles Technologies de l'Information et de la Communication), les médias sociaux (Web communautaire) sont devenus un canal majeur pour les consommateurs et les entreprises. La multiplication des forums, des blogs, des avis de produits, etc., a considérablement augmenté la quantité d'opinions disponibles en ligne . Cette abondance de données a contribué à l'émergence et au développement du domaine de la fouille d'opinion. La recherche a commencé dans les années 2000 [59] [188] et couvre différentes problématiques liées à l'opinion, comme la classification de textes subjectifs, la détection de l'orientation sémantique des mots et des expressions, ou la construction de ressources lexicales.

14. Ces termes sont souvent employés de manière interchangeable. Selon Pang and Lee [128], ces termes se réfèrent à l'étude et au traitement informatique de l'opinion, du sentiment, et de la subjectivité dans les textes. Ils englobent des notions comme l'analyse de l'affect, des émotions, de la subjectivité, fouille de sentiment ou *review mining* [101].

Langues les plus présentes sur le web	**% contenu web**
Anglais	**58,8**
Russe	5,3
Espagnol	4,3
Français	3,7
Allemand	3,7
Japonais	3
Turc	2,8
Perse	2,3
Chinois	1,7
Italien	1,6
Portugais	1,5
Vietnamien	1,4
Néerlandais/Flamand	1,2
Polonais	1,1
Total : **14 langues**	**92,4**

TABLE 3.3 – Pourcentage des langues des sites web, en *italique* celles qui ne font pas partie de top 10 de langues les plus parlées.

La classification de la polarité des sentiment, souvent exprimée en termes d'aimer ou de ne pas aimer, a été largement étudiée. Les contextes peuvent varier, comme dans l'analyse des discours politiques où la polarité peut signifier le soutien ou l'opposition à une question [128]. Les études actuelles utilisent principalement quatre catégories de caractéristiques : syntaxiques [129], sémantiques [176], basées sur les liens [40], et stylistiques [189].

Trois grandes approches classent la fouille d'opinions : supervisée [129], non supervisée [176] et hybride [7, 8]. Une nouvelle méthode, appelée approche semi-supervisée [200], ou faiblement supervisée, a émergé, combinant des données non étiquetées et partiellement annotées pour construire de meilleurs classificateurs. L'approche supervisée nécessite un jeu de données étiquetées où les opinions sont classées comme positives, négatives ou neutres. L'approche non supervisée utilise des lexiques [15], et n'a donc pas besoin de données étiquetées. [182] comparent des méthodes supervisées pour l'analyse de sentiment dans un environnement multilingue sans réseau profond. L'approche hybride combine les deux précédentes, en utilisant une combinaison de données non étiquetées et partiellement annotées.

15. Chaque mot du lexique a un score qui représente sa polarité, à savoir positif, négatif ou neutre. Le lexique peut être créé à partir de dictionnaires existants ou d'un corpus.

Régions du monde	**% de la population**	**% utilisateurs Internet 2017**	**% utilisateurs Internet 2023**
Asie	54,9	48,7	54,2
Afrique	17,6	10,9	11,2
Europe	10,6	17,0	13,9
Amérique Latine	8,4	10,5	9,99
Amérique du Nord	4,7	8,3	6,5
Moyen Orient	3,4	3,9	3,8
Océanie/Australie	0,5	0,7	0,6
Total monde	**100 %**	**100 %**	**100 %**

TABLE 3.4 – Statistiques sur la proportion de la population mondiale par région et le pourcentage d'utilisateurs d'internet en fonction de la population mondiale pour les années 2017 et 2023.

Nous citons également quelques approches représentatives testées en d'autres langues, comme l'analyse de la subjectivité basée sur la traduction automatique de l'anglais vers d'autres langues, pour surmonter les limites de ressources pour les langues moins étudiées, (les tests ont été effectué en roumain et en espagnol) [13]. [181] a proposé la classification de la polarité sur les tweets en espagnol avec la mise en relation des informations lexicales, syntaxiques et psychométriques.
[5] explorent l'analyse de la subjectivité et du sentiment (Subjectivity Sentiment Analysis (SSA)) au niveau de la phrase sur des textes en arabe standard moderne en explorant l'impact de différents niveaux de prétraitement sur la tâche de classification SSA et présentant un nouveau corpus annoté manuellement et un lexique de polarité spécifique à la langue [4].
Les premiers résultats, selon les auteurs[43], sur l'analyse de sentiment en danois, viennent d'une approche sur l'adaptation du domaine.

Les réseaux profonds ont déjà été utilisés avec succès pour l'analyse de sentiment monolingue, principalement sur des ensembles de données en anglais. [166] prédisent les distributions de sentiment en utilisant des autoencodeurs récursifs avec un Réseau Neuronal Tensoriel Récursif [167]. [67] explorent l'application de réseaux neuronaux récurrents profonds à la tâche d'extraction d'expression d'opinion au niveau de la phrase, alors que [30] [83] sur de multiples ensembles de données.
Des variantes de RNN et CNN ont été développées et utilisées avec succès dans l'analyse de sentiment [36, 29, 83, 186, 75]. [37] effectuent une analyse de sentiment pour des textes courts en utilisant un ConvNet allant de l'information du caractère au niveau de la phrase, nommé (CharSCNN), en utilisant deux couches convolutionnelles. [159] ont utilisé un Réseau Neuronal Convolutionnel profond

pour l'analyse de sentiment de textes courts sur Twitter. [186] ont utilisé des LSTMs pour prédire la polarité de tweets. Pour d'autres langues, [75] comparentCNNet LSTM dans l'analyse de sentiment de tweets russes. [90] décrivent une tentative de construction d'un système d'analyse de sentiment pour les tweets en indonésien, en utilisant un modèle basé sur le LSTM sans normaliseur, soulignant ainsi l'importance des tweets comme source de recherche pour l'analyse de sentiment en Indonésie. [9] explorent quatre architectures différentes [16] : DNN, DBN, et une combinaison d'autoencodeur [17] avec DBN pour l'analyse de sentiment dans des textes en arabe.

3.4 Corpus utilisés

Travailler dans un environnement multilingue implique de faire face à un manque de ressources, telles que des données textuelles étiquetées dans des langues moins dotées. Construire ce type de ressources pour toutes les langues ciblées est chronophage et coûteux. Pour pallier ce problème, nous utilisons des modèles de langage neuronaux et des techniques d'apprentissage profond dans un environnement multilingue, sans nous appuyer sur des connaissances préalables [18] ou sur un lexique, un dictionnaire bilingue [119], une indication de changement de langue ou l'utilisation d'une langue pivot.

Le web 2.0 communautaire est le canal interactif de prédilection pour des internautes souhaitant partager leurs avis. Ces avis sont consultables, disponibles et explicitement notés, et constituent un corpus d'opinions conséquent. Nous nous sommes intéressés aux avis de deux domaines : la restauration et l'hôtellerie. En plus du défi multilingue, nous ajoutons une complexité supplémentaire avec les entrées de domaines sémantiquement liés.
Les corpus constitués pour les besoins de cette recherche sont créés avec REVS-

16. Deep Neural Network (DNN) : Deep Neural Network (Réseau Neuronal Profond). Il s'agit d'un réseau neuronal artificiel avec plusieurs couches entre la couche d'entrée et la couche de sortie. Ces réseaux tentent d'apprendre des représentations de niveau supérieur et des abstractions des données d'entrée, et sont une partie essentielle de l'apprentissage profond [92].
Deep Belief Network (DBN) : Deep Belief Network (Réseau de Croyance Profond). Un DBN est un type de réseau neuronal profond qui contient plusieurs couches de variables cachées stochastiques. Les DBN peuvent être considérés comme un empilement de Machines de Boltzmann Restreintes (RBM) ou d'Auto-Encodeurs, où chaque couche sert d'entrée cachée pour la couche suivante [61].

17. Autoencodeur : est un type de réseau neuronal utilisé pour apprendre une représentation compressée des données, généralement dans un but de réduction de dimensionnalité ou de "débruitage". Un autoencodeur est typiquement formé pour *mapper* l'entrée dans une forme compressée, puis décompresser cette forme dans quelque chose qui correspond étroitement à l'original [16].

18. Notre modèle n'a pas besoin de connaissances a priori sur la langue ou la structure des phrases.

CRAP présenté dans le chapitre précédent 2.2.1. Le corpus utilisé est en français mais les modèles sont également testés sur les données ouvertes IMDB [107]. Les résultats sont satisfaisants et l'expérience montre que le choix d'hyperparamètres est plus important que la profondeur des couches.

Notre corpus principal est composé d'environ 102 460 avis, subdivisés comme suit pour les besoins expérimentaux :

1. Un corpus composé de 26 804 avis en français
2. Un corpus composé de 57 176 avis en anglais
3. Un corpus composé de 83 980 avis en anglais et en français
4. Un corpus composé de 91 816 avis en anglais,français et grec
5. Un corpus composé de 102 460 avis en anglais, arabe, français et grec

Ces avis sont classifiés par thématique (restauration ou hôtellerie) et par polarité (positive ou négative)

A titre d'exemple, le corpus en français est composé de 65242 opinions notées de 1 à 5 (1 étant la note la plus basse). Afin de maximiser l'effet de la polarité, nous avons choisi d'éliminer les opinions notées 3, ce qui constitue un corpus de travail composé de 47 818 opinions (23 909 positives et 23 909 négatives). La taille moyenne des séquences est de 429 caractères. Les couches de convolution exigent des séquences à taille fixe, nous les avons établies à 500, par la technique de *pad*, les séquences les plus longues sont coupées et les plus courtes sont complétées par des 0. Le corpus n'a subi aucun autre pré-traitement. Les modèles sont testés une première fois avec le corpus brut vectorisé, puis par des vecteurs-mots de taille 50 et 300 appris sur la Wikipédia française avec l'outil word2vec [19] [121].

Les *word embeddings*, ou plongements de mots, représentent une approche fondée sur la sémantique distributionelle, qui est une théorie selon laquelle le sens d'un mot est déterminé par son contexte d'utilisation.

Étant donné que la représentation du contexte sémantique des données textuelles constitue un aspect important de notre étude, nous discuterons des différentes approches qui ont été développées pour aborder ce défi dans la section suivante.

3.4.1 Représentations vectorielles des mots

Les données textuelles sont composées de mots liés par la grammaire, la syntaxe, et le contexte sémantique, formant une structure complexe. Cette section examine

19. Pour plus d'informations sur le code de word2Vec : `https://code.google.com/archive/p/word2vec/`

plusieurs méthodes pour transformer ces données en représentations numériques capables de capturer ces nuances, constituant ainsi un défi majeur dans le TALN.

Les méthodes classiques, basées sur la fréquence des termes, peuvent traduire les mots en vecteurs numériques, mais elles ne parviennent souvent pas à saisir les relations subtiles et les nuances de sens entre les mots. Ces représentations, bien que simples, peuvent être creuses et de grande dimension, limitant leur utilité.

Face à ces limitations, les modèles basés sur l'approche distributionnelle ont émergé, cherchant à capturer les caractéristiques des mots voisins. Cette approche s'inspire en partie de l'hypothèse distributionnelle formulée par Harris [20] [57], bien qu'elle utilise des techniques basées sur des modèles statistiques de la co-occurrence des mots, des techniques d'optimisation, et des modèles de réseaux de neurones, pour créer une représentation vectorielle des mots.

Ces méthodes peuvent être classées en trois catégories :

- **Clusters (clusters de Brown)** : Utilisant des algorithmes de clustering pour regrouper les mots [23]. Ces méthodes sont utiles pour le partitionnement mais manquent de finesse sémantique.
- **Distributionnelles** : Projetant les mots dans un espace de faible dimension en utilisant des mesures comme TF-IDF ou PMI, associées à des algorithmes de réduction de dimensions tels que LDA [19] ou LSA [171]. Elles sont efficaces mais peuvent manquer de représentations contextuelles.
- **Représentations distribuées** : Connues sous le nom de "word embeddings", ces méthodes associent à chaque mot un vecteur dense de faible dimensionnalité, capturant des traits latents du mot. Elles sont très efficaces pour capturer des relations sémantiques complexes.

La révolution dans ce domaine est venue avec l'adoption des réseaux de neurones artificiels, qui ont permis de créer des outils tels que word2vec [120] et GloVe [21] [138]. Ces techniques ont redéfini l'état de l'art, offrant une représentation plus riche et plus précise du contexte sémantique, tout en réduisant la dimensionnalité. Elles sont devenues essentielles pour le deep learning et le machine learning en général.

Les inconvénients spécifiques de word2vec et GloVe sont également notables. word2vec, par exemple, ne prend pas en compte l'ordre des mots dans le contexte et a du mal à gérer la polysémie. GloVe, d'autre part, peut nécessiter une grande

20. L'idée « the amount of meaning correspond[s] roughly to the amount of difference in their environments »(Harris 1954 :157) signifie que si deux mots apparaissent dans des contextes très similaires, ils sont probablement similaires en signification. À l'inverse, si leurs contextes sont très différents, alors leurs significations sont probablement aussi très différentes.

21. GloVe : Global Vectors, matrice de co-occurrence utilisant le corpus entier

quantité de mémoire pour construire la matrice de cooccurrence et partage les mêmes limitations en ce qui concerne la polysémie et l'ordre des mots.

Les *embeddings* permettent d'utiliser le calcul vectoriel pour effectuer des transformations sémantiques et servent de première couche de représentation pour entraîner des classificateurs supervisés. Dans notre étude, nous avons travaillé avec des combinaisons d'architectures et présenté des données brutes et pré-entraînées avec différentes tailles de vecteurs.

3.5 Modèles d'analyse de sentiment et de topics multilingue

Dans le domaine de l'analyse d'opinions multilingues [22], l'un des défis majeurs est le traitement de multiples langues sans recours à des méthodes de traduction, des langues pivots, ou des modules spécifiques par langue. L'état de l'art dans ce domaine aborde souvent ce défi en utilisant ces méthodes pour créer une représentation commune entre différentes langues, associant souvent des modules complémentaires tels que la traduction [165, 201].

Le manque de ressources est un autre problème majeur en fouille d'opinions [23] multilingue. Afin de le combattre, les études dans ce domaine se concentrent souvent sur le transfert de connaissances depuis une langue "riche," comme l'anglais, vers une autre. Ce transfert peut être réalisé par la traduction [13], l'utilisation de dictionnaires bilingues [119], ou de corpus parallèles [185] pour inférer des correspondances.
Cette recherche explore une approche différente, en testant la capacité d'un réseau à apprendre plusieurs langues en un seul flux, sans ces mécanismes de traduction ou d'adaptation. Notre méthode vise à simplifier le processus et à exploiter les capacités d'apprentissage des réseaux profonds pour une analyse multilingue plus directe. Dans ce contexte, nous avons proposé des modèles qui permettent de s'affranchir de ces limitations plutôt que de les résoudre directement par des méthodes

22. Le terme "analyse d'opinion" ou *opinion analysis* [33] est employé ici de manière délibérée, pour souligner que les méthodes et les approches présentées peuvent servir de base à des travaux futurs dans l'analyse des aspects, un sujet qui sera exploré plus en détail dans le chapitre suivant 4. Une clarification sur la terminologie associée à l'analyse de sentiment est fournie dans dès les premières lignes du présent chapitre.

23. L'emploi du terme "fouille d'opinion" ou *opinion mining*, même si souvent est employé de manière interchangeable dans le domaine de TALN, ici désigne l'extraction d'informations d'une grande quantité de données qui permettent d'identifier et de structurer des opinions à partir des données non structurées, souvent la détection de la polarité peut être considérée comme extraction d'une des caractéristiques de "fouille d'opinion". Pour plus de précisions sur les termes liés à l'analyse de sentiment, voir le début du chapitre.

traditionnelles.

Nous avons choisi de travailler sur un corpus multilingue composé de critiques extraites en trois langues (l'anglais, le français et le grec) qui constituent l'entrée de deux modèles de réseaux profonds. Les corpus utilisés sont présentés dans la section 3.4.
Nous avons transposé l'apprentissage multilingue par la présentation d'un flux multilingue en entrée du réseau de neurones. Ce flux est composé des avis collectées dans les trois langues choisies. La figure 3.1 illustre cette idée. Les avis dans les trois langues sélectionnées (français, anglais et grec) sont présentées à l'entrée du réseau.

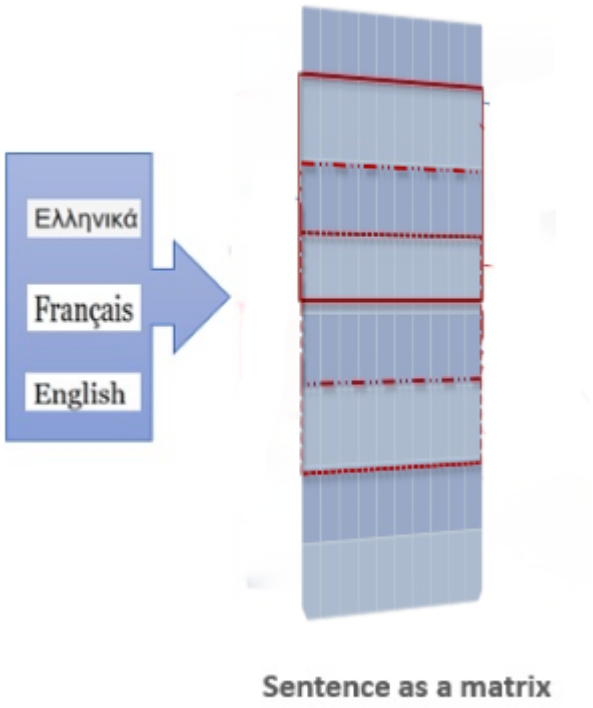


FIGURE 3.1 – Illustration de la présentation des entrées à nos modèles d'apprentissage

Les entrées sont du texte brut ; les phrases ont simplement été vectorisées en une matrice de taille $X \in R^{s \times d}$, où d représente la taille des vecteurs ligne et s représente la taille des phrases. Cette approche offre une manière unique et efficace de traiter les données multilingues, sans les complications associées aux méthodes traditionnelles de traduction ou d'adaptation.

Nous avons réalisé nos tests avec la librairie open source Keras [27] et pour les calculs matriciels, nous avons travaillé avec les unités de traitement graphique *(GPUs) NVIDIA 940M*.

3.5.1 Modèle ConvNet

Notre premier modèle est un réseau convolutionnel à une seule couche. Lorsqu'il s'agit de classification de texte, il a été prouvé par Kim [83] qu'un ConvNet à une couche de convolution est aussi performant qu'un ConvNet à plusieurs couches. Néanmoins, [31] ont proposé un modèle à 29 couches appliqué à une représentation atomique plus petite que le "mot", les caractères, en appliquant de *max pooling* avec une fenêtre de taille *trois*. Leurs résultats surpassent l'état de l'art pour la classification de texte en utilisant des données très volumineuses.

Dans cette étude, nous travaillons sur une granularité de type "mot". Ainsi, nous avons élaboré une série de tests en multipliant les couches afin de valider notre choix final. Dans notre cas, la multiplication de couches n'améliore pas l'apprentissage. Nous avons donc choisi de rester sur une seule couche de convolution.

Notre réseau est un ConvNet classique. La nouveauté vient de la multiplication des langues en entrée unique. La figure 3.2 illustre notre modèle ConvNet en exposant les différentes couches qui le constituent :

- Une entrée
- Une couche de convolution
- Une couche d'agrégation (pooling)
- Une couche entièrement connectée (fully-connected)
- La sortie (prédiction finale)

La dernière couche est une fonction d'activation sigmoïde permettant d'obtenir des probabilités d'appartenance à chaque classe (la prédiction) :

$$f(x) = (1 + e^{-x})^{-1}, f : \mathbb{R} \to [0, 1] \tag{3.1}$$

La couche d'agrégation, dite *pooling*, permet principalement de détecter des motifs récurrents dans une phrase [76]. Un pooling est une forme de sous-échantillonnage qui compresse l'information en réduisant sa taille, permettant ainsi de réduire le nombre de paramètres à calculer et contrôler aussi le sur-apprentissage.

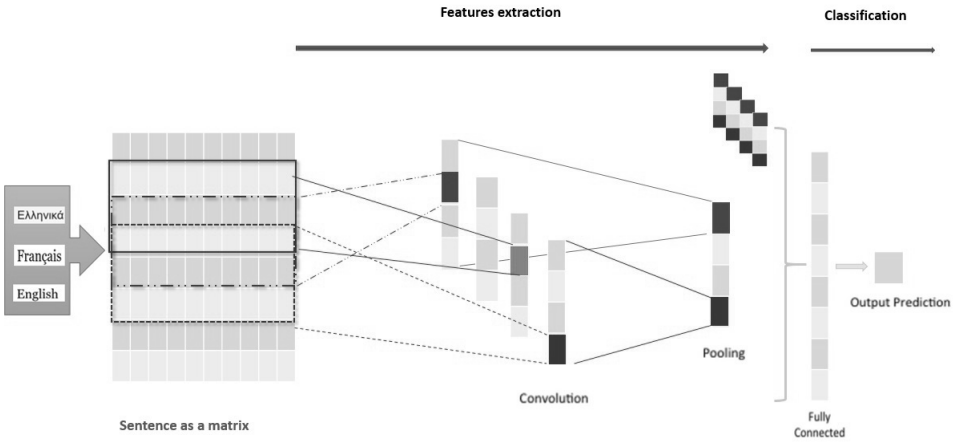


FIGURE 3.2 – Illustration du modèle ConvNet

3.5.2 Modèle LSTM

Le second modèle est un réseau Long Short-Term Memory (LSTM). Contrairement au modèle ConvNet qui utilise des convolutions, le modèle LSTM prend en compte la séquence et la dépendance temporelle des données en utilisant des cellules de mémoire.

Le modèle LSTM est composé des couches suivantes :

- Une entrée
- Une couche LSTM
- Une couche entièrement connectée (fully-connected)
- La sortie (prédiction finale)

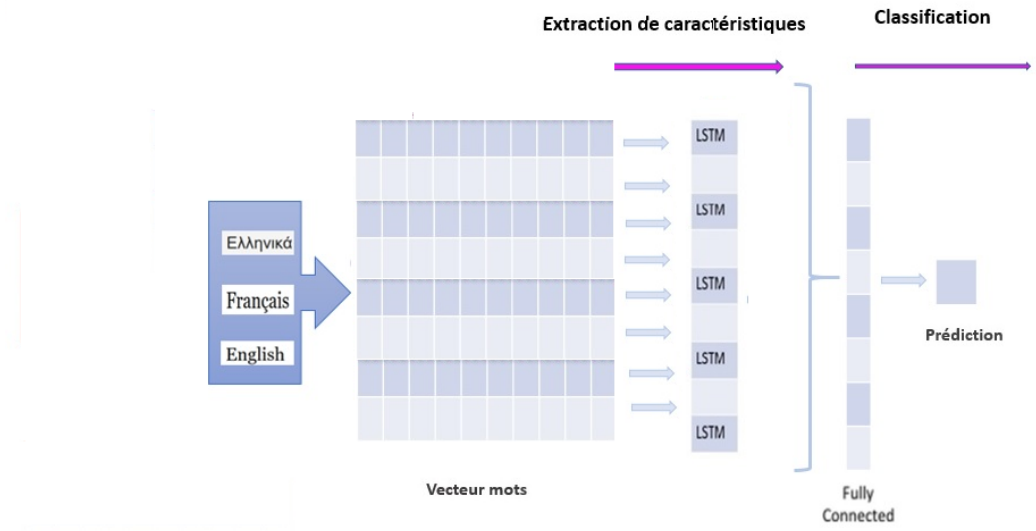


FIGURE 3.3 – Illustration du modèle LSTM

3.5.3 Modèle ConvLSTM

La combinaison de deux modèles le ConvNet hiérarchique et le RNN séquentiel est une tentative de relever le défi de la prédiction de la polarité et de la classification thématique sur corpus multilingue. La première tentative relévée est de [199] avec un modèle C-LSTM. Nous retrouvons cette combinaison dans la tentative où la complexité s'accroît, pour la tâche de classification de sentiment en langue arabe, vu sa morphologie riche [10]. Le modèle proposé se décompose comme suit :

— La couche d'entrée,
— Une couche de régularisation de type Dropout à 0.25 [24]
— Une couche convolutionnelle suivi d'une activation de type *ReLU*
— Une couche d'agrégation de type *Maxpooling*
— Une deuxième couche convolutionnelle suivi d'une activation de type *ReLU*.
— Une couche agrégation-max (max-pooling) avec taille de sous-région à 4
— Une couche de régularisation de type Dropout
— Une couche LSTM
— Une couche totalement connectée dite "FC fully connected" à 1 correspondant au nombre de classe attendu.

24. Voir section sélection des hyperparamètres 3.5.4.

— La dernière couche est une fonction d'activation sigmoïde
$f(x) = (1 + e^{-x})^{-1}$
$f : R \rightarrow [0, 1]$ permettant d'obtenir des probabilités d'appartenance à chaque classe.

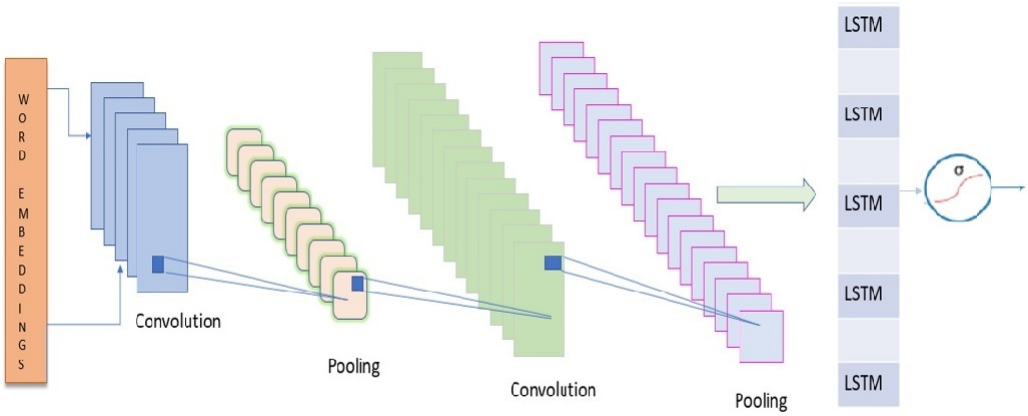


FIGURE 3.4 – Architecture (Bi)-ConvLSTM

L'apprentissage est réalisé par la méthode de descente de gradient stochastique, optimisée par la technique Adam [84] qui calcule un taux d'apprentissage pour chaque paramètre.

Bi-ConvLSTM
Pour nos essais nous avons testé aussi la bidirectionalité à l'architecture combinant la convolution à la mémoire à long terme, donnant un autre modèle hybride le Bidirectional Convolutional Long-Short Term Memory (Bi-CNN-LSTM). Les architectures, dites bidirectionnelles, peuvent prendre en compte les données futures, car elles considèrent le contexte à la fois en amont et en aval de chaque point de la séquence. Cela est utile pour "comprendre" le sens global d'une phrase ou d'un paragraphe, et grâce à cette vision "double", le modèle a une meilleure représentation de chaque élément de la séquence. Ce qui peut améliorer la performance dans des tâches comme l'analyse de sentiment où le rôle du contexte est crucial.
Les modèles présentés ont été initialisés avec des hyperparamètres qui sont présentés dans la section suivante 3.5.4.

3.5.4 Sélection des hyperparamètres

Dans le paragraphe précédent, nous avons présenté l'architecture générale des trois systèmes proposés sans entrer dans les détails des couches. Au-delà du choix d'une architecture adaptée, l'apprentissage est dépendant des hyperparamètres choisis.

Les paramètres d'un réseau sont les poids W et les biais b. Les hyperparamètres sont les facteurs déterminant la structure du réseau, tels que le nombre de couches cachées, le nombre de neurones par couches, les fonctions d'activations, ainsi que les variables qui déterminent l'apprentissage du réseau tel que le taux d'apprentissage et le nombre d'itérations. Ces hyperparamètres sont les variables qui contrôlent les poids et les biais du réseau.

Le problème du choix des hyperparamètres est très critique pour les réseaux, au vu de leur très fort impact sur les performances système. Il existe des produits open source tels que HyperOpt, Spearmint, BayesOpt, SMAC, et MOE basés sur des optimisations bayésiennes qui permettent de trouver certains hyperparamètres (tels que le taux d'apprentissage) pour des algorithmes comme SVM, random forest, et des réseaux neuronaux. Pour l'apprentissage profond, les deux méthodes utilisées sont :

— La méthode Grid Search : cette méthode permet de tester automatiquement une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage. Pour chaque paramètre, on détermine un ensemble de valeurs que l'on souhaite tester. Grid Search croise simplement chacune de ces hypothèses et crée un modèle pour chaque combinaison de paramètres. Si nous prenons un réseau profond exigeant de trouver 5 hyperparamètres, en prenant en compte 4 hypothèses par hyperparamètre, le modèle testera 4^5 donc 1024 évaluations sur une architecture profonde.

— La méthode manuelle : cette méthode correspond au même processus que la méthode Grid, mais en testant un ou plusieurs croisements à la fois en fonction du programme et/ou de la capacité de la machine.

La méthode Grid Search présentée dans le paragraphe précédent étant trop gourmande en ressources, nous avons choisi de sélectionner nos hyperparamètres optimaux par un processus manuel [25].

Pour trouver les hyperparamètres efficaces de nos modèles, nous avons établi un protocol de tests pour nous permettre de choisir les mêmes hyperparamètres pour les différentes architectures. Notre critère de sélection était de ne retenir que les configurations qui permettaient d'atteindre des performances supérieures à 85%, tout en montrant une bonne capacité de généralisation.
Par exemple, pour le filtre linéaire des couches de convolutions, nous avons testé les différentes architectures avec un filtre de taille 3, puis 4, et ainsi de suite jusqu'à 7, pour choisir finalement le filtre de taille 5 avec lequel la performance de différents modèles était toujours supérieure à 85% avec une capacité de généralisation du

25. Le nombre de tests pour arrêter le choix de ces hyper-paramètres a été réalisé par un processus heuristique manuel ne dépassant pas huit configurations différentes pour chaque modèle, et mettant les limites dans les meilleures valeurs vues dans la littérature

même ordre.
Les expérimentations ont été réalisées sur un corpus bilingue français-anglais d'avis d'hôtels et de restaurants, reflétant ainsi la nature multilingue des problèmes que nous souhaitons résoudre. Chaque test a été exécuté pendant 4 epochs, (suffisant pour atteindre un équilibre entre performance et généralisation selon nos critères). En examinant les pertes de nos ensembles de tests nous avons choisi les hyperparamètres des résultats les plus équilibrés. Le tableau 3.5 met en évidence nos choix optimaux qui ont été appliqués à toutes nos expériences.
Les modèles sont entraînés par descente de gradient stochastique sur des mini-lots de longueur 64, via la rétropropagation pour minimiser la perte d'entropie croisée binaire en utilisant l'optimiseur Adam[84].

Hyperparamètres	**Intervalle Expérimental**	**Choix**
Nombre de filtres de convolutions	30-100	64
Taille filtre de convolution	3-7	5
Régularisation-Dropout	0-1	0.25 - 0.4
Agrégation Max Pooling	2-4	4
Dimension LSTM	50-200	70
Optimisation	Adam, RMSprop	Adam

TABLE 3.5 – Hyper-paramètres sélectionnés pour les modèles convNet, LSTM, ConvLSTM, BiConvLSTM

Pour l'agrégation, nous avons choisi de travailler avec un pooling de type *max* qui ne conserve que la plus grande valeur de la taille de la sous-région sélectionnée, dans notre cas quatre. Le dropout [62] [169] est une technique de régularisation, aussi connue sous le terme d'inhibition. En apprentissage, à chaque itération, lors de la propagation, le réseau désactive un nombre de neurones choisi aléatoirement, simulant un ensemble de modèles différents et apprenant à chaque itération des sous-réseaux contenant moins de paramètres. Cette technique permet de pallier le sur-apprentissage et oblige également les neurones à apprendre indépendamment des autres, évitant la co-adaptation. Le nombre de neurones à désactiver est un hyperparamètre sous forme d'une probabilité prédéfinie. Dans le réseau ConvNet, nous avons appliqué un dropout équivalent à 40% du réseau. En résumé, nos choix d'hyperparamètres ont été soigneusement calibrés pour équilibrer les performances et la capacité de généralisation des modèles sur un ensemble de tâches diverses et multilingues. Le tableau 3.5 détaille les hyperparamètres qui ont été appliqués de manière uniforme à toutes nos expérimentations.

3.6 Expérimentations et résultats pour les modèles ConvNet et LSTM

Nous avons réalisé nos expérimentations sur 5 corpus composés de critiques monolingues, bilingues et multilingues afin de pouvoir analyser le comportement des réseaux au travers des différentes performances. Les corpus ont été décomposé comme suit :

1. Un corpus composé de 91 816 avis de restaurants et d'hôtels en anglais, en français et en grec, pour un apprentissage multilingue sur un échantillon de 64 272 avis et un test sur 27 544 avis.
2. Un corpus composé de 83 980 avis de restaurants d'hôtels en anglais et en français pour un apprentissage sur 58 786 avis bilingues et un test sur 25 194 avis.
3. Un corpus composé de 57 176 avis de restaurants et d'hôtels en anglais pour un apprentissage sur 40 024 avis et une validation sur 7 152 avis.
4. Un corpus composé de 26 804 avis de restaurants et d'hôtels en français, pour un apprentissage sur un échantillon de 18 763 avis et une validation sur 8 041 avis.
5. Un corpus composé de 102 460 avis de restaurants et d'hôtels en anglais, arabe, français et grec, pour un apprentissage multilingue sur un échantillon de 92 214 avis et une validation sur 10 246 avis.

Nous avons ajouté les corpus de avis d'opinion d'hôtel et restaurant en langue arabe de [44]. Le corpus hotel est composé de 15 572 avis, le corpus restaurant de 10 970 avis.

3.6.1 Évaluation de la performance de la classification multilingue

Les résultats de classification des thèmes (topics) pour chaque modèle sont exceptionnellement élevés, dépassant ou égalant 98% pour les deux modèles. Le tableau 3.6 présente les résultats du modèle ConvNets, tandis que le tableau 3.7 montre les résultats du modèle LSTM.
Performance est une métrique qui mesure la proportion des prédictions correctes par rapport au nombre total de prédictions. Elle est donnée par la formule suivante :

$$\text{Accuracy (Exactitude)} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}}$$

Ces résultats mettent en évidence le fait que l'entrée en langues mélangées a donné des performances équivalentes par rapport à une entrée monolingue pour les deux modèles. Ils reflètent également le fait que les modèles ConvNets et LSTM ont donné des résultats équivalents en utilisant des entrées monolingues, bilingues et multilingues.

Langues des corpus	**Accuracy (%)**
Anglais et français et grec	98.42
Anglais et français	98.42
Français	98.94
Anglais	98.16

TABLE 3.6 – Résultats de performance du modèle ConvNets pour la tâche de classification mutlithématique sur des avis d'hôtels et de restaurants en français, anglais et grec, mettant en évidence le fait que l'entrée en langues mélangées a fonctionné aussi bien qu'une entrée monolingue

Langues des corpus	**Accuracy (%)**
Anglais et français et grec	98.43
Anglais et français	98.37
Français	99.05
Anglais	97.81

TABLE 3.7 – Résultats de performance du modèle LSTM pour la tâche de classification multithématique sur des avis d'hôtels et de restaurants en français, anglais et grec, mettant en évidence des résultats équivalents pour l'entrée en langues mélangées par rapport à une entrée monolingue

Les résultats de la classification multithématique de deux modèles montrent qu'une entrée multilingue a une performance aussi élevée qu'une entrée monolingue. Il ressort également qu'il n'y a pas de différence significative en terme de performance entre le modèle ConvNet et le modèle LSTM dans le traitement de différents corpus présentés.
Les Tableaux 3.8 et 3.9 présentent respectivement les résultats de l'analyse de sentiment (et classification de la polarité) obtenus avec les modèles ConvNets et LSTM sur les corpus multilingues. Les performances sont satisfaisantes pour tous les corpus et pour les deux modèles, comme le montrent les taux ci-dessous.

Comme pour les résultats de la classification multi-domaines, les performances de l'analyse de sentiment et de la classification de la polarité de deux modèles démontrent qu'une entrée multilingue est aussi efficace qu'une entrée monolingue.

Langues des corpus	Accuracy (%)
Anglais, Français et Grec	91,25
Anglais et Français	91,74
Français	92,76
Anglais	91,27

TABLE 3.8 – Résultats de l'analyse de sentiment et de la classification de la polarité avec le modèle ConvNets ; les performances sont satisfaisantes pour tous les corpus

Langues des corpus	Accuracy (%)
Anglais, Français et Grec	91,27
Anglais et Français	91,16
Français	92,68
Anglais	90,94

TABLE 3.9 – Résultats de l'analyse de sentiment et de la classification de la polarité avec le modèle LSTM ; les performances sont satisfaisantes pour tous les corpus

De plus, il n'y a pas de différence significative entre le modèle ConvNets et le modèle LSTM pour les entrées monolingues, bilingues ou multilingues. Néanmoins, les résultats de la classification multi-domaines sont supérieurs à ceux de l'analyse de sentiment, avec une différence d'environ 7% pour toutes les configurations des corpus. Cette différence peut indiquer que l'analyse de sentiment est une tâche plus complexe que la classification multi-domaines.

3.6.2 Indicateurs et rapports de classification

Les résultats présentés dans la section précédente montrent qu'un corpus multilingue réussit aussi bien qu'un corpus monolingue en analyse de sentiment, en termes de classification de la polarité et de classification multi-domaines. Nous proposons ici des métriques complémentaires afin de mieux démontrer l'efficacité des modèles proposés en analyse et classification multilingue.

Les trois mesures proposées sont la précision, le rappel et le F1-score. La précision représente la proportion d'avis correctement classés parmi l'ensemble des avis classés comme positifs ou négatifs, mesurant la capacité du modèle à identifier correctement la polarité. Le rappel mesure la capacité du système à identifier tous les avis pertinents de chaque catégorie (positif et négatif). Le F1-score est la moyenne harmonique de la précision et du rappel, offrant une mesure globale de la performance du modèle en tenant compte à la fois de la précision et du rappel.

PREDICTION			
Classes		Positif	Négatif
	Positif	**VP Vrais Positifs**	**FN Faux Négatifs Erreur Type II**
	Négatif	**FP Faux Positifs Erreur Type I**	**VN Vrais Négatifs**

TABLE 3.10 – Matrice de confusion, définition Erreur type I et II

Ces indicateurs sont calculés comme suit :

$$\text{Précision} = \frac{VP}{VP+FP} \tag{3.2}$$

Explication : La précision est le rapport des vrais positifs (VP) aux vrais positifs et faux positifs (FP).

$$\text{Rappel} = \frac{VP}{VP+FN} \tag{3.3}$$

Explication : Le rappel est le rapport des vrais positifs (VP) aux vrais positifs et faux négatifs (FN).

$$\text{F1-score} = 2 \times \frac{\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}} \tag{3.4}$$

Explication : Le F1-score est la moyenne harmonique de la précision et du rappel.

Les Tableaux 3.11 et 3.12 montrent les rapports de classification pour la classification multi-domaines et l'analyse des avis multilingue.

Domaine	**Nombre d'avis**	**Précision**	**Rappel**	**F1-score**
Restaurant	18 653	0,99	0,99	0,99
Hôtel	8 891	0,98	0,98	0,98

TABLE 3.11 – Rapport de classification du modèle ConvNet pour la classification multi-domaines, corpus multilingue

Le tableau 3.11 montre le résultat obtenu sur un échantillon composé de 27 544 avis multilingues, le nombre d'avis incorrectement classifiés est comme suit :

— 217 avis d'opinions FP,

— 218 avis d'opinions FN.

Le tableau 3.12 montre le résultat obtenu sur un échantillon composé de 27544 avis multilingues, le nombre des avis incorrectement classifiés est comme suit :

— 1 218 avis classés FP,

Label	**Nombre d'avis**	**Précision**	**Rappel**	**F1-score**
Négative	12 688	0,91	0,90	0,90
Positive	14 856	0,92	0,92	0,92

TABLE 3.12 – Rapport de classification du modèle ConvNet pour la classification d'opinions : corpus multilingue

Label	**Nombre d'avis**	**Précision**	**Rappel**	**F1-score**
Négatives	3 746	0,92	0,93	0,92
Positives	4 295	0,94	0,92	0,93

TABLE 3.13 – Rapport de classification avec le modèle convNet : Classification d'opinions corpus français

— 1 192 avis classés FN.

Le tableau 3.13 montre le résultat obtenu sur un échantillon composé de 26 804 avis en français et testés sur 8 041 avis, relevant un nombre d'avis incorrectement classifiée :

— 323 avis classés FP,
— 259 avis classés FN.

Domaine	**Nombre d'avis**	**Précision**	**Rappel**	**F1-score**
Restaurant	10 634	0,98	0,99	0,99
Hôtel	6 518	0,98	0,98	0,98

TABLE 3.14 – Rapport de classification du modèle ConvNet : Classification multi-domaines corpus anglais

Le tableau 3.14 montre le résultat obtenu sur un échantillon composé de 57 176 avis en anglais, testé sur 17 152 avis, relevant un nombre d'avis incorrectement classées est comme suit :

— 161 avis classés FP,
— 154 avis classés FN.

3.6.3 Classification d'opinions multilingues, incluant la langue arabe

Le corpus des avis sur des restaurants en langue arabe est composé de 10 970 avis, dont 2 675 sont négatifs. Afin de travailler avec un nombre d'avis équilibré entre le positif et le négatif, nous avons retenu 5 300 avis composés de 37 901 tokens uniques pour les tests de classification d'opinions en langue arabe. Ce corpus a ensuite été ajouté aux corpus en anglais, français et grec pour obtenir un corpus total de 102 460 avis avec 180 322 tokens uniques.

Corpus des langues	Accuracy
Arabe, Anglais, Français, et Grec	91.81
Arabe	86.79

TABLE 3.15 – Résultats de la classification d'opinions multilingues (arabe, anglais, français, grec) avec le modèle ConvNet

Avis	Nombre d'avis	Précision	Rappel	F1-score
Négatifs	4 840	0,93	0,90	0,91
Positifs	5 406	0,91	0,94	0,92

TABLE 3.16 – Rapport de classification d'un modèle ConvNet pour l'analyse d'opinions sur un corpus multilingue incluant la langue arabe

Le tableau 3.16 montre les résultats obtenus sur un échantillon composé de 102 460 avis en arabe, anglais, français et grec, testé sur 10% du corpus (soit 10 246 avis). Le nombre d'avis incorrectement classés est réparti comme suit :

— 488 avis classés comme de faux positifs (FP),
— 351 avis classés comme de faux négatifs (FN).

3.6.4 Classification multi-domaines en langues multiples, incluant la langue arabe

Dans le cadre de tests de classification multi-domaines en langue arabe, nous avons pris en compte les corpus relatifs aux hôtels et aux restaurants dans leur intégralité. Le corpus combiné des hôtels et restaurants comprend 24 127 avis avec 141 697 mots uniques. En y ajoutant les langues anglais, français et grec, le corpus multilingue totalise 118 358 avis et 244 892 mots uniques.

Le tableau 3.18 présente les résultats obtenus sur un échantillon composé de 118 358 avis en arabe, anglais, français et grec, testé sur 10% du corpus (soit 11 835 avis). Le nombre d'avis incorrectement classés se répartit comme suit :

Corpus des langues	Accuracy
Arabe, Anglais, Français, et Grec	98.87
Arabe	98.21

TABLE 3.17 – Résultats de la classification multi-domaines en langues multiples (arabe, anglais, français, grec) avec le modèle ConvNet

Domaine	Nombre d'avis	Précision	Rappel	F1-score
Restaurant	7 224	0,99	0,99	0,99
Hôtel	4 611	0,99	0,98	0,98

TABLE 3.18 – Rapport de classification d'un modèle ConvNet pour la classification multi-domaines sur un corpus multilingue incluant la langue arabe

— 58 avis classés comme de faux positifs (FP),
— 75 avis classés comme de faux négatifs (FN).

3.6.5 Comparaison avec IMDB

Pour évaluer et comparer la performance de nos modèles, nous avons utilisé le corpus IMDB pour la classification de sentiment [107]. Ce corpus contient 100 000 critiques de films, réparties en 25 000 critiques positives, 25 000 critiques négatives, et 50 000 non annotées. Nous présentons les résultats sur les données IMDB brutes et pré-entraînées par des vecteurs-mots GloVe de taille 300.

Modèles	Accuracy
LSTM[64]	89.1
Model Unlabeled + Bag of words (bnc)[107]	88.89
LSTM	87.06
LSTM - GloVe	89.57
ConvNet	89.16
ConvNet - GloVe	89.54
ConvNet-LSTM	89.18
ConvNet-LSTM - GloVe	89.90
Bi-ConvLSTM	89.48
Bi-ConvLSTM - GloVe	90.34

TABLE 3.19 – Résultats obtenus pour les données IMDB

Dans notre configuration du modèle hybride CNN-LSTM, nous examinons deux aspects de la polarité des critiques : l'orientation intrinsèque des mots (positive ou

négative) et le contexte dans lequel ils sont utilisés. Certains mots peuvent changer d'orientation en fonction du contexte, ou être affectés par des éléments syntaxiques distants, tels qu'une négation.
Pour notre évaluation, nous avons utilisé des vecteurs-mots GloVe pré-entraînés [138]. GloVe génère des vecteurs de mots de 300 dimensions basés sur des données issues de Wikipédia, capturant ainsi des relations sémantiques entre les mots.
Les résultats montrent que le modèle Bi-ConvLSTM pré-entraîné avec GloVe obtient les meilleures performances. L'amélioration est relativement faible, avec un $\triangle$ de seulement $0,5\%$.

Modèles	**Accuracy (%)**
Bi-ConvLSTM 300	94.86
Bi-ConvLSTM 50	93.08
Bi-ConvLSTM	94.90
ConvNet LSTM 300	94.41
ConvNet LSTM 50	93.43
ConvNet LSTM	94.74
LSTM 300	94.03
LSTM 50	92.80
LSTM	94.33
ConvNet 300	93.96
ConvNet 50	92.76
ConvNet	94.71

TABLE 3.20 – Performances sur le corpus français pour une classification binaire

Le tableau 3.20 présente les performances moyennes des différents modèles entraînés sur notre corpus en français. Les modèles ont été testés avec des vecteurs de mots word2vec de différentes dimensions (50 ou 300), ou sans ces vecteurs (mentionné comme "brut").
Afin d'obtenir une mesure plus fiable, chaque modèle a été entraîné plusieurs fois pour tenir compte de l'initialisation aléatoire des poids et des effets stochastiques introduits par la technique du dropout. Les modèles ont été entraînés sur un ensemble de 33 473 avis et validés sur 14 345 avis.

En général, toutes les architectures montrent une excellente performance, avec des taux d'accuracy supérieurs à 92%. Il est intéressant de noter que le modèle Bi-ConvLSTM a légèrement surperformé les autres architectures, notamment la ConvNet, avec une différence de 0,18%. En ce qui concerne les temps d'exécution, le modèle ConvNet simple a nécessité 3,6 minutes, tandis que les modèles Bi-ConvLSTM de taille 300^2 ont requis entre 14,33 et 25,3 minutes.

Les tableaux 3.21 et 3.22 sont des rapports de classification pour les modèles :

— CNN300, et

— Bi-ConvLSTM 300

évalués sur un ensemble de validation en français composé de 14 345 avis (7 156 positifs et 7 189 négatifs). Les métriques de précision, rappel, et F1-score confirment la robustesse de ces modèles.

CNN	**Précision précision**	**Rappel**	**F1-score**	**Nbr d'avis de test**
Négatif	0.93	0.95	0.94	7 189
Positif	0.95	0.93	0.94	7 156
	0.94	0.94	0.94	14 345

TABLE 3.21 – Rapport de classification : Précision, Rappel et F1-score pour le ModèleCNN300WV

	Opinions Positives	Opinions Négatives
Opinions Positives	6 853	336
Opinions Négatives	521	6 635

TABLE 3.22 – Matrice de confusion ModèleCNN300 pour 14 345 avis

Exemples d'avis mal qualifiés avec un modèleCNN(relevés sans correction) :
"bonne ambiance, bonne adresse, belle clientèle, mais nourriture tres moyenne. un endroit que l'on fréquente que pour l'ambiance pas pour le contenu de l'assiette...j'ai joué le jeu un soir, je n'y retournerai pas...mes amis non plus. je ne m'attendais pas à un gastro mais compte tenu des prix, je pensais que cela serait meilleur."
Cet avis a été annoté en "Négatif" et a été prédit en Positif.
"emplacement au top, à quelques pas de l'arc de triomphe, cuisine goûteuse et recherchée, on sent la maîtrise en cuisine, proportions correctes cependant, ce resto a des points à améliorer pour prétendre être gastro et digne de ses promesses. la déco n'est pas raffiné, l'ambiance n'est pas intimiste, je vous en supplie, arrêter le set de table en papier !...".
Cet avis a été annoté en "Positif" et a été prédit en Négatif.

Nous relevons dans le premier exemple de 'phrase mal qualifiée' que le modèle LSTM etCNNont traité cet avis de la même façon. Dans ce cas, le modèle LSTM

n'a pas mieux géré les dépendances longue distance qu'un modèle CNN. Un même constat est à porter pour d'autres résultats de type faux positifs et faux négatifs obtenus.

3.7 Conclusion

Nous avons évalué plusieurs architectures neuronales, notamment LSTM, ConvNet et leurs combinaisons, pour l'analyse de sentiment multilingue et multithématique. Les modèles hybrides LSTM-ConvNet ont été particulièrement performants, chaque architecture contribuant de manière unique au traitement des données. Les réseaux convolutionnels excellaient dans l'extraction de caractéristiques locales, tandis que les LSTM géraient efficacement les dépendances temporelles à long terme. Le Bi-ConvLSTM, qui intègre la bidirectionnalité, a montré une capacité accrue à contextualiser les données, même si l'amélioration en termes de performance était modeste. Une autre observation importante a été que nos modèles ont montré de robustes performances lorsqu'ils ont été formés sur des données textuelles brutes, sans prétraitement ni étiquetage. Ce résultat est particulièrement encourageant, compte tenu de la variabilité et des imprécisions souvent rencontrées dans les avis en ligne, y compris les fautes de frappe et les erreurs grammaticales.

Les réseaux LSTM et CNN, bien que supplantés dans certains domaines par les modèles de langage basés sur les Transformers, restent pertinents pour des applications spécifiques nécessitant des modèles plus compacts ou moins coûteux en ressources. Bien que les grands modèles de langage offrent une performance accrue grâce à leur capacité à modéliser des dépendances complexes et à généraliser à partir de grandes quantités de données, leur applicabilité peut être limitée par les exigences en matière de calcul et de stockage. Dans le contexte spécifique de l'analyse de sentiments multilingue pour la restauration et l'hôtellerie, les LSTM et CNN peuvent offrir un équilibre efficace entre précision et coût opérationnel. Envisager une intégration des caractéristiques apprises par ces réseaux dans des architectures plus récentes pourrait ouvrir la voie à des avancées hybrides, combinant le meilleur des deux approches pour une efficacité accrue dans des applications spécialisées.

Contributions
Ce projet apporte plusieurs contributions significatives au domaine de l'analyse de sentiment sur corpus multilingue et multithématique. D'abord, nous avons démontré la viabilité des architectures hybrides LSTM-ConvNet pour ce type d'analyse,

en particulier dans des langues morphologiquement complexes comme l'arabe. De plus, une évaluation comparative a été réalisée pour mettre en lumière l'impact de la bidirectionnalité sur la performance des modèles. Enfin, cette recherche enrichit la compréhension de l'importance relative des différents composants architecturaux et hyperparamètres, offrant ainsi des lignes directrices pour le développement futur dans ce domaine. L'approche proposée relève le défi d'utiliser des données textuelles brutes, sans aucun prétraitement ni connaissances préalables, et sans recourir à une langue pivot ou à des lexiques de mots liés aux sentiments. Cette flexibilité la rend extrêmement adaptable et applicable à un large éventail de sources de données et de langues.

Perspectives

En se basant sur la littérature actuelle, nous identifions plusieurs voies pour de futurs travaux. Étant donné les différents niveaux de granularité dans l'analyse de sentiment (niveau du document, de la phrase, et aspect), une avenue intéressante pourrait être d'explorer des modèles capables de traiter ces différentes échelles de manière intégrée. De plus, l'étude de l'analyse de sentiment au niveau de l'aspect, qui offre une analyse plus fine, pourrait révéler des nuances qui échappent aux modèles basés uniquement sur le niveau du document ou de la phrase. En outre, l'ajout de tâches associées, telles que la détection de sarcasme ou l'analyse émotionnelle, pourrait enrichir les fonctionnalités de nos modèles.

Chapitre 4

Annotation des aspects par apprentissage profond : transfert de connaissances et apprentissage actif

Sommaire

Dans le chapitre précédent, nous avons présenté des modèles hybrides de l'apprentissage profond, pour l'analyse de sentiment et la classification thématique en contexte multilingue, et nous avons démontré leur efficacité en soulignant une caractéristique clé de cette démarche : notre concentration sur des données multilingues non annotées, illustrant ainsi la puissance de notre méthode pour extraire des caractéristiques pertinentes sans nécessiter de prétraitement ni d'annotation.

Ce chapitre prolonge cette recherche en abordant une question complémentaire mais tout aussi essentielle : l'annotation des aspects sur corpus multilingues non annotés [1]. Alors que l'analyse de sentiment, même celle basée sur les aspects [2], se focalise sur l'évaluation globale d'un texte pour la classification de la polarité, l'annotation des aspects vise à cibler des éléments plus spécifiques, comme les attributs et les caractéristiques d'un produit ou d'un service. Tout comme dans notre travail précédent, nous continuons à privilégier des données multilingues non annotées, mais cette fois, nous nous engageons dans la quête de méthodes pour les enrichir et les transformer en un ensemble de données plus structuré et exploitable.

Nous introduisons une méthode hybride qui combine l'apprentissage par transfert [3] [126] avec l'apprentissage actif [4] [**?**] afin de réaliser l'annotation des aspects sur des corpus textuels multilingues et sans aucun traitement préalable. L'annotation des aspects est essentielle pour l'élaboration de modèles de traitement du langage naturel plus avancés et plus fiables. Nous cherchons à optimiser l'efficacité des techniques d'apprentissage profond dans l'annotation de corpus, surtout pour les langues qui manquent de ressources annotées, pour affiner ou améliorer les performances des modèles linguistiques en TALN. Le chapitre est composé des parties suivantes :

- Introduction 4.1 : Un aperçu de l'importance et des défis liés à l'annotation des aspects dans les corpus textuels.
- Problématique 4.2 : Exploration des questions sous-jacentes relatives à la qualité, l'efficacité et la transférabilité de l'annotation d'aspects.
- État de l'art 4.3 : Revue des méthodes existantes d'annotation et d'identification des aspects, y compris les limites des approches traditionnelles.
- Identification des aspects à partir de corpus non labellisés 4.4 : Discussion sur notre approche de pré-labelisation par transfert de connaissances et l'amélioration des pré-labels par apprentissage actif.
- Expérimentation et évaluation des performances 4.5 : Description des données utilisées pour l'évaluation, des métriques choisies et des résultats obtenus à chaque étape de l'annotation.
- Conclusion 4.6 : Synthèse des contributions, implications pour les futures

1. Ces travaux de recherche font partie de la thèse cifre (contrat avec Novagen `https://www.novagen.tech/`) de Mme Maroua Boudabous sous ma direction (en cours).

2. Aspect Based Sentiment Analysis (Aspect Based Sentiment Analysis (ABSA)), nous en détaillons plus dans la section "état de l'art" 4.3

3. Dans le contexte de notre recherche l'apprentissage par transfert signifie que nous utilisons des modèles pré-entraînés, ou des caractéristiques extraites d'un ensemble de données, pour faciliter l'annotation d'aspects dans un nouvel ensemble de données multilingue.

4. Dans notre recherche cette méthode choisit spécifiquement des exemples dans les données multilingues où l'annotation d'aspect est difficile ou ambiguë, et utilise ces exemples pour améliorer le modèle grâce à des cycles d'annotation et de réentraînement.

recherches, et discussion sur l'apport de notre approche peut dans le vaste domaine de l'annotation de corpus et de l'ADT.

En intégrant les avancées du chapitre précédent sur l'analyse de sentiment via l'apprentissage profond, avec les contributions actuelles sur l'annotation des aspects, nous formulons ici un cadre unifié pour une exploitation plus complète des données textuelles multilingues.

4.1 Introduction

Dans le domaine de l'intelligence artificielle appliquée à la linguistique, l'annotation de corpus [5] est souvent considérée comme un élément clé pour l'analyse sémantique et pragmatique de textes. L'annotation permet non seulement d'enrichir un corpus avec des métadonnées utiles, mais elle sert également de tremplin pour des méthodes plus avancées d'Analyse des Données Textuelles (ADT), y compris des calculs de cooccurrence et des analyses factorielles.

Au cours de la dernière décennie, l'annotation des corpus linguistiques a connu une évolution significative, passant d'une approche centrée principalement sur la structure et la grammaire à des niveaux d'analyse beaucoup plus complexes, notamment sémantiques et pragmatiques. Il y a une vingtaine d'années, dans [118] l'annotation était décrite en trois catégories principales : l'annotation (ou "marquage" selon l'anglicisme "markup") structurelle, l'annotation grammaticale, et l'annotation en parties du discours ("PoS tagging"). Ces types d'annotation fournissaient des bases solides pour la recherche en linguistique [6], mais ils étaient limités dans leur capacité à capturer des nuances plus subtiles telles que le sens, les sentiments, et les aspects d'un texte.

[93] décrit, outre les premiers outils et méthodes pour annoter des corpus, la nécessité d'annotation pour extraire des informations des corpus en ajoutant des étiquettes contenant des informations explicites sur le contenu. [144] expliquent également l'importance de l'annotation pour approfondir le sens et lever les ambiguïtés entre autres, en détaillant la méthodologie employée pour annoter et par conséquent enrichir les données linguistiques ("datasets" pour des corpus annotés) afin de les utiliser pour les besoins de l'apprentissage automatique (Machine

5. Les termes annotation/balisage/étiquetage/labellisation sont interchangeables, utilisés en IA et en apprentissage automatique. Ces termes servent à élaborer des jeux de données destinés à des tâches de TALN. Le texte est annoté, ou étiqueté, dans l'objectif de clarifier les termes clés ou les mots significatifs pour les algorithmes, facilitant ainsi une réponse cohérente de la part des machines

6. Souvent l'annotation était manuelle afin d'annoter tous les phénomènes appartenant à la linguistique textuelle ne pouvant pas être annotés de manière automatique : les coréférents textuels, les connecteurs et marqueurs logico-temporels et discursifs, etc.

Learning). Annoter un texte est un moyen de capturer les informations qui sont précieuses aux algorithmes d'apprentissage automatique, pour améliorer leur capacité à comprendre et à analyser le langage humain.
Avec l'avènement des techniques d'apprentissage profond, le domaine a vu une transition vers des formes d'annotation plus sophistiquées. Par exemple, des travaux récents dans le domaine de la désambiguïsation du sens des mots [173] ont souligné l'importance d'ensembles de données annotés en "sens" pour former des systèmes supervisés performants. De même, l'annotation des aspects dans les avis ou l'analyse de sentiment va bien au-delà des capacités des approches d'annotation traditionnelles , abordant des dimensions qui étaient auparavant difficiles à quantifier ou à catégoriser de manière fiable. Alors que les algorithmes deviennent plus sophistiqués, les besoins pour des annotations plus fines et plus complexes augmentent également. Malgré l'existence de plusieurs datasets disponibles en ligne [7], il est nécessaire d'en créer de nouveaux, pour compléter le manque de données spécifiques, ou le manque d'annotation, permettant d'explorer de nouvelles facettes des données.

Il y a plusieurs travaux décrivant méthodes et outils [8] pour différent types d'annotation comme l'annotation linguistique des corpus [54] ou l'annotation sémantique textuelle [100]. L'annotation des aspects est une tâche sémantique fine, souvent orientée vers l'analyse de sentiment [198]. Plusieurs de ces travaux pour l'annotation des sentiments, ont été évalués dans [178]. Nous détaillerons plus le contexte de la recherche sur l'annotation ainsi que les méthodes proposées dans la section état de l'art 4.3.

Dans ce travail, nous nous intéressons à l'identification et annotation des aspects, qui caractérisent des particularités des produits et/ou des services, décrites dans les avis des utilisateurs. L'identification des aspects permet la reconnaissance des termes utilisés pour désigner ces attributs et caractéristiques. Dans la littérature nous distinguons deux types d'aspects : explicites et implicites. L'aspect est dit explicite quand la cible de l'opinion exprimée peut être associée à un mot (ou un ensemble de mots) dans l'avis utilisateur. L'aspect est qualifié d'implicite si le trait cible de l'opinion est indirectement déduit de la sémantique de l'avis.

Ci-dessous un exemple d'avis illustrant la différence entre les deux types d'aspects.

7. Voir par exemple `https://www.kaggle.com/datasets?tags=13204-NLP`, qui propose des datasets gratuits et payants pour les tâches de machine learning en NLP.

8. A titre d'exemple quelques outils d'annotation open source : pour les pages web `http://annotatorjs.org/`, pour différentes tâches ADT dont analyse de sentiment `https://github.com/doccano/doccano`, ou différentes librairies comme spacy `https://spacy.io/`

« Design top ! J'adore les nouvelles couleurs, je l'ai en mauve !»
→ *Aspect explicite* : couleur, *Valeur* : mauve

« Je préfère le dernier modèle, il est génial je peux le mettre dans ma poche ! »
→ *Aspect implicite* : taille, *Valeur* : petite

FIGURE 4.1 – Exemples illustrant les notions "explicite", "implicite"

Dans ce travail, nous nous intéressons à l'identification des aspects explicites. Cette tâche généralement considérée comme un problème d'étiquetage séquentiel supervisé. Pour former des modèles d'apprentissage pour cette tâche, un corpus de données annotées est indispensable. Néanmoins, la collecte de telles données est un processus à la fois long et exigeant en termes d'expertise humaine [9].
Ce contexte nous a incités à développer une méthode d'identification des aspects qui fonctionne avec des corpus initialement non annotés.
Dans le but de faciliter la compréhension et l'interprétation de ce travail, il est important de préciser l'emploi de plusieurs termes liés à l'annotation des données, fréquemment utilisés en TALN.
Bien que les termes "annotation", "labellisation", et "étiquetage" puissent souvent sembler interchangeables, nous proposons dans le contexte de cette étude une distinction plus nuancée entre ces concepts. Cette distinction est fondée sur les descriptions faites dans [50, 193], ainsi que sur les pratiques empiriques observées dans nos travaux.
Annotation est utilisé pour décrire le processus général par lequel des métadonnées ou des informations supplémentaires sont ajoutées aux données textuelles. L'annotation peut englober diverses activités telles que l'attribution de catégories, l'identification d'entités, et même l'ajout de commentaires explicatifs.
Labellisation est un terme plus spécifique principalement utilisé pour décrire le processus d'attribution d'étiquettes, des "labels" à des unités discrètes du texte. Cette action n'implique pas nécessairement un contexte spécifique (souvent il s'agit de contexte séquentiel). Il est souvent employé dans des scénarios de classification.
Étiquetage est utilisé pour décrire l'annotation des données en prenant en compte un contexte ou une séquence. Il est souvent employé en TALN pour des tâches comme l'étiquetage morphosyntaxique ou le "Part-of-Speech" (PoS) tagging.
Nous soulignons cette nuance terminologique car l'identification des aspects dans

9. Nous parlons ici de langues pour lesquelles il existe un manque de données annotées en ce qui concerne la reconnaissance des aspects. Dans ce contexte, le français est également concerné. Bien que certains corpus annotés en aspects existent pour l'analyse de sentiment, ces corpus se limitent souvent à indiquer le sentiment associé à un aspect donné, sans fournir d'information sur les caractéristiques spécifiques de cet aspect.

ce travail peut impliquer une attention particulière au contexte et à la séquence, ainsi que son classement dans une catégorie donnée, soit de manière implicite, soit explicite, dans le cadre du processus d'annotation.

Notre approche se déroule en trois phases. La première utilise les champs aléatoires conditionnels (CRF) pour une "pré-labellisation par transfert" permettant une annotation préliminaire des données "cible" 4.4.1. La seconde phase utilise une stratégie de sélection basée sur un indice de confiance pour corriger et améliorer la labellisation initiale, et la troisième utilise l'apprentissage actif pour traiter l'incertitude des étiquettes, affiner les labellisations déjà effectuées et en proposer des nouvelles.
Ces étapes facilitent l'application de modèles supervisés aux langues à faibles ressources et proposent des approches originales pour la gestion du déséquilibre de fréquence entre termes "aspects" et "non-aspects".
Le modèle proposé, génère des données que l'expert vérifie, corrige et enrichit, contribuant ainsi à une amélioration continue de la performance du modèle d'identification des aspects.

4.2 Problématique

L'annotation de corpus [10] demeure une base essentielle de la recherche en TALN, contribuant continuellement à l'avancement des modèles de langage et des algorithmes d'analyse. Cette démarche est particulièrement importante pour l'extraction d'informations, d'où la nécessité de l'encodage préalable d'analyses linguistiques dans le corpus.
L'univers des annotations s'est significativement diversifié, englobant non seulement les étiquetages structurels et morphosyntaxiques, mais s'étendant également à des niveaux plus complexes de l'analyse du langage comme l'annotation sémantique et pragmatique, d'aspect et de relation ou d'entité. Comment garantir alors une qualité d'annotation à la hauteur des besoins spécifiques des recherches en TALN ? Quels outils et méthodes s'offrent à nous pour naviguer à travers cette complexité ?
L'importance de créer continuellement de nouveaux ensembles de données annotées est primordiale, particulièrement dans le contexte multilingue et spécialisé

10. Cette recherche fait partie d'un contrat cifre. L'annotation de texte est un enjeu important pour le domaine de la data science puisqu'elle permet de fournir des données exploitables et analysables. Les données peuvent contenir des informations spécifiques, être dans des langues peu utilisées ou encore être sous un format très particulier. Face à cette singularité, il est souvent nécessaire de labelliser son propre jeu de données. De plus, selon les moyens financiers et temporels de chaque contexte, toutes les solutions d'annotation ne sont pas pertinentes.

où la rareté de données bien annotées constitue un défi majeur. Les annotations jouent le rôle de métadonnées qui, en s'ajoutant à un texte, facilitent des utilisations ultérieures variées, allant de l'entraînement de modèles d'apprentissage automatique à des analyses plus pointues. L'adaptabilité de ces modèles d'annotation à des langues ou des domaines moins dotés en ressources représente un autre défi significatif.
En réponse à ces enjeux, la présente recherche propose une méthode hybride qui combine l'apprentissage par transfert et l'apprentissage actif pour l'annotation d'aspects relatifs aux caractéristiques de produits ou de services dans des corpus multilingues. Nous visons ainsi à enrichir les ensembles de données avec des annotations d'aspects plus précises et plus utiles pour les futurs travaux en ML.
C'est là qu'interviennent les différentes approches et techniques de l'annotation de jeux de données, que nous détaillerons dans la section des travaux connexes. Ces techniques cherchent à atténuer le problème de la rareté des données et à optimiser l'utilisation des données disponibles.
Afin de mieux cerner les enjeux et les objectifs de cette recherche, il est utile de formaliser le problème d'annotation séquentielle comme suit :

Définition du Problème

Étant donné un vocabulaire d'entrée fixe de mots σ, un ensemble de sortie de K étiquettes T et une séquence d'entrée $x(x_1, x_2, ..., x_M)$ composée de M mots de σ, dans la tâche d'étiquetage séquentiel, notre objectif est d'apprendre une fonction $h : \sigma \rightarrow T$ qui attribue une étiquette de T à chaque élément de la séquence d'entrée x. La fonction h *mappe* la séquence d'entrée à la séquence d'étiquettes la plus probable $t^* = (t_1^*, ..., t_M^*)$ où $t \in K$. C'est-à-dire, nous cherchons à trouver :

$$t^* = \arg\max_{t \in T} P(t|x) \tag{4.1}$$

Dans un cadre supervisé, la fonction de mappage h est apprise en entraînant un modèle d'apprentissage sur un ensemble de données de N séquences étiquetées sous la forme de paires de séquences $(x^{(i)}, t^{(i)})$ pour $i \in \{1, ..., N\}$.
Nous considérons l'extraction de termes d'aspect (ATE) comme un problème d'étiquetage séquentiel où chaque jeton (mot) est étiqueté comme un aspect ou non. Notez que nous considérons à la fois les aspects en plusieurs mots et en un seul mot et utilisons le format d'étiquetage IOB2 [147] à cet effet. Dans ce format, trois étiquettes sont définies : "I" indique que le jeton est à l'intérieur de l'expression d'aspect, "O" signifie que le jeton n'est pas un aspect et "B" exprime que le jeton est le début d'une expression d'aspect.

4.3 État de l'art

L'identification des aspects, les termes qui font référence aux attributs et aux caractéristiques d'un produit ou d'un service, constitue une étape essentielle dans l'analyse de sentiment et la fouille d'opinions. Cette tâche est d'autant plus complexe pour les langues avec des ressources limitées en données labellisées. Elle s'inscrit dans le cadre plus large de l'annotation de jeux de données.
Les méthodes d'annotation varient en degré d'automatisation, allant de l'annotation purement manuelle à des approches hautement automatisées.
Ces méthodes présentent des écarts significatifs en termes de temps et de ressources nécessaires. Par ailleurs, le niveau d'implication humaine souhaité dans le processus d'annotation n'est pas anodin et mérite donc de s'y arrêter.

Dans certains cas, les chercheurs optent pour la création de jeux de données spécifiques à leur travail de recherche en exploitant des outils automatisés. Par exemple, des API proposées de réseaux sociaux comme Twitter sont souvent utilisées pour collecter des données en temps réel, sans avoir besoin de recourir à des plateformes de *crowdsourcing* [11].
Ce type de méthode permet une plus grande maîtrise sur la nature et la qualité des données recueillies, même si elle n'élimine pas totalement la nécessité d'une vérification manuelle pour assurer leur pertinence et leur fiabilité. Un exemple en est l'étude menée sur la détection du discours haineux aux tweets en langue arabe par [187], où les données ont été collectées via l'API de Twitter. De cette manière, la rigueur dans l'annotation est garantie.
Cependant, ce type d'approche peut s'avérer très coûteux et chronophage pour des jeux de données de grande envergure, d'où l'émergence de plateformes de *crowdsourcing* [12] L'efficacité et l'éthique de ces plateformes restent toutefois un sujet de débat, comme souligné par [163].

Nous nous intéresserons particulièrement aux méthodes automatiques d'identification et de labellisation des données textuelles. Deux types d'approches dominent dans la littérature :

- les approches supervisées, qui consistent à labelliser les textes. Un exemple classique est la reconnaissance *spam/non spam* qui nécessite en amont une qualification humaine, et
- les approches non-supervisées, qui consistent à fournir des données sans

11. Les plateformes de crowdsourcing sont des plateformes en lignes sur lesquelles on peut faire appel à des travailleurs appelés crowdworkers afin qu'ils réalisent une tâche, en leur soumettant un jeu de données et les labels à assigner, de manière rémunérée.

12. A titre d'exemple nous citons le service de micro-travail lancé par Amazon.com fin 2005 `https://www.mturk.com/` qui est parmi les plus connus et les plus critiqués [48].

labellisation pour des tâches de regroupement (clustering), classification ou de détection d'anomalies.

Parmi les méthodes non supervisées, nous citons les travaux de [145], qui utilisent conjointement la double propagation et les relations syntaxiques entre les attributs et les noms. Cette approche ne nécessite pas de données labellisées. Cependant, elle manque de précision en raison de la complexité de définir toutes les règles nécessaires pour décrire les relations syntaxiques entre les termes. Ces règles sont complexes et diffèrent d'une langue à une autre.

Une autre technique statistique non supervisée pour l'analyse de sentiment modélise explicitement la relation entre les aspects et les phrases subjectives [151], comme un problème d'inférence conjointe et comparée à une architecture de pipeline pure. La question est dans quelle mesure un modèle conjoint surpasse un modèle de pipeline en termes d'extraction d'aspects, de phrases subjectives et de la relation entre eux. Les résultats montrent une légère amélioration de la prédiction des cibles à l'aide d'un modèle d'inférence conjointe, mais les performances de détection de phrases subjectives et d'extraction de relations diminuent.

Les méthodes, qui utilisent un système de règles de gestion, identifient les caractéristiques linguistiques et grammaticales des aspects et les relations syntaxiques entre eux. Le travail de [66] est l'un des travaux pionniers dans cette catégorie. Il considère les noms et les phrases nominales fréquents comme des aspects.

Des améliorations de ces travaux ont été proposées par [104] qui emploient la sémantique pour enrichir les données initiales.

L'inconvénient de ces approches est le manque de précision nécessitant de changer les règles quand le domaine d'application change. De plus, elles dépendent fortement de l'expertise humaine.

Pour d'autres approches, l'utilisation d'apprentissage statistique supervisé a suscité de l'intérêt pour l'identification des aspects en définissant cette tâche sous forme d'un problème de labellisation séquentielle, comme les travaux de [68], qui ont adopté le modèle de champs aléatoires conditionnels (CRF).

Différentes architectures de réseaux profonds ont été utilisées pour résoudre le problème d'identification des aspects à savoir les réseaux récurrents de type LSTM [103], le mécanisme d'attention [98], les réseaux de convolutionCNN[194] et les transformateurs [154]. Les performances de toutes ces architectures dépendent de la disponibilité de données d'apprentissage labellisées, ce qui rend leur utilisation problématique pour les langues à faibles ressources.

La rareté des données labellisées pour entraîner les modèles d'apprentissage supervisé a été considérée par [97] qui ont eu recours à l'augmentation des données. Ils génèrent des données supplémentaires à travers un modèle de génération masqué conditionnel appliqué sur les données d'apprentissage. Cette méthode est assez incontrôlable car elle risque de changer la sémantique des opinions générées. [26] ont utilisé des prototypes logiciels appris sur les données externes générées

par des modèles de langage pré-entraînés pour pallier ce problème.

Avant de présenter les approches qui nous ont inspirés pour notre méthode, nous citons à titre d'exemple un des outils commercialisés, qui effectue l'annotation entre d'autres tâches. Prodigy [1] est un outil d'annotation qui assure l'entraînement et l'évaluation de modèles pour certaines tâches de TALN, comme la reconnaissance des entités nommées, la classification des textes et l'annotation morpho-syntaxique. Il emploie l'apprentissage actif pour limiter les coûts d'annotation manuelle et profiter de l'interactivité avec l'utilisateur dans le but de faciliter à la fois l'annotation et l'apprentissage. Cependant, il s'agit d'un outil commercial pas disponible en open-source.
Notons également l'émergence des Modèles de Langage à Grande Échelle (Large Language Model (LLM)) comme ChatGPT-4 [124] ou Llama [177] comme des outils puissants pour l'annotation de textes [13]. [174] a constaté que l'utilisation de ChatGPT-4 surpassait les experts et les crowdworkers en termes de précision, de fiabilité et de biais. Cette observation est renforcée par [53], qui ont mis en évidence la flexibilité et le faible coût de ChatGPT pour l'annotation en comparaison avec des méthodes humaines traditionnelles, comme le constatent aussi [125], qui ont automatisé l'étiquetage des scénarios en utilisant des modèles linguistiques ML. L'intégration de LLM est une avancée récente, dont l'application et l'efficacité continueront à être étudiées et affinées dans les années à venir, notamment en comparant leur efficacité (celle des LLM) à celle des ensembles d'annotations d'experts et en explorant leur rôle en tant que oracles dans des scénarios d'apprentissage actif.
Pour notre recherche deux approches de l'apprentissage profond ont attiré notre attention : l'apprentissage par transfert et l'apprentissage actif.

L'apprentissage par transfert consiste à utiliser des connaissances provenant d'une ou de plusieurs tâches sources pour améliorer les performances sur une tâche cible, et constitue un domaine de recherche actif en apprentissage automatique. Il améliore l'efficacité de l'apprentissage en réutilisant des modèles pré-entraînés.
[126] considèrent que l'apprentissage par transfert offre une assistance automatisée aux annotateurs dans des contextes exploratoires [14], en particulier pour les

13. Notons que l'efficacité de ces modèles peut être influencée par la langue utilisée. Les LLM sont principalement entraînés sur des données en anglais, ce qui peut poser des problèmes pour les applications dans d'autres langues, comme le français, où la traduction de mauvaise qualité peut affecter les performances. Néanmoins, cette limitation est progressivement surmontée par l'émergence de jeux de données multilingues comme OSCAR `https://oscar-project.org/`, remettant ainsi en question la prédominance de la langue anglaise.

14. Les contextes exploratoires se réfèrent à des environnements de recherche où les méthodologies, les jeux de données ou les ensembles d'étiquettes ne sont pas encore bien définis, nécessitant

langues sous-documentées et, démontrent à travers des simulations que même des méthodes de transfert simples peuvent améliorer significativement la qualité des pré-annotations.
Plus récemment, [143] abordent le défi de la cartographie d'annotations, passant de descriptions génériques de processus au niveau de la langue et du domaine à des données de processus spécifiques à une implémentation, en utilisant une approche basée sur l'apprentissage par transfert.

L'Active Learning (AA) [157] représente une alternative pour améliorer la performance des modèles d'apprentissage lorsque l'annotation manuelle devient trop coûteuse. L'apprentissage actif est un processus itératif qui sélectionne un échantillon des données non labellisées selon une stratégie de sélection prédéfinie et invite un expert à les labelliser.
Étant donné que des modèles de pré-annotation précis sont très efficaces pour réduire l'effort d'annotation, de nombreux travaux ont été réalisés pour entraîner des modèles de haute qualité avec le moins de données possible. La littérature sur l'apprentissage actif cherche à réduire le coût nécessaire pour entraîner un modèle en sélectionnant des instances pour l'annotation qui sont susceptibles d'être les plus informatives [158]. Les techniques faiblement supervisées tentent d'accélérer l'entraînement du modèle en apprenant à partir d'instances non étiquetées, ou en permettant aux annotateurs de communiquer leur savoir au modèle en spécifiant des étiquettes ou des contraintes applicables à de grandes classes d'instances de données [152, 39, 99, 49].
Dans le domaine de TALN, AA a été considéré pour différentes tâches de labellisation séquentielle à savoir la reconnaissance des entités nommées (NER) [161] et l'étiquetage morpho-syntaxique (PoS) [150]. Dans ce contexte, l'AA a été, dans un premier temps, utilisé avec des modèles d'apprentissage automatique standards [158]. Ensuite, l'AA a été associé à des modèles d'apprentissage profonds. Cette association a révélé des résultats intéressants en termes de coût de calcul et de labellisation [155, 149, 162].

Dans le cadre de notre recherche, nous adaptons les deux techniques pour annoter nos corpus multilingues : l'apprentissage par transfert et l'apprentissage actif. Premièrement, nous utilisons l'apprentissage par transfert pour effectuer une pré-annotation automatique des données. Deuxièmement, nous mettons en œuvre un processus d'apprentissage actif pour enrichir ces données pré-labellisées. Dans cette étape, un expert intervient pour vérifier et, si nécessaire, corriger l'exactitude de la labellisation des données. Ce processus permet non seulement d'augmenter la quantité de données labellisées mais aussi d'améliorer la précision des annotations.

une adaptation et une évolution constantes pour parvenir à des résultats fiables.

4.4 Identification des aspects à partir de corpus non labellisés

Dans cette section, nous abordons la problématique liée au manque de données labellisées pour l'identification d'aspects explicites dans les avis d'utilisateurs. Plus spécifiquement, nous nous intéressons à un cas extrême où le domaine d'application cible ne dispose d'aucune donnée labellisée.
Pour relever ce défi, notre approche commence par une étape de pré-étiquetage via l'apprentissage par transfert. Cette étape initiale fournit une base solide pour l'entraînement ultérieur du modèle, lequel sera ensuite affiné à l'aide de techniques d'apprentissage actif.
Nous avons choisi le français comme une langue représentative des langues faiblement dotées en ressources [15], puisqu'il existe peu de jeux de données labellisées pour la tâche d'identification des aspects dans cette langue.

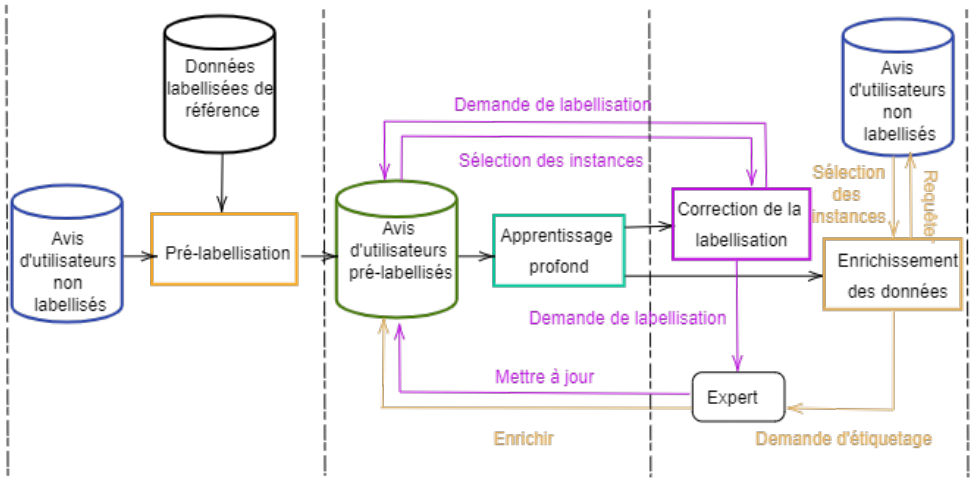


FIGURE 4.2 – Schéma global du processus proposé pour l'identification des aspects explicites

La figure 4.2 illustre l'architecture globale de notre processus en trois étapes. Tout d'abord, nous étiquetons les données de manière pseudo-étiquetage, c'est la "pré-labellisation par transfert". Durant cette phase, nous utilisons un modèle de Champs Aléatoires Conditionnels (CRF) [88] pour transférer des connaissances d'un domaine à l'autre (nous détaillons ce type de transfert dans la section suivante

15. Nous discutons du seul jeu de données en français doté de labels d'aspects que nous utilisons pour l'évaluation de notre méthode dans la section **??**

4.4.1), ce qui nous permet d'obtenir des annotations préliminaires sur les données cibles.
Durant la deuxième étape, nous entraînons un modèle d'apprentissage profond pour l'identification de terme d'aspect sur les données précédemment étiquetées dans un contexte supervisé.
Dans la dernière étape, nous utilisons l'apprentissage actif pour traiter l'incertitude des étiquettes, pour affiner les labellisations déjà effectuées. Le modèle génère des données que l'expert vérifie, corrige et enrichit, contribuant ainsi à une amélioration continue de la performance du modèle d'identification des aspects.

4.4.1 Pré-labellisation des aspects par transfert de connaissances

L'apprentissage par transfert de connaissances est une méthode qui consiste à utiliser les connaissances acquises sur une tâche spécifique pour les transférer à une autre [126]. En transférant les connaissances d'un jeu de textes source déjà labellisé à un second, que nous appelons cible, nous pouvons ainsi l'annoter sans effort humain. Cette technique est particulièrement utile lorsque la quantité de données disponibles pour une nouvelle tâche est limitée.
L'apprentissage par transfert est utilisé avec un jeu de données source déjà labellisé et similaire (parfois différent) à notre jeu cible afin de pré-entraîner un modèle. Il est ensuite possible en le réglant - c'est-à-dire en ajustant les poids, les paramètres et la couche de sortie - de l'utiliser pour prédire les classes de notre jeu cible et ainsi former notre jeu labellisé. Certains modèles déjà pré-entrainés existent et sont mis à disposition tels que Bert [35] qui a déjà été un support pour effectuer des transferts de connaissances. [122] présente un fine-tuning du modèle Bert en détaillant les modifications sur les différentes couches du modèle qui a pour but la détection de discours haineux.

[202] classifie ces techniques d'apprentissage par transfert de connaissances en deux catégories : le *l'apprentissage par transfert homogène* et le *l'apprentissage par transfert hétérogène*. Le transfert homogène s'appuie sur un transfert de connaissances à partir d'un jeu de données similaire au notre donc avec un même espace de caractéristiques. La méthode d'apprentissage par transfert homogène s'apparente à l'apprentissage semi-supervisé mais il est à noter que les données annotées et non-annotées proviennent obligatoirement de la même source.

L'apprentissage par transfert hétérogène détermine l'utilisation d'un jeu de données différent au nôtre, ayant par conséquent un espace de caractéristiques (appelé *feature space*) différent. [126] mettent en garde contre le "transfert négatif",

où le transfert de connaissances pourrait en fait nuire aux performances sur le jeu de données cible. Cette dégradation pourrait survenir lorsque les jeux de données source et cible sont trop différents.

Dans la cas de notre étude le transfert s'effectue à partir des labels du corpus restaurants vers le corpus des produits de beauté.
En effet, les deux corpus ont des domaines d'application très différents et sont susceptibles d'avoir des espaces de caractéristiques différents, notamment en ce qui concerne les aspects discutés, par exemple, *la qualité* de la nourriture dans le cas des restaurants et *l'efficacité* d'un produit de beauté dans l'autre cas.
Notons que l'aspect de "qualité" peut être commun aux deux domaines (restaurants et produits de beauté), mais la façon dont cette "qualité" est évaluée et discutée pourrait varier considérablement. Par exemple, la qualité dans le contexte d'un restaurant peut englober des facteurs comme le goût, la présentation, et la fraîcheur des ingrédients. En revanche, dans le contexte des produits de beauté, la qualité pourrait être évaluée en termes d'efficacité, de durabilité, ou même de composition chimique.
Ce genre de nuances fait que même des aspects qui semblent "communs" aux deux domaines peuvent en fait avoir des espaces de caractéristiques différents, ce qui justifie que dans le cadre de notre recherche, l'approche est l'apprentissage par transfert hétérogène. Il faut noter également que certains aspects, comme "l'efficacité," ne sont pas directement transposables entre les deux domaines. Dans ce genre de transfert hétérogène, l'un des défis majeurs est de trouver un moyen de "traduire" ou de "*mapper*" les caractéristiques d'un domaine à l'autre. Ceci peut être complexe, surtout si les caractéristiques ou les aspects à annoter varient considérablement entre les deux domaines.

L'utilisation d'un modèle CRF (Conditional Random Field) peut être intéressante dans ce cas, surtout si nous pouvons adapter les caractéristiques du modèle pour qu'elles soient pertinentes dans les deux domaines. Nous devons préalablement effectuer un ajustement fin du modèle sur le nouveau corpus pour obtenir des résultats optimaux.
En somme, le transfert de connaissances est une approche potentiellement puissante, mais elle comporte des risques et des défis. La sélection d'un jeu de données source approprié est donc essentielle, et dans certains cas, il peut être nécessaire d'annoter une partie du jeu de données cible pour l'utiliser comme source.

La phase de pré-labellisation sert de préliminaire pour l'annotation de données non labellisées. Pour cette étape, nous utilisons les Champs Aléatoires Conditionnels (CRF), un modèle de prédiction probabiliste adapté à la segmentation et à la

labellisation de données séquentielles comme le texte. Le CRF prédit le label de chaque terme en fonction de son contexte dans la séquence. Cette architecture permet d'optimiser la probabilité globale des étiquettes attribuées, en tenant compte des dépendances intrinsèques entre les termes.

Formellement, on définit une donnée séquentielle en entrée par $t = t_1, , t_n$ où t_i est le vecteur de caractéristiques du i^{eme} mot et par $l = l_1, , l_n$ la séquence de labels correspondants à la donnée en entrée.
On dénote par $\lambda(t)$ l'ensemble des séquences de labels possibles pour t.

Le modèle probabiliste de CRF définit un ensemble de probabilités conditionnelles $p(l|t;W,b)$ par rapport à toutes les séquences de labels possibles l étant donné t comme suit :

$$p(l|t;W,b) = \frac{\prod_{i=1}^{n} \psi_i(l_{i-1}, l_i, t)}{\sum_{l' \in \lambda(t)} \prod_{i=1}^{n} \psi_i(l'_{i-1}, l_i, t)} \tag{4.2}$$

où $\psi_i(l', l, t) = exp(W^T_{l',l} t_i + b_{l',l})$ sont des fonctions potentielles, $W^T_{l',l}$ et $b_{l',l}$ sont respectivement les vecteurs de poids et de biais correspondant au couple de séquences de labels (l', l).

Concrètement, les données en entrée pour le CRF sont des vecteurs de caractéristiques qui représentent différentes informations à propos du terme à savoir sa valeur littérale, le nombre de lettres, s'il s'agit d'un terme en majuscule ou minuscule, s'il s'agit d'un signe de ponctuation ou d'un chiffre. Nous avons inclus l'étiquette morpho-syntaxique (PoS) et nous considérons également les mêmes informations pour les six mots voisins.

Le modèle CRF (Champ Aléatoire Conditionnel) a besoin d'un ensemble initial de données labellisées pour son apprentissage. C'est là que le transfert de connaissances entre domaines intervient. Nous entraînons notre modèle CRF sur un ensemble de données labellisées d'un domaine source D, qui peut être plus facilement accessible ou déjà existant. Une fois que le modèle est bien entraîné sur ce domaine, nous l'appliquons ensuite au domaine cible D'.
Pour renforcer l'indépendance de notre modèle CRF par rapport au domaine d'application, nous avons omis la caractéristique "valeur littérale" des mots dont l'étiquette morpho-syntaxique correspond aux catégories noms ("Noun"), verbes ("Verb") et les propositions nominales ("PropN"), puisqu'il s'agit de catégories les plus dépendantes du domaine.

Dans notre cas spécifique, nous avons utilisé les seules données disponibles en français pour l'identification des aspects explicites (Jeu de données sur les restaurants de SemEval 2016) pour entraîner notre modèle CRF. Ces données jouent le rôle du domaine source D, et notre domaine cible D'. Les valeurs des paramètres retenues sont décrites dans la table 4.1.

	Paramètre	Valeur
Modèle CRF	Coefficient de Régularisation L1	0.5
	Coefficient de Régularisation L2	0.1
	Nombre maximal d'itérations	50
	Algorithme d'optimisation	lbfgs
Modèle BiLSTM-CNN-CRF	Initialisation des plongements	U(-0.5,0.5)
	Nombre des couches convolutions	3
	Taille des couches convolutions	30
	Dimension de la couche BiLSTM	200
	Taux d'abandon (Dropout)	0.5
	Momentum	0.9
	Taux d'apprentissage	0.01
	Taille du batch	32

TABLE 4.1 – Définition des paramètres d'apprentissage pour le modèle CRF et le modèle BiLSTM-CNN-CRF

La figure 4.3 montre la comparaison entre l'approche utilisant tous les descripteurs de CRF (Version 1, en bleu) à celle favorisant l'adaptation du domaine en omettant le descripteur "valeur littérale" aux termes ayant comme étiquette morpho-syntaxique "Noun" et "Verb" (version 2, en orange). Les résultats confirment que la deuxième approche favorise le transfert des connaissances.
Nous remarquons une amélioration de $13,3\%$, $17,5\%$, et $12,5\%$ en termes de F1-score respectivement sur le corpus des musées, celui des appareils technologiques et le corpus des produits de beauté.

À l'issue de la phase de pré-labellisation, nous avons évalué les étiquettes générées par notre modèle en les comparant à un corpus de référence, que nous avons annoté manuellement. Dans cet exercice, l'éloignement entre les espaces de caractéristiques du domaine source (restaurants) et du domaine cible (produits de beauté) a été pris en compte. Sur notre corpus cible des produits de beauté, qui n'avait pas été annoté préalablement, nous avons observé un taux de labels incorrects de 40% et $59,5\%$ des labels désignant les aspects manquants.

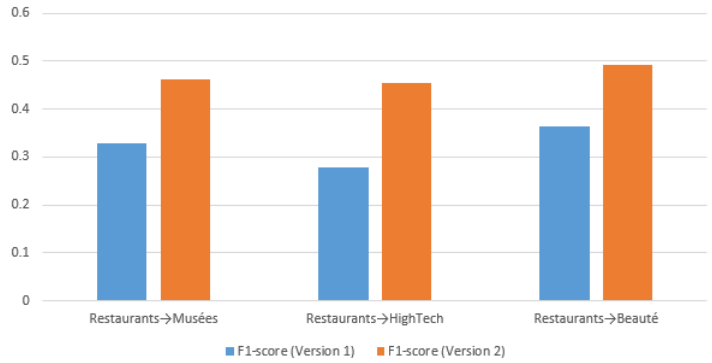


FIGURE 4.3 – Évaluation de la pré-labellisation par adaptation de domaines en prenant en compte tous les descripteurs (Version 1) et en ajustant les descripteurs (Version 2)

4.4.2 Amélioration des pré-labels par apprentissage actif

Suite à la phase de pré-labellisation, nous disposons d'un ensemble de données labellisées L qui serviront comme point de départ pour l'apprentissage d'un modèle supervisé $M(.,\theta_M)$ destiné à l'identification des aspects.
Le processus est itératif et prend en entrée un ensemble initial de données labellisées L (appelé ensemble d'apprentissage) ainsi qu'un grand ensemble U de données non étiquetées (appelé ensemble de *pool*).
Le modèle d'apprentissage $M(.,\theta_M)$ que nous utilisons pour cette tâche est le BiLSTM-CNN-CRF, initialement proposé par [106] pour les tâches de Named Entity Recognition (NER) et de Part-of-Speech (PoS).
Une fonction de stratégie de requête est utilisée pour sélectionner des instances dans l'ensemble de données non étiqueté, comme l'échantillonnage d'incertitude ou l'échantillonnage de diversité. Les instances sélectionnées sont étiquetées par l'oracle et ajoutées à l'ensemble de données étiqueté. La stratégie de sélection θ_{LC} que nous utilisons pour guider l'apprentissage actif est la stratégie de sélection par incertitude, définie comme suit :

$$\theta_{LC}(x) = 1 - \frac{1}{N_{Aspect}} \sum_{i \in C_{Aspect}} max(p(x_i)) \tag{4.3}$$

où C_{Aspect} est l'ensemble des tokens identifiés comme aspect, $N_{Aspect} = |C_{Aspect}|$, et $max(p(x_i))$ est la probabilité maximale du label affecté au i^{eme} token.

Ce processus est répété de manière itérative jusqu'à ce que le niveau requis d'amélioration des performances soit atteint, comme le critère d'arrêt le détermine. De nombreuses stratégies pour interroger les instances de données non étiquetées

sont expliquées dans [87], y compris l'échantillonnage basé sur l'entropie (*entropy based sampling*, l'échantillonnage marginalisé *margin sampling*, et l'échantillonnage basé sur un comité *committee-based sampling*.

```
1 fonction active_learning ( 𝓛 , 𝓤 , θ_LC , m )
2     ℰ ← ∅ ;
3     while needed do
4         I_1, 𝓛 ← query_and_annotate ( θ_LC , 𝓛 , m ) ;
5         ℰ ← ℰ ∪ I_1 ;
6         I_2, 𝓤 ← query_and_annotate ( θ_LC , 𝓤 , m ) ;
7         ℰ ← ℰ ∪ I_2 ;
8         𝓜 ← train ( 𝓛 ) ;
9     return 𝓜 , ℰ ;
```

Alg. 7: Apprentissage actif pour des données labelisées $\mathscr{L}$ et non labelisées $\mathscr{U}$ en partant d'un modèle initial $\mathscr{M}$ avec θ_{LC} une stratégie de sélection par incertitude, m une taille de lot fixée et $\mathscr{E}$ un ensemble de modifications apportées pendant l'apprentissage, tant que les critères définis ne sont pas satisfaits.

L'algorithme 7 détaille le processus global d'apprentissage actif définit par l'étape de correction des labels et de l'enrichissement des données.

Dans ce travail, nous avons adapté le processus standard de l'apprentissage actif :

— Le processus est itératif. A chaque itération, le modèle sélectionne deux ensembles d'instances I_1 et I_2 issus des ensembles $\mathscr{L}$ et $\mathscr{U}$, permettant respectivement de corriger les données déjà labellisées $\mathscr{L}$ et d'ajouter des données de $\mathscr{U}$ dans $\mathscr{L}$. Le modèle est alors ré-entraîné sur $\mathscr{L}$ jusqu'à ce que certains critères d'arrêt soient satisfaits [16]. Les instances sélectionnées sont ensuite ajoutées à $\mathscr{E}$, l'ensemble de toutes les modifications apportées. L'idée est d'utiliser les annotations de l'oracle pour évaluer les incertitudes détectées résultant d'étiquettes erronées dues au pseudo-étiquetage ou à la mauvaise classification du modèle.

— Le processus de sélection de l'instance est implémenté pour s'adapter à la tâche d'étiquetage séquentiel et pour gérer le déséquilibre des classes. Nous avons utilisé l'échantillonnage basé sur l'incertitude en utilisant la stratégie de moindre confiance [17] (LC) (Least Confidence) [95], où les

16. Comme un nombre maximum d'itérations ou un seuil de performance désiré.

17. Nous avons choisi la stratégie de LC pour la sélection des instances en raison de sa rapidité

instances ayant la plus faible confiance dans leur étiquette la plus probable sont sélectionnées. L'incertitude est définie comme $1 - \max(p)$, où p est la distribution de probabilité catégorielle sur les étiquettes. Comme les instances de texte sont composées de séquences de jetons, la valeur de moindre confiance est mise à jour en calculant la moyenne des incertitudes au niveau du jeton, comme le montre l'équation 4.4.

$$LC = 1 - \frac{1}{N}\sum_{i=1}^{N}\max(p_i) \tag{4.4}$$

Le score d'incertitude est ensuite ajusté comme dans l'équation 4.5 en omettant tous les jetons qui n'ont pas été prédits comme un aspect (c'est-à-dire avec une étiquette "O") afin de réduire le biais lié à la dominance de la classe non-aspect et ainsi améliorer l'informativité des échantillons sélectionnés en direction des jetons d'aspect. L'objectif principal du processus de l'apprentissage actif est d'accélérer l'apprentissage du modèle et de minimiser l'effort d'annotation.

$$LC = 1 - \frac{1}{N_{BI}}\sum_{i \in C_{BI}}\max(p_i) \tag{4.5}$$

où $C_{BI} = \{i/\hat{y}_i \in \{'Aspect'\}, 1 \leq i \leq N\}$, $N_{BI} = |C_{BI}|$ et $\hat{y}_i$ est l'étiquette prédite pour le $i^{\text{ème}}$ jeton par le modèle d'apprentissage avec p_i la probabilité de prédiction du jeton considéré.

La Figure 4.4 présente l'architecture du modèle BiLSTM-CNN-CRF. Le modèle se compose de deux couches de plongements de mots (*word embeddings*) : la première est dédiée à l'obtention de plongements de mots tandis que la seconde emploie un réseau neuronal convolutif (CNN) pour extraire et encoder des caractéristiques morphologiques au niveau des caractères. Ces plongements sont ensuite transmis à une couche LSTM bidirectionnelle, permettant de capturer des informations contextuelles sur la séquence de mots. Enfin, une couche CRF est ajoutée à l'extrémité pour prédire la séquence de labels la plus probable.

L'algorithme 8 explicite la fonction *query_and_annotate* qui résume l'étape de sélection des instances et de leur annotation par l'expert.

Dans notre approche, le processus se décompose en trois étapes majeures, illustrées dans la Figure 4.5. La première étape implique deux niveaux de plongements

computationnelle, et de sa capacité de se consacrer uniquement sur l'incertitude la plus immédiate du modèle ce qui s'adapte bien à notre tâche d'étiquetage séquentiel. Cette approche offre une manière efficace de choisir des instances qui sont susceptibles d'améliorer significativement la performance du modèle.

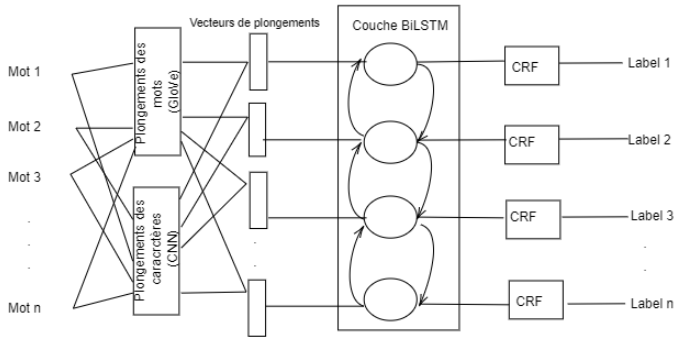


FIGURE 4.4 – Architecture avec *word embeddings*

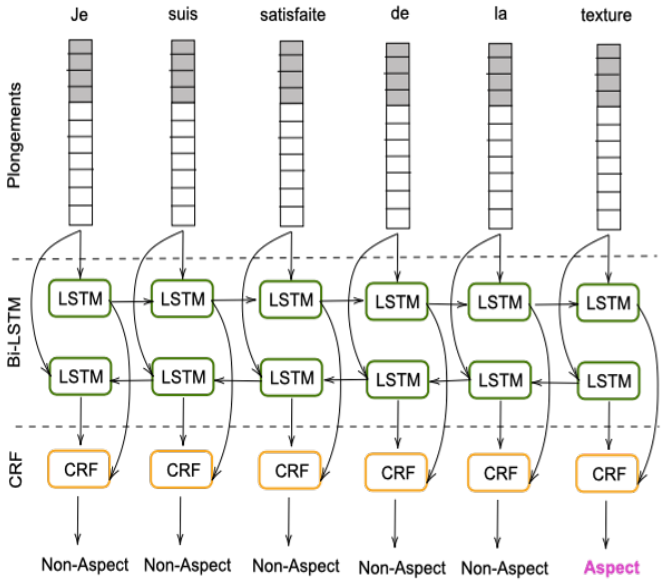


FIGURE 4.5 – Architecture du modèle BiLSTM-CNN-CRF

lexicaux. Le premier niveau recourt à l'algorithme GloVe, préalablement entraîné sur le corpus Wikipédia en français, pour générer des embeddings au niveau du mot. Le second niveau utilise une couche de réseau de neurones convolutifs (CNN) pour capter les caractéristiques morphologiques des mots et les transposer en représentations neurales.

Lors de la deuxième étape, les représentations issues des deux niveaux de plongements sont acheminées vers une couche LSTM bidirectionnelle. Cette couche est conçue pour capturer les dépendances temporelles en considérant à la fois les mots précédents et suivants dans la séquence.

```
fonction query_and_annotate ( θ , X , m )
    Ŷ ← predict (X , θ) ; // Labeliser la donnée X avec
     le modèle d'apprentissage
    I ← ∅ ;
    for m do
        x_b ← argmax_{x∈X} θ(Ŷ) ; // Sélectionner les données
         dont l'indice de confiance est le plus
         faible à partir de X
        y*_b ← label (x_b) ;   // Demander la labellisation
         de l'expert
        I ← I ∪ {(x_b, y*_b)} ;   // Mettre à jour les données
         labellisées
        X ← X − {x_b} ;   // Mettre à jour les données
         initiales
    return I, X
```

Alg. 8: Fonction `query_and_annotate` pour θ une stratégie de sélection et X un ensemble d'avis utilisateur, avec une taille de lot m.

Enfin, la troisième étape consiste en une couche de champ aléatoire conditionnel (CRF) qui assigne la séquence de labels la plus probable aux tokens d'entrée. Le modèle est entraîné sur un ensemble de données étiqueté, scindé selon un ratio de 80/20 pour l'entraînement et la validation. Pour l'optimisation du modèle, nous avons utilisé l'optimiseur Adam et mis en place une technique d'arrêt précoce (*early stopping*) pour prévenir le surapprentissage.

4.5 Expérimentation et évaluation des performances

4.5.1 Description des données

Pour évaluer les différentes étapes du processus proposé, nous nous sommes appuyés essentiellement sur deux types de jeux de données : des jeux de données labellisés des avis mis à disposition dans la compétition SemEval (SemEval-2016 Tâche 5 « Aspect-Based Sentiment Analysis »), et des corpus non labellisés extraits à partir du web dont des sous-ensembles ont été manuellement labellisés et utilisés comme un jeu de données d'évaluation. Les données SemEval sont labellisées avec des entités, des aspects et des valeurs de polarité pertinents. Le processus de labellisation complet est détaillé dans [12].

	Restaurants	Musées
Nombre d'instances	Données d'apprentissage	
	337	-
Nombre d'instances	Données de Test	
	120	162

TABLE 4.2 – Description des données SemEval 2016

	Corpus	
	Appareils Technologiques	Produits de Beauté
Apprentissage	4000	4000
Validation	200	190
Pool	2000	2000
Test	303	300

TABLE 4.3 – Description des corpus non labellisés

Le tableau 4.2 résume les informations sur les jeux de données de SemEval. Ils sont composées d'un ensemble de données d'entraînement de 337 d'avis de clients sur des restaurants, un ensemble de données de validation de 120 d'avis utilisateurs sur les restaurants et un troisième ensemble de données, utilisé pour la validation hors domaine, comportant 162 avis clients sur des musées.

Corpus non labellisés extraits du web

Pour l'évaluation de notre modèle, nous avons utilisé un corpus composé de commentaires d'utilisateurs en langue française depuis des sites web. Ces commentaires portent sur deux domaines spécifiques : les produits technologiques, principalement des appareils et accessoires électroniques, ainsi que les produits de beauté. La méthodologie d'extraction d'avis des utilisateurs est conforme au processus établi dans l'article [21].
Le corpus non labellisé relatif au domaine des appareils électroniques est composé de 4000 instances destinées à l'entraînement du modèle, auxquelles s'ajoutent 2000 instances constituant le pool de données. Pour la phase de validation, ce corpus comprend 200 instances. Concernant le domaine des produits de beauté, la distribution est similaire, à l'exception de la phase de validation qui comprend 190 instances. Le tableau 4.3 décrit la répartition des instances pour chacun des deux corpus.
Cette structuration du corpus vise à fournir une base solide pour l'entraînement et l'évaluation de notre modèle dans des contextes variés.

Critères d'évaluation

Dans ce travail, nous recourons à des métriques de performance standards, à savoir la précision, le rappel et le F1-score, pour évaluer la qualité de chaque étape de notre processus algorithmique. Ces métriques sont définies comme suit :

$$Prcision = \frac{VP}{(VP+FP)},\ Rappel = \frac{VP}{(VP+FN)},$$
$$F1\text{-}score = 2.\frac{Prcision\ .\ Rappel}{(Prcision+Rappel)}$$

Dans ces formules, VP correspond aux Vrais Positifs, soit les instances correctement labellisées ; FP représente les Faux Positifs, c'est-à-dire les instances incorrectement identifiées comme des aspects ; et FN se réfère aux Faux Négatifs, soit les aspects qui n'ont pas été détectés. Ces mesures serviront de fondement pour l'évaluation quantitative de notre modèle.
Dans ce qui suit, nous nous intéressons à la mesure de F1-score pour l'analyse des résultats d'évaluation. Cette mesure résume les performances du modèle en équilibrant la précision et le rappel sur les classes cibles et s'adapte mieux aux tâches d'apprentissage où les classes sont déséquilibrées notamment l'identification des aspects.

4.5.2 Évaluation de la pré-labellisation

La première étape de notre pipeline consiste en une technique de transfert de connaissances à travers une adaptation de domaine. Pour ce faire, nous utilisons un modèle pré-entraîné sur des données issues du domaine des restaurants pour générer des étiquettes sur des données relevant de domaines cibles différents, tels que les produits de beauté et les appareils électroniques.
Afin d'ajuster notre modèle pour cette tâche, nous avons entraîné le CRF en utilisant un ensemble de données en français, spécifiquement le jeu de données des restaurants de SemEval 2016. Les hyperparamètres du modèle ont été ajustés à : le coefficient de régularisation L1 a été fixé à $0,5$, le coefficient de régularisation L2 à $0,1$, le nombre maximum d'itérations à 35, et nous avons utilisé l'algorithme d'optimisation "$lbfgs$" [18] pour le processus d'entraînement.

Le tableau 4.4 illustre les performances du CRF appliqué à trois différentes configurations en utilisant comme domaines cibles les jeux de données de musées, d'appareils technologiques et de produits de beauté respectivement.
Nous comparons l'approche utilisant tous les descripteurs de CRF (Version 1) à celle favorisant l'adaptation du domaine en omettant le descripteur "valeur litté-

18. L'algorithme L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) est plus efficace pour de petits jeux de données et des problèmes peu dimentionnels

rale" aux termes ayant comme étiquette morpho-syntaxique "Noun" ou "Verb" ou "PropN" (version 2).
Les résultats confirment que la deuxième approche favorise le transfert de l'apprentissage. En effet, dans la première approche, la performance de modèle CRF se dégrade considérablement quand le domaine cible est différent du domaine de corpus d'apprentissage. La deuxième approche obtient des meilleures performances avec un F1-score sur la configuration "Restaurants $\rightarrow$ Beauté" comparable à celui de "Restaurants $\rightarrow$ Restaurants".
Cette approche (Version 2) permet d'équilibrer la performance entre les domaines source et cibles et offre une amélioration de $2,3\%$, 11%, et $8,7\%$ en termes de F1-score respectivement sur le corpus des musées, des appareils technologiques et des produits de beauté.
Il est à noter que deux sous-ensembles de corpus (303 des avis sur les appareils technologiques et 300 des avis sur les produits de beauté) ont été manuellement labellisées par un expert pour former le "corpus de référence" nécessaire pour évaluer la pré-labellisation, le modèle BiLSTM-CNN-CRF ainsi que la performance de l'apprentissage actif.

	Précision		Rappel		F1-score	
Source $\rightarrow$ *Cible*	Version 1	Version 2	Version 1	Version 2	Version 1	Version 2
Restaurants $\rightarrow$ *Restaurants*	0.714	0.428	0.563	0.632	0.63	0.51
Restaurants $\rightarrow$ *Musees*	0.597	0.448	0.252	0.326	0.354	0.377
Restaurants $\rightarrow$ *HighTech*	0.361	0.531	0.249	0.447	0.29	0.40
Restaurants $\rightarrow$ *Beaute*	0.509	0.602	0.334	0.463	0.432	0.519

TABLE 4.4 – Évaluation de la pré-labellisation par adaptation de domaines en prenant en compte tous les descripteurs (Version 1) et en ajustant les descripteurs (Version 2)

4.5.3 Évaluation de l'apprentissage actif pour l'identification des aspects

Dans la suite, nous adoptons le modèle BiLSTM-CNN-CRF comme architecture d'apprentissage profond pour identifier les aspects explicites via l'apprentissage actif. Pour initialiser les plongements de mots, nous utilisons les plongements

		Corpus Appareils technologiques			Corpus Produits de Beauté		
Configuration	Stratégie de sélection	Précision	Rappel	F1-score	Précision	Rappel	F1-score
TCL=0%, TED=0%*	—-	0.505	0.372	0.429	0.699	0.405	0.513
TCL=10%, TED=0%	Echantillonage par incertitude	**0.465**	**0.452**	**0.458**	0.684	**0.476**	**0.561**
	Échantillonage aléatoire	0.455	0.438	0.446	**0.686**	0.448	0.542
TCL=20%, TED=0%	Échantillonage par incertitude	**0.487**	**0.513**	**0.50**	**0.724**	**0.547**	**0.623**
	Échantillonage aléatoire	0.483	0.464	0.473	0.714	0.48	0.575
TCL=30%, TED=0%	Échantillonage par incertitude	**0.550**	0.495	**0.52**	**0.718**	**0.625**	**0.668**
	Échantillonage aléatoire	0.474	**0.534**	0.502	0.688	0.493	0.574
TCL=30%, TED=10%	Échantillonage par incertitude	**0.552**	**0.634**	**0.551**	**0.748**	**0.60**	**0.666**
	Échantillonage aléatoire	0.536	0.511	0.523	0.724	0.514	0.601
TCL=30%, TED=20%	Échantillonage par incertitude	**0.596**	**0.586**	**0.591**	**0.774**	**0.631**	**0.695**
	Échantillonage aléatoire	0.586	0.5	0.54	0.751	0.592	0.561

TABLE 4.5 – Évaluation de l'apprentissage actif en terme de précision, rappel et F1-score sous différentes configurations

pré-entraînés Glove en 300 dimensions (GloVe.840B.300d) fournis par [139]. Ce choix se justifie par la qualité et la richesse sémantique de ces plongements.
Quant aux plongements de caractères, ils sont initialisés aléatoirement selon une distribution uniforme $U(-0.5, 0.5)$. Nous suivons ici la configuration établie par [106], fixant le nombre et la taille des couches de convolutions à 3 et 30, respectivement.
La dimension de la couche Bi-LSTM est fixée à 200, un choix basé sur des expérimentations préliminaires indiquant un bon compromis entre performance et complexité calculatoire. Nous utilisons un taux d'abandon (*dropout*) de 30% pour la couche Bi-LSTM et de 10% pour les plongements de mots, afin de prévenir l'*overfitting*.
L'optimisation des paramètres est réalisée par la descente de gradient stochastique (SGD) avec un taux d'apprentissage de 10^{-2}. La taille du *batch* est fixée à 32, ce qui permet d'équilibrer efficacité computationnelle et précision de l'entraînement. Le modèle BiLSTM-CNN-CRF est initialement formé sur 30 *epochs*. Nous itérons ensuite le processus d'apprentissage actif 10 fois. Cette décision est motivée par l'utilisation de la technique d'arrêt précoce pour éviter l'*overfitting*.
Pour évaluer les contributions individuelles de la correction de pré-labels et de l'enrichissement des données d'apprentissage, notre plan d'expérimentation est organisé en deux phases :

1. Dans un premier temps, l'apprentissage actif est employé uniquement pour la correction des labels. Nous avons conduit trois expérimentations avec un taux de correction des labels (TCL) fixé à 10%, 20%, et 30%. Ces valeurs ont été choisies pour étudier l'impact d'une correction croissante sur la performance du modèle.

Corpus des produits de beauté		Classe réelle	
		Non-aspect	Aspect
Classe prédite	Non-aspect	8495	250
	Aspect	125	428
	Nombre total d'aspects		678

TABLE 4.6 – Matrice de confusion pour le corpus des produits de beauté

2. Dans un second temps, nous avons considéré l'enrichissement des données d'apprentissage. Dans ce contexte, le TCL est maintenu à 30% tandis que le taux d'enrichissement de données (TED) varie de 10% à 20%. Cette phase vise à évaluer le bénéfice potentiel d'un ensemble de données plus volumineux.

En limitant les valeurs de TCL et TED à 30% et 20% respectivement, nos observations empiriques indiquent un gain de 50% en termes d'effort et de temps pour la labellisation complète du jeu de données. Ce choix est principalement guidé par le désir de minimiser les erreurs de pré-labellisation qui pourraient affecter la performance du modèle.

Le tableau 4.5 présente les mesures d'évaluation obtenues lors des différentes expérimentations sur les avis portant sur les produits de beauté et les appareils électroniques. Nos résultats mettent en évidence la supériorité de la stratégie de sélection basée sur l'incertitude par rapport à une approche aléatoire. Plus précisément, nous observons une amélioration substantielle du F1-score, justifiant ainsi le choix de cette stratégie de sélection pour maximiser l'efficacité de l'apprentissage actif.
De plus, nous remarquons que le processus itératif de correction des labels améliore les performances du modèle. Ceci est corroboré par une augmentation du rappel et du F1-score en augmentant le TCL, démontrant ainsi l'efficacité de ce paramètre. En ce qui concerne l'enrichissement des données, une progression similaire du F1-score est observée à mesure que le TED augmente.

Globalement, nos résultats attestent de l'utilité de l'application d'un processus d'apprentissage actif. Nous notons des augmentations respectives de 16.9% et 18.2% du F1-score sur les avis relatifs aux appareils électroniques et aux produits de beauté. Ces augmentations, bien que spécifiques à notre jeu de données, soulignent le potentiel de la méthode pour des applications en environnements réels, tout en reconnaissant la nécessité d'études supplémentaires pour valider ces résultats dans des contextes variés.
Les tableaux 4.6 et 4.7 illustrent les matrices de confusion pour les corpus des

Corpus des appareils Tech		Classe réelle	
		Non-aspect	Aspect
Classe prédite	Non-aspect	17229	317
	Aspect	304	450
	Nombre total d'aspects		767

TABLE 4.7 – Matrice de confusion pour le corpus des produits technologiques

produits de beauté et des appareils technologiques à la suite de l'application de l'apprentissage actif. Nous observons que 428 sur un total de 678 aspects ont été correctement labellisés par le modèle sur le corpus des produits de beauté et 450 sur un total de 767 aspects ont été identifiés pour le corpus des appareils technologiques.
L'écart de performance entre les deux corpus s'explique par la complexité du domaine des appareils technologique due à l'hétérogénéité des termes aspects utilisés selon la catégorie des objets considérés (e.g : son, image, batterie, stockage, accessoires ...)

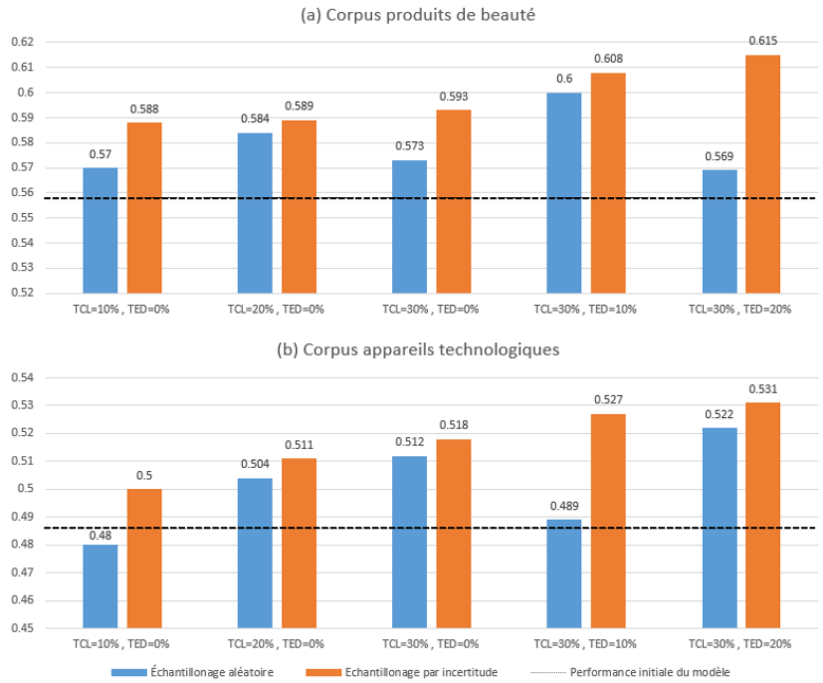


FIGURE 4.6 – Évaluation de l'apprentissage actif en terme de F1-score sous différentes configurations sur les corpus : (a) produits de beauté et (b) appareils technologiques

La figure 4.6 illustre les résultats d'évaluation, en termes de F1-score, de différentes expérimentations décrites ci-dessus sur les jeux de données des avis concernant les produits de beauté et les produits technologiques. Elle montre la comparaison entre l'approche utilisant tous les descripteurs de CRF (Version 1, en bleu) à celle favorisant l'adaptation du domaine en omettant le descripteur "valeur littérale" aux termes ayant comme étiquette morpho-syntaxique "Noun" et "Verb" (version 2, en orange). Les résultats confirment que la deuxième approche favorise le transfert des connaissances. Nous remarquons une amélioration de $13,3\%$, $17,5\%$, et $12,5\%$ en termes de F1-score respectivement sur le corpus des musées, celui des appareils technologiques et le corpus des produits de beauté.
Nous comparons la stratégie de sélection par incertitude que nous avons adaptée à la stratégie de sélection aléatoire. Nous désignons par TCL le taux de correction de labels et par TED le taux d'enrichissement des données. Nous limitons les valeurs de TCL et le TED à 30% et 20% respectivement, pour garantir un gain de 50% en termes d'effort et de temps nécessaires pour une labellisation totale du jeu de données d'apprentissage. Nous avons également favorisé la correction des labels afin de réduire l'impact des erreurs de pré-labellisation sur la performance du modèle. La ligne en pointillé représente la performance du modèle BiLSTM-CNN-CRF avant l'application de l'apprentissage actif.

4.6 Conclusion et perspectives

Dans cette étude, nous avons présenté une approche intégrée pour identifier les aspects dans des corpus non annotés, en mettant un accent particulier sur les langues peu dotées en ressources comme le français pour ce cas là. En utilisant des corpus composés d'avis utilisateurs sur des produits variés, du maquillage aux appareils électroniques, notre méthodologie s'appuie fondamentalement sur l'apprentissage par transfert et l'apprentissage actif. Ces techniques nous permettent d'abord de pré-étiqueter les données de manière autonome pour compenser le déficit d'annotations, puis d'affiner ces prédictions grâce à des cycles itératifs d'apprentissage actif. Les données empiriques recueillies démontrent une amélioration notable des performances du modèle, lorsque nous observons un gain de près de 50% en termes d'effort et de temps nécessaires pour une labellisation totale du jeu de données d'apprentissage. Dans le prolongement de cette recherche, nous considérons l'extension de notre méthodologie aux corpus multilingues et envisageons l'ajout d'une étape supplémentaire de catégorisation des aspects identifiés, ce qui augmentera la complexité du système. Nous sommes curieux d'explorer également l'efficacité des LLMs dans des scénarios d'apprentissage actif. Cette exploration pourrait nous amener à considérer des métriques supplémentaires comme l'alpha de Krippendorff pour évaluer la cohérence entre les annotations humaines et les

prédictions du modèle, notamment pour apporter un éclairage complémentaire au score F1 traditionnellement utilisé. L'adéquation de notre approche avec les Modèles de Langage à Grande Échelle (LLMs) renforcera son utilité non seulement pour l'annotation des données, mais aussi comme oracle dans le cadre de l'apprentissage actif dans le domaine de l'annotation et de la labellisation de données. Néanmoins, il est important de préciser que leur efficacité peut être influencée par le *fine-tuning* sur des corpus spécifiques.

Contributions

Nos contributions enrichissent divers domaines de la recherche. D'abord, nous offrons une nouvelle méthode robuste pour créer des corpus annotés à partir de données non structurées, ce qui a une utilité manifeste pour les langues à faibles ressources. De plus, notre approche bimodale de labellisation se distingue par son efficacité économique et sa précision, validée par des tests empiriques.
En somme, cette recherche ouvre la voie à des avancées dans la labellisation des données et dans l'application des modèles de langage à des problématiques en traitement automatique du langage naturel.

Chapitre 5

Conclusion et perspectives

Sommaire

Pour conclure, nous synthétisons les principaux enjeux et solutions abordés, et mettons en lumière la manière dont nos approches contribuent aux avancées dans le domaine du TALN.
Nos travaux se structurent autour de deux axes complémentaires, reflétant une approche holistique qui englobe à la fois les défis posés et les solutions apportées dans notre domaine de recherche. Le premier axe se focalise sur la conception et le développement d'outils spécifiques pour la création de corpus et de datasets. Ces ressources textuelles sont conçues pour être robustes, multilingues et adaptées aux tâches d'apprentissage machine. Le second axe, en revanche, capitalise sur ces corpus pour développer des modèles hybrides basés sur des techniques d'apprentissage profond. Ces modèles sont orientés vers des tâches spécifiques, telles que l'analyse de sentiment et l'annotation sémantique d'aspects. Ensemble, ces deux axes forment une vision intégrée qui vise à aborder de manière complète et efficace les diverses problématiques de notre domaine de recherche.
Dans le premier axe nous abordons l'enjeu toujours d'actualité de la création de corpus à partir de sources web. Ce défi découle de la difficulté inhérente à la conception d'un outil polyvalent apte à isoler du texte pertinent et sans bruit, indépendamment de la structure ou de la langue du site web source. Nous avons exploré diverses approches et développé des outils qui, selon l'approche, satisfont en grande partie ou complètement ces critères. Ce qui suit est un récapitulatif des méthodes et outils que nous avons proposés.
Le premier outil présenté, REVSCRAP, est conçu pour extraire des commentaires

en identifiant des nœuds HTML répétitifs dans des sous-blocs spécifiques. Cette approche s'appuie sur un seuil de répétition et une taille moyenne du contenu. Bien que REVSCRAP soit semi-automatique (requérant des URL prédéfinies) et incorpore plusieurs mécanismes de filtrage pour éliminer le bruit, il est efficace et demande un temps de calcul important, comme présenté dans [115]. Sa capacité à composer un corpus "propre", exempt de bruit, constitue une contribution importante. Le corpus créé a notamment servi de base pour l'entraînement et les tests dans le cadre de la tâche d'analyse de sentiment.
Le deuxième outil, DYCORC [109], est un extracteur de contenu non supervisé qui se caractérise par une analyse fine de la structure DOM des pages web. Ce processus d'analyse s'appuie sur diverses distances de chaîne pour identifier et isoler les éléments textuels pertinents. Conçu pour parcourir les forums en ligne, l'outil récupère le code source de chaque page et procède à son analyse de manière individuelle. La structure HTML et le texte sont nettoyés des erreurs et des incohérences avant d'être convertis en un arbre DOM, qui est ensuite enregistré au format XML. Le principe directeur de cette méthode repose sur l'idée que les structures les plus fréquentes, contenant des textes diversifiés, sont susceptibles d'inclure du contenu pertinent. Une évaluation rigoureuse a permis de sélectionner les distances de chaîne les plus performantes parmi les sept examinées. En comparaison avec d'autres modèles, DYCORC a démontré une qualité d'extraction supérieure, tout en étant plus rapide en termes de temps moyen de traitement.
Le dernier outil proposé WEBT-IDC [21] fusionnant des techniques de *crawling* et de *scraping*, rivalise efficacement avec les méthodes avancées du domaine. Indépendant de l'architecture intrinsèque des pages web, il s'adapte facilement à diverses structures de forums et de blogs qui intègrent un élément de pagination, comme décrit dans la section 8. Ce facteur d'indépendance lui permet de générer un corpus thématique multilingue sans bruit, en conservant uniquement les données cibles pertinentes. La performance globale de l'outil est renforcée par l'utilisation du traitement *multithread* et l'accès direct aux sections utiles des pages web [1].

Le deuxième axe de nos travaux de recherche a abordé la problématique de l'analyse de sentiment et l'annotation des aspects dans un contexte multilingue et multithématique. Nous résumons nos approches en mettant l'accent sur l'efficacité de modèles hybrides d'apprentissage profond proposés pour leur aptitude à répondre aux analyses de données textuelles.
La première question abordée est celle de la classification de la polarité des avis dans leur langue d'origine, la distinction de différents types d'opinions et la démons-

1. Les corpus générés avec cet outil sont utilisés pour l'entraînement et tests des travaux de thèse de Maroua Boudabous.

tration de la robustesse des architectures de réseaux profonds. Les modèles LSTM, Convolutional Network (ConvNet), et surtout leurs combinaisons hybrides, se sont avérés être particulièrement efficaces. En utilisant ces architectures, nous avons pu extraire des caractéristiques importantes du texte tout en gérant des dépendances temporelles complexes. Chaque architecture a contribué d'une manière unique au traitement des données. Les réseaux convolutionnels se sont avérés excellents pour l'extraction de caractéristiques locales, tandis que les LSTM ont démontré leur aptitude à gérer les dépendances à long terme. L'intégration de la bidirectionnalité dans le Bi-ConvLSTM a offert une capacité accrue à contextualiser les données, bien que l'augmentation de la performance reste modeste. L'un des résultats les plus encourageants a été la robustesse des modèles lorsqu'ils ont été formés sur des données brutes (sans prétraitement ou étiquetage spécifique). Cela suggère non seulement la flexibilité de ces architectures, mais aussi leur applicabilité à des contextes où les ressources linguistiques sont limitées, y compris dans les langues moins dotées en ressources.

Par la suite, nous avons mis l'accent sur l'annotation sémantique, en ce qui concerne les aspects dans les langues moins dotées en ressources. L'utilisation de l'apprentissage par transfert et de l'apprentissage actif a montré une efficacité dans le pré-étiquetage des données, permettant de combler le déficit d'annotations dans ces langues. Nos analyses empiriques valident une amélioration significative du modèle lorsque près d'un tiers des étiquettes initiales ont été corrigées grâce à des cycles itératifs d'apprentissage actif. L'un des atouts majeurs de cette recherche est la formalisation du problème d'annotation séquentielle, qui vise à étiqueter une séquence de mots en fonction d'un ensemble prédéfini d'étiquettes 4.4.1. Cette formalisation est exploitable dans une variété de techniques d'apprentissage automatique, notamment l'apprentissage par transfert et l'apprentissage actif, pour améliorer la qualité et la précision des annotations.

En employant une approche intégrée, notre recherche réussit à identifier des aspects spécifiques dans des corpus non annotés. Cette méthode se révèle particulièrement efficace pour les langues moins dotées en ressources, car elle nécessite seulement un petit ensemble de données annotées pour obtenir des résultats probants. Sa polyvalence lui permet de s'appliquer à différents domaines, allant des produits de beauté aux appareils électroniques, ce qui est particulièrement utile lorsque les données spécifiques à un domaine sont rares. Nos analyses empiriques confirment que cette méthodologie améliore significativement les performances du modèle. Cette amélioration est particulièrement notable lorsque les annotations initiales sont optimisées au cours de cycles itératifs d'apprentissage actif, ce qui se traduit par une réduction de 50 % de l'effort et du temps nécessaires pour la labellisation complète du jeu de données.

En somme, l'efficacité et la flexibilité de notre approche intégrée se manifestent à la fois dans l'analyse de sentiment et l'annotation d'aspects, tout en s'adaptant

particulièrement bien aux langues moins dotées en ressources. Nos résultats empiriques montrent une réduction substantielle de l'effort et du temps de labellisation, marquant ainsi une avancée importante en annotation de texte. Ce travail contribue significativement à la collecte et à l'annotation de corpus web multilingues, tout en abordant les défis liés à la labellisation des données et à l'utilisation de modèles de langage. Il adresse également des défis clés dans le domaine du TALN, notamment le problème de la rareté des données et les limites des méthodes d'annotation classiques. Ce triptyque de contributions – la collecte de corpus, l'analyse de sentiment et l'annotation séquentielle – forme une approche holistique pour le traitement des données textuelles. Des perspectives de recherches futures s'étendent non seulement à l'optimisation de ces outils et à leur application dans des corpus multilingues, mais aussi à des applications pratiques dans des domaines variés. L'annotation sémantique pourrait bénéficier d'un grain encore plus fin à travers l'intégration de filets sémantiques [89], qui offrent une représentation structurée pour manipuler des connaissances complexes. Cette approche pourrait également être testée et étendue pour améliorer l'efficacité de l'apprentissage par transfert et de l'apprentissage actif pour l'annotation des aspects.
Les deux dimensions de ce travail se complètent mutuellement et posent des jalons pour des recherches futures, notamment l'optimisation des modèles et l'exploration d'autres architectures et techniques pour enrichir davantage les caractéristiques des données et améliorer leur exploitation. Un élément clé de cette optimisation concerne le choix judicieux des architectures de modèles.

5.0.1 Critères de choix des modèles d'apprentissage profond

Le choix des modèles d'apprentissage profond dans notre recherche a été largement influencé par la complexité des tâches à accomplir, allant de l'analyse de sentiment à l'apprentissage actif et par transfert. Le modèle CNN a été employé pour sa capacité à extraire efficacement des traits locaux, ce qui est crucial pour comprendre le contexte des mots dans l'analyse de sentiment. Le modèle RNN et sa variante plus avancée, LSTM, ont été utilisées pour capturer les dépendances temporelles dans les séquences de texte. La combinaison de CNN-LSTM et BiConv-LSTM a été mise en œuvre pour tirer parti à la fois des caractéristiques spatiales et temporelles du texte. Pour des tâches plus avancées comme l'apprentissage par transfert et actif, des modèles tels que CRF et BiLSTM-CNN-CRF ont été adoptés pour leur capacité à capturer des dépendances complexes et à améliorer la séquence d'étiquetage.
L'apprentissage profond offre plusieurs avantages par rapport aux modèles classiques de l'apprentissage automatique. Il est capable de :

— modéliser des relations complexes, puisque les réseaux neuronaux profonds sont conçus pour apprendre des représentations hiérarchiques de données.

— gérer de grandes dimensions de fonctionnalités, puisque les méthodes d'apprentissage profond sont capables de traiter efficacement des espaces de fonctionnalités de grande dimension. Ceci est crucial dans des domaines comme l'analyse de sentiment, où le vocabulaire peut être très vaste et où chaque mot ou phrase peut être considéré comme une caractéristique.
— travailler avec des représentations des données moins élaborées manuellement, car contrairement aux méthodes traditionnelles qui nécessitent souvent un prétraitement des données et l'extraction manuelle des caractéristiques, les réseaux neuronaux profonds peuvent apprendre des représentations pertinentes directement à partir des données brutes.

Le lien entre l'apprentissage profond et le langage naturel s'est considérablement renforcé, avec les modèles profonds qui ont fait leurs preuves dans un large éventail de tâches liées à l'analyse et au traitement du langage. Leur utilisation a souvent apporté des améliorations significatives en termes de performance et de précision. Les modèles basés sur les Transformers se sont imposés comme le standard *de facto* en raison de leur polyvalence et leur aptitude à gérer des séquences complexes à grande échelle. Ils ont en effet révolutionné trois domaines clés du TALN. Premièrement, ils ont amélioré l'extraction de texte et la recherche d'information grâce à des représentations internes sophistiquées. Deuxièmement, ils ont amélioré l'analyse linguistique en gérant efficacement les dépendances à longue distance, allant des relations syntaxiques [96] à la conceptualisation des connaissances [18]. Enfin, leur puissance prédictive a considérablement amélioré la génération de texte. Cette architecture polyvalente, est comparable à un "véritable couteau suisse" de l'ingénieur linguiste selon [195].

L'efficacité des architectures de type Transformer dans la mise en œuvre de l'apprentissage par transfert est particulièrement pertinente dans notre contexte [191]. En effet, nous disposons d'un ensemble de données partiellement annotées grâce à une stratégie d'apprentissage actif [2]. Le *fine-tuning* de modèles tels que BERT permet d'exploiter un riche ensemble de fonctionnalités déjà apprises par le modèle pré-entraîné [35]. Ainsi, l'apprentissage par transfert nécessite généralement beaucoup moins de ressources que la formation d'un modèle à partir de zéro [3], ce qui pourrait nous permettre d'obtenir des résultats significatifs même avec un ensemble de données de taille modeste.

En résumé, les architectures de type Transformer ont redéfini les normes en matière de TALN et leur flexibilité les rend particulièrement adaptées aux expérimentations

2. Cette annotation partielle permet un *fine-tuning* du modèle avec des performances qui pourraient être légèrement inférieures en termes de précision. Une approche mixte, combinant ce corpus avec un autre entièrement annoté, est également envisageable pour améliorer les performances du modèle.

3. Il est à noter que la réduction de l'utilisation des ressources a aussi un impact environnemental positif. Pour une discussion sur l'empreinte carbone des modèles de ML, voir par exemple [170]

dans des domaines variés. Cela va au-delà des simples avantages techniques et s'étend aux possibilités de collaboration interdisciplinaire, comme nous le verrons dans la section suivante concernant l'application de ces technologies dans un projet mêlant théâtre classique, art et intelligence artificielle.

5.0.2 Dialogues du théâtre classique avec un chatbot

Cette transdisciplinarité trouve une application concrète dans notre participation à un projet collaboratif qui regroupe chercheurs, artistes, et documentalistes. Le projet illustre la puissance des modèles génératifs dans la simulation du langage humain [148] et il soulève des questions à la frontière de l'IA et l'art. Nous souhaitons le présenter brièvement avant de conclure ce manuscrit. LITTE_BOT est un projet ArTeC [4] visant à transposer des personnages du théâtre du XVIIe siècle, notamment Don Juan, en chatbot. Ce projet a donné des contributions présentées dans diverses conférences [5] et publications [136, 55, 25]. Pour commencer nous avons utilisé le modèle Seq2Seq [183, 156], cherchant à générer des dialogues qui sont à la fois naturels et théâtralement authentiques. Ce choix de modèle offre un compromis entre performance et coût computationnel, bien que des problèmes de cohérence sémantique subsistent.
Un second modèle, GPT-fr, un modèle de type GPT-2 est également testé pour un nouveau chatbot le MoliAIre. L'objectif d'apprentissage est de reconstruire les extraits de dialogue en minimisant le logarithme de la perplexité. Chaque extrait de dialogue est divisé en une séquence de jetons U = {u_1, ..., u_n} $U = \{u_1, \ldots, u_n\}$ et les paramètres du modèle Θ sont optimisés en minimisant :

$$\mathscr{L}(U, \Theta) = -\sum_{i=1}^{n} \log P(u_i | u_1, \ldots, u_{i-1}, \Theta)$$

Pour l'inférence, nous employons une génération stochastique avec la méthode "top-k" pour favoriser diversité et improvisation. Le modèle attribue des probabilités aux jetons suivants en fonction du contexte, et le prochain jeton est choisi parmi les k plus probables, avec $k = 40$.
Pour l'évaluation, le modèle atteint une perplexité de 14.88. Nous utilisons le score BLEU comme indicateur de style. Il est difficile de définir un texte de référence en improvisation théâtrale, donc nous comparons les répliques générées aux œuvres originales de Molière.

4. Voir https://eur-artec.fr/projets/litte_bot/

5. Notamment présenté lors de "Affects, Compagnons Artificiels et Interactions-ACAI" AFIA 2022https://ci.mines-stetienne.fr/pfia2022/Ateliers/ACAI/, et "Futurs Fantastiques 2021" https://bnf.hypotheses.org/files/2021/11/FF21_Programme_VF5_20211118.pdf.

— Générez n répliques sans contexte initial pour simuler la génération libre par MoliAIre.
— Utilisez les répliques des 32 œuvres de Molière comme références.
— Calculez le score BLEU pour évaluer les répliques générées.

Cette méthode est sensible à la stratégie de génération et au choix de n. Nous avons effectué 5 itérations avec $k = 40$ et $n = 1000$.

Itération 1	Itération 2	Itération 3	Itération 4	Itération 5	Moyenne
30,54	37,99	31,55	29,74	22,92	30,63

TABLE 5.1 – Score BLEU calculé vec top-k=40 et n=1000 sur 5 itérations

Le modèle excelle en improvisation et ne surapprend pas. Ses répliques sont souvent inédites par rapport aux textes de Molière. Le style demeure constant, même quand sollicité avec du français moderne. Sa principale limite est la cohérence dans des contextes éloignés de l'œuvre de Molière. MoliAIre peine à créer des rimes en raison de son architecture autorégressive. Pour pallier ce problème, nous utilisons la modélisation de langage inversée dans un nouveau modèle, MoliAIre-VERSE. Ce dernier est spécifiquement entraîné sur des vers, et la tokenisation des lignes est inversée.

Vers original								
Parbleu	je	ne	vois	pas,	lorsque	je	m'	examine
1	2	3	4	5	6	7	8	9
Où	prendre	aucun	sujet	d'	avoir	l'	Âme	chagrine
10	11	12	13	14	15	16	17	18
Vers inversé								
examine	m'	je	lorsque	pas,	vois	ne	je	Parbleu,
9	8	7	6	5	4	3	2	1
chagrine	Âme	l'	avoir	d'	sujet	aucun	prendre	Où
18	17	16	15	14	13	12	11	10

TABLE 5.2 – Méthode d'inversion des vers utilisée pour la formation de MoliAIre-RIME. Les jetons de chaque vers sont inversés, mais l'ordre des vers est conservé.

Les résultats sont prometteurs : le modèle génère facilement des rimes et des alexandrins. Nous avons développé des agents conversationnels pour le théâtre, en explorant notamment la génération stylisée et la création de rimes. Nous sommes conscients que le chemin vers une génération de texte parfaitement stylisée et théâtralement cohérente est encore long. Cependant, les résultats sont encourageants et incitent à une exploration plus approfondie. Par ailleurs, nous avons transposé ce modèle avec une version en allemand où le chatbot génère des dialogues des pièces de théâtre de Brecht. Avant d'intégrer les œuvres de Brecht, nous avons

préparé notre modèle et ajusté ses paramètres sur un corpus des pièces de théâtre de Molière. Ce premier pas a servi de base solide pour le développement ultérieur du modèle, qui a ensuite été adapté à l'univers thématique et linguistique de Brecht. Nous envisageons de poursuivre cette recherche en augmentant la taille et la diversité des ensembles de données, en expérimentant avec d'autres architectures de modèle, et en affinant les mécanismes de génération de rime et de métrique.
En définitif, ce manuscrit est une étape plus qu'une finalité dans la recherche d'une intelligence artificielle à la fois respectueuse de l'environnement et capable de capturer la complexité et la diversité du langage humain, le tout dans un cadre éthique responsable.

Bibliographie

[1] Prodigy · an annotation tool for ai, machine learning nlp. `https://prodi.gy`. 93

[2] M.-A. Abchir, Isis Truck, and Anna Pappa. Dealing with natural language interfaces in a geolocation context. In *Proceedings of the The 10th International FLINS Conference on Computational Intelligence in Decision and Control*, pages 806–811, Istanbul, Turquie, 2012. 1, 4

[3] Mohammed-Amine Abchir, Isis Truck, and Anna Pappa. *Fuzzy Semantics in Closed Domain Question Answering*, pages 171–188. Atlantis Press, Paris, 2013. 1, 4

[4] Muhammad Abdul-Mageed and Mona Diab. Awatif : A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA), 2012. 61

[5] Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. Subjectivity and sentiment analysis of Modern Standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 587–591, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 61

[6] Kilgarriff Adam and Grefenstette Gregory. Introduction to the special issue on the web as corpus. *Comput. Linguist.*, 29(3) :333–347, sep 2003. 10

[7] Basant Agarwal and Namita Mittal. *Machine Learning Approach for Sentiment Analysis*, pages 21–45. Springer International Publishing, Cham, 2016. 60

[8] Munir Ahmad, Shabib Aftab, Iftikhar Ali, and Noureen Hameed. Hybrid tools and techniques for sentiment analysis : A review. *International Journal of Multidisciplinary Sciences and Engineering*, 8 :31–38, 06 2017. 60

[9] Al Sallab Ahmad A., Gilbert Ramy Baly, Hazem Hajj Badaro, El Hajj Wassim, and Khaled B. Shaban. Deep learning models for sentiment analysis in arabic, 2015. 62

[10] Abdulaziz M. Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. A combined CNN and LSTM model for arabic sentiment analysis. In *Lecture Notes in Computer Science*, pages 179–191. Springer International Publishing, 2018. 69

[11] Halevy Alon, Norvig Peter, and Pereira Fernando. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24 :8–12, 2009. 11

[12] Marianna Apidianaki, Xavier Tannier, and Cécile Richart. Datasets for aspect-based sentiment analysis in french. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1122–1126, 2016. 104

[13] Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135. Association for Computational Linguistics, 2008. 61, 65

[14] Marco Baroni and Silvia Bernardini. Bootcat : Bootstrapping corpora and terms from the web. In *LREC*, page 1313, 2004. 14, 23, 32

[15] Marco Baroni and Adam Kilgarriff. Large linguistically-processed web corpora for multiple languages. In *Demonstrations*, 2006. 10, 14

[16] Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1) :1–127, 2009. 62

[17] Vincent Berment. *Méthodes pour informatiser des langues et des groupes de langues 'peu dotées'*. PhD thesis, 2004. Thèse de doctorat dirigée par Boitet, Christian Informatique Grenoble 1 2004. 57

[18] S. Bhatia and R. Richie. Transformer networks of human conceptual knowledge. *Psychological Review*, 2022. Advance online publication. 117

[19] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3 :993–1022, March 2003. 64

[20] Cristina Bosco, Viviana Patti, and Andrea Bolioli. Developing corpora for sentiment analysis : The case of irony and senti-tut (extended abstract). In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, page 4188, 2015. 22

[21] Maroua Boudabous and Anna Pappa. Webt-idc : A web tool for intelligent dataset creation a use case for forums and blogs. *Procedia Computer Science*, 192 :1051–1060, 2021. Knowledge-Based and Intelligent Information Engineering Systems : Proceedings of the 25th International Conference KES2021. 5, 105, 114

[22] Maroua Boudabous and Anna Pappa. Apprentissage actif pour l'extraction des aspects explicites : application à des avis non annotés en français. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, pages 20–28, 2022. © 2022 CNRS. 7

[23] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4) :467–479, December 1992. 64

[24] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv :2005.14165*, 2020. 9

[25] Tristan Cazenave, Guillaume Grosjean, Baptiste Rozière, and Anna Pappa. Moliaire, a theatrical agent which speaks like molière's characters. In *Proceedings of the XXV Generative Art Conference, GA2022*, pages 286–294. Domus Argenia Publisher, 2022. 118

[26] Zhuang Chen and Tieyun Qian. Enhancing aspect term extraction with soft prototypes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2107–2117, 2020. 92

[27] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015. 67

[28] James Clark, Steve DeRose, et al. Xml path language (xpath) version 1.0, 1999. 41

[29] Ronan Collobert and Jason Weston. A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, New York, NY, USA, 2008. ACM. 61

[30] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12, 2011. 61

[31] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann LeCun. Very deep convolutional networks for natural language processing. *CoRR*, abs/1606.01781, 2016. 67

[32] Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. Roadrunner : Towards automatic data extraction from large web sites. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pages 109–118, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. 23

[33] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA, 2003. ACM. 65

[34] Kerstin Denecke. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 507–512. IEEE, 2008. 56

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018. 47, 96, 117

[36] Marco Dinarelli and Tellier Isabelle. Étude de réseaux de neurones récurrents pour l'étiquetage de séquences. In *JEP-TALN-RECITAL 2016, volume 2 : TALN*, pages 98–111, Paris, France, 2016. Association pour le Traitement Automatique des Langues. 61

[37] Cícero Nogueira dos Santos and Maíra A. de C. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, 2014. 61

[38] Mehdy Dref and Anna Pappa. An interaction approach between services for extracting relevant data from tweets corpora. In Antonio Moreno Ortiz and Chantal Perez-Hernandez, editors, *CILC2016. 8th International Conference on Corpus Linguistics*, volume 1 of *EPiC Series in Language and Linguistics*, pages 97–110, 2016. 22

[39] G. Druck, B. Settles, and A. McCallum. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 81–90. Association for Computational Linguistics, 2009. 94

[40] Miles Efron. Cultural orientation : Classifying subjective documents by cociation analysis. In *AAAI Fall Symposium on Style and Meaning in Language, Art, and Music*, pages 41–48, 2004. 60

[41] Gil Elbaz. Common crawl. `https://commoncrawl.org/`, 2014. Consulté le : 2022-11-18. 9

[42] Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2) :179–211, 1990. 55

[43] Jakob Elming, Barbara Plank, and Dirk Hovy. Robust cross-domain sentiment analysis for low-resource languages. pages 2–7, 06 2014. 61

[44] Hady ElSahar and Samhaa R. El-Beltagy. Building large arabic multi-domain resources for sentiment analysis. In *CICLing*, 2015. 73

[45] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques : A survey. *Knowledge-Based Systems*, 70 :301 – 323, 2014. 22

[46] William H. Fletcher. Making the web more useful as a source for linguistic corpora. In *Corpus Linguistics in North America*, pages 191–205. Rodopi, 2004. 10

[47] Jerry A. Fodor. *The Language of Thought*. Harvard University Press, 1975. 2

[48] Karen Fort, Gilles Adda, and Kevin Bretonnel Cohen. Amazon mechanical turk : Gold mine or coal mine ? *Computational Linguistics*, pages 413–420, 2011. hal-00569450. 91

[49] K. Ganchev, J. A. Graça, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 11 :2001–2049, 2010. 94

[50] Roger Garside, Geoffrey Leech, and Anthony Mark McEnery. Corpus annotation : Linguistic information from computer text corpora. 1997. 88

[51] Mohsen Ghadessy, Alex Henry, and L. Roseberry, Robert. *Preface. In Small Corpus Studies and ELT : Theory and Practice*. Studies in Corpus Linguistics. J. Benjamins Publishing Company, Amsterdam,Philadelphia, 2001. 11

[52] Rayid Ghani, Rosie Jones, and Dunja Mladenić. Mining the web to create minority language corpora. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 279–286, New York, NY, USA, 2001. ACM. 14

[53] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), jul 2023. 93

[54] Stefan Gries and Andrea Berez. *Linguistic Annotation in/for Corpus Linguistics*, pages 379–409. 06 2017. 87

[55] Guillaume Grosjean, Anna Pappa, Baptiste Roziere, and Tristan Cazenave. Dialogue avec molière. In *Traitement Automatique des Langues Naturelles*, pages 6–7, 2022. `hal-03701466`. 118

[56] Jiawei Han and Kevin Chang. Data mining for web intelligence. *Computer*, 35(11) :64–70, November 2002. 10

[57] Zellig Harris. Distributional structure. *Word*, 10(23) :146–162, 1954. 2, 64

[58] Nabil Hathout and Ludovic Tanguy. Webaffix : une boîte à outils d'acquisition lexicale à partir du web. *Revue québécoise de linguistique*, 32(1) :61–84, 2005. 12

[59] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting*

of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98, pages 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics. 59

[60] F. Herrera and L. Martinez. A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems*, 8(6) :746–752, 2000. 3

[61] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7) :1527–1554, 2006. 62

[62] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, 2012. 72

[63] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8) :1735–1780, November 1997. 55

[64] James Hong and Michael Fang. Sentiment analysis with deeply learned distributed representations of variable length texts. 2015. 79

[65] Thanda Htwe and Nan Saing Moon Kham. Extracting data region in web page by removing noise using dom and neural network. In *3rd International Conference on Information and Financial Engineering*, 2011. 37

[66] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004. 92

[67] Ozan İrsoy and Claire Cardie. Opinion mining with deep recurrent neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 720–728, 2014. 61

[68] Niklas Jakob and Iryna Gurevych. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1035–1045, 2010. 92

[69] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406) :414–420, 1989. 28

[70] Jiang Jingtian, Yu Nenghai, and Lin Chin-Yew. Focus : Learning to crawl web forums. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 33–42, New York, NY, USA, 2012. ACM. 22

[71] M. Sinclair John. *The automatic analysis of corpora*, pages 379 – 397. De Gruyter, 1991. 9, 11

[72] Faustina Johnson and Santosh Kumar Gupta. Web content mining techniques : a survey. *International Journal of Computer Applications*, 47(11), 2012. 36

[73] Dan Jurafsky and James H. Martin. *Speech and Language Processing*. Pearson/Prentice Hall, 2009. 36

[74] Lars Nygaard Jörg Tiedemann. The opus corpus - parallel & free. `https://opus.nlpl.eu/`, 2004. Consulté le : 2022-11-18. 9

[75] Arkhipenko K., Kozlov I., Trofimovich J., Gomzin A., and Turdakov D. Skorniakov K. Comparison of neural network architectures for sentiment analysis of russian tweets. *Computational Linguistics and Intellectual Technologies : Proceedings of the International Conference "Dialogue 2016"*, 2016. 61, 62

[76] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *CoRR*, 2014. 67

[77] Andrew Kehoe and Renouf Antoinette. WebCorp : Applying the Web to Linguistics and Linguistics to the Web. In *WWW2002 Conference*, 2002. 14

[78] Andrew Kehoe and Matt Gee. New corpora from the web : making web text more 'text-like'. volume Volume 2. VARIENG, 2007. 11

[79] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online, November 2020. Association for Computational Linguistics. 55

[80] R. Khare, Cutting D., Sitaker K., and A. Rifkin. Nutch : A Flexible and Scalable Open-Source Web Search Engine. In *CommerceNet*, CN-TR-04-04, November 2005. 23, 32

[81] Adam Kilgarriff. Googleology is bad science. *Comput. Linguist.*, 33(1) :147–151, 2007. 10

[82] Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. A corpus factory for many languages. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). 14

[83] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014. 61, 67

[84] Diederik P. Kingma and Jimmy Ba. Adam : A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 70, 72

[85] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, New York, NY, USA, 2010. ACM. 8, 23

[86] Raymond Kosala and Hendrik Blockeel. Web mining research : A survey. *ACM SIGKDD Explorations Newsletter*, 2, 12 2001. 36

[87] Punit Kumar and Atul Gupta. Active learning query strategies for classification, regression, and clustering : A survey. *Journal of Computer Science and Technology*, 35 :913–945, 2020. 101

[88] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. 2001. 95

[89] Aurélien Lamercerie, David Rouquet, Valérie Bellynck, Christian Boitet, and Vincent Berment. Extraction de contenus sémantiques pour la vérification d'exigences systèmes. In *TextMine 2022 (EGC'22 - Atelier Fouille de Textes)*, Blois, France, Jan 2022. `hal-03789280`. 116

[90] Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohku. Sentiment analysis for low resource languages : A study on informal indonesian tweets. *Proceedings of the 12th Workshop on Asian Language Resource*, page 123–131, 2016. 62

[91] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. In Michael A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 255–258. MIT Press, Cambridge, MA, USA, 1998. 55

[92] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553) :436–444, 2015. 62

[93] Geoffrey Leech. *Linguistic Information from Computer Text Corpora*, chapter Introducing corpus annotation, pages 1 – 19. Routledge, London and NY, 1997. 86

[94] Vladimir Levenshtein. Binary codes capable of correcting deletions and insertions and reversals. pages 707–710, 1966. 28

[95] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994. 101

[96] Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. Assessing the Capacity of Transformer to Abstract Syntactic Representations : A Contrastive Analysis Based on Long-distance Agreement. *Transactions of the Association for Computational Linguistics*, 11 :18–33, January 2023. 117

[97] Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics. 92

[98] Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018. 92

[99] P. Liang, M. I. Jordan, and D. Klein. Learning from measurements in exponential families. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 641–648. ACM, 2009. 94

[100] Xiaofeng Liao and Zhiming Zhao. Unsupervised approaches for textual semantic annotation, a survey. *ACM Computer. Surv. 52*, 4, Article 66 (August 2019), 45 pages, 2019. 87

[101] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan and Claypool, 2012. 59

[102] Bing Liu. *Sentiment Analysis : Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015. 57

[103] Pengfei Liu, Shafiq Joty, and Helen Meng. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1433–1443, 2015. 92

[104] Qian Liu, Bing Liu, Yuanlin Zhang, Doo Soon Kim, and Zhiqiang Gao. Improving opinion aspect extraction using semantic similarity and aspect associations. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2986–2992. AAAI Press, 2016. 92

[105] Shao Lu. *Fuzzy Language in Literature and Translation*. Routledge, New York, NY, 2023. 4

[106] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, 2016. 100, 108

[107] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis.

In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, pages 142–150, 2011. 63, 79

[108] Sanjay Kumar Malik and SAM Rizvi. Information extraction using web usage mining, web scrapping and semantic annotation. In *2011 International Conference on Computational Intelligence and Communication Networks*, pages 465–469. IEEE, 2011. 37

[109] Otman Manad, Anna Pappa, and Gilles Bernard. A cleaning algorithm for noiseless opinion mining corpus construction. 10 2018. 5, 37, 45, 114

[110] Damon Mayaffre. Corpus et web-corpus. réflexion sur la corporalité numérique. *Cahiers de praxématique [En ligne]*, 54-55, 2010. 10

[111] T. McEnery and A. Wilson. *Corpus linguistics*. Edinburgh textbooks in empirical linguistics. Edinburgh University Press, 1996. 10

[112] Andrew J. McMinn, Yashar Moshfeghi, and Jose Joemon M. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 409–418, New York, NY, USA, 2013. ACM. 22

[113] Lisa Medrouk and Anna Pappa. Deep learning model for sentiment analysis in multi-lingual corpus. In *Neural Information Processing - 24th International Conference, ICONIP*, pages 205–212, 2017. 6, 55

[114] Lisa Medrouk and Anna Pappa. Do deep networks really need complex modules for multilingual sentiment polarity detection and domain classification ? In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2018. 6, 55

[115] Lisa Medrouk, Anna Pappa, and Jugurtha Hallou. Review web pages collector tool for thematic corpus creation. In Antonio Moreno Ortiz and Chantal Perez-Hernandez, editors, *CILC2016. 8th International Conference on Corpus Linguistics*, volume 1 of *EPiC Series in Language and Linguistics*, pages 274–282, 2016. 5, 6, 13, 24, 114

[116] Olga. Melekhova, M.-A. Abchir, Pierre. Châtel, Jacques. Malenfant, Isis. Truck, and Anna. Pappa. Self-adaptation in geotracking applications : Challenges, opportunities and models. In *Proceedings of the 2nd International Conference on Adaptive and Self-adaptive Systems and Applications*, pages 68–77, 2010. 1, 4

[117] Sylvie Mellet. Corpus et recherches linguistiques. *Corpus [En ligne], 1*, mis en ligne le 15 décembre 2003. 11

[118] Charles F. Meyer. *Annotating a corpus*, page 81–99. Studies in English Language. Cambridge University Press, 2002. 86

[119] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, June 2007. 62, 65

[120] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. 64

[121] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013. 63

[122] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media. pages 928–940, 11 2019. 96

[123] Daniel Myers and James W McGuffee. Choosing scrapy. *Journal of Computing Sciences in Colleges*, 31(1) :83–89, 2015. 37, 41

[124] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023. 93

[125] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means ? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *To be Published*, 2023. 93

[126] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10) :1345–1359, 2010. 85, 93, 96

[127] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2) :1–135, 2008. 55

[128] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2) :1–135, January 2008. 57, 59, 60

[129] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up ? sentiment classification using machine learning techniques. *CoRR*, cs.CL/0205070, 2002. 60

[130] Anna. Pappa. A 3-step algorithm for morphological disambiguation using untagged corpora. In *Proceedings of IC-AI'03 International Conference on Artificial Intelligence*, pages 921–926, Las Vegas, USA, June 2003. 1, 2

[131] Anna. Pappa. Etiquetage syntaxique automatique des parties du discours en français et en grec. In *Proceedings of 8th International Symposium of Social Communication and Applied Linguistics*, pages 512–517, Santiago, Cuba, January 2003. 1

[132] Anna. Pappa. Start : analyseur syntaxique de surface – présentation et évaluation. In *ATALA : EVANS : Méthodes et outils pour l'évaluation des analyseurs syntaxiques*, pages 18–24, Paris, May 2004. 1, 2

[133] Anna. Pappa. Robust tagging system for lexicon creation. In *Proceedings of the 2006 World Congress in Computer Science, Computer Engineering and Applied Computing*, USA, June 2006. 1, 2

[134] Anna. Pappa. Automatic acquisition of lexical entries with morphological features from non annotated corpora - (presentation). In *International NooJ Conference*, Univ. Autonomous of Barcelona, Spain, 7 to 9 June 2007. 1, 2

[135] Anna. Pappa. Constructing lexicon with morpho-syntactic features from untagged corpora. In *Proceedings of the 3rd International Conference on European Computing Conference*, 2009. Best Paper. 1, 2

[136] Anna Pappa. Litte_bot : le bot qui donne la réplique dans le style molière. In *Journée d'étude Technologies du Langage Humain et Accès Interactif à l'Information (JAII)*, pages 22–23, Paris, France, 2022. ©2022. 118

[137] Anna. Pappa, Gilles. Bernard, and Hind. Oukerradi. Détection automatique des frontières des phrases – un système adaptatif multi-langues. *ISDM (Informations, Savoirs, Décisions et Médiations)*, (13), February 2004. Permanent online Journal of Information and Communication Technologies. 1, 2

[138] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 64, 80

[139] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 108

[140] Resnik Philip and Elkiss Aaron. The linguist's search engine : An overview. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, ACLdemo '05, pages 33–36, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. 14

[141] Jakub Piskorski and Roman Yangarber. Information extraction : Past, present and future. In Thierry Poibeau, Horacio Saggion, Jakub Piskorski, and Roman Yangarber, editors, *Multi-source, Multilingual Information Extraction*

and Summarization, Theory and Applications of Natural Language Processing, chapter 2, pages 23–49. Springer Berlin Heidelberg, 2013. 14

[142] Jan POMIKÁLEK. *Removing Boilerplate and Duplicate Content from Web Corpora [online]*. Doctoral theses, dissertations, Masaryk University, Faculty of Informatics, Brno, 2011. 24

[143] Anbumunee Ponniah, Swati Agarwal, Sharanya Milind Ranka, and Shashank Madhusudhan. A transfer learning framework for annotating implementation-specific corpus. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2022. 94

[144] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*, chapter 1, page 342. O'Reilly Media, Inc., 2012. 86

[145] G. Qiu, J. Liu, B.and Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. In *Computational Linguistics*, volume 37, page 9–27, 2011. 92

[146] Sophie Raineri and Camille Debras, editors. *Corpora and Representativeness*, volume 19, 2019. 10

[147] Lance A Ramshaw and Mitchell P Marcus. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer, 1999. 90

[148] Amon Rapp, Lorenzo Curti, and Arianna Boldi. The human side of human-chatbot interaction : A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151 :102630, July 2021. 118

[149] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9) :1–40, 2021. 94

[150] Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. Active learning for part-of-speech tagging : Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, pages 101–108, 2007. 94

[151] Klinger Roman and Cimiano Philipp. Joint and pipeline probabilistic models for fine-grained sentiment analysis : Extracting aspects, subjective phrases and their relations. *Proc. IEEE 13th Int. Conf. Data Mining Workshops (ICDMW)*, page 937–944, Dec. 2013. 92

[152] D. Roth and W.-t. Yih. A linear programming formulation for global inference in natural language tasks. Technical report, Defense Technical Information Center, 2004. 94

[153] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Neurocomputing : Foundations of research. pages 696–699, 1988. 55

[154] Brucce Neves Dos Santos, Ricardo Marcondes Marcacini, and Solange Oliveira Rezende. Multi-domain aspect extraction using bidirectional encoder representations from transformers. *IEEE Access*, 9 :91604–91613, 2021. 92

[155] Christopher Schröder and Andreas Niekler. A survey of active learning for text classification using deep neural networks. *arXiv :2008.07267*, 2020. 94

[156] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3776–3783. AAAI Press, 2016. 118

[157] Burr Settles. Active learning literature survey. 2009. 94

[158] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008. 94

[159] Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 959–962, New York, NY, USA, 2015. ACM. 61

[160] Serge Sharoff. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus. Gedit*, 2006. 14

[161] Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Irina Fedulova, and Dmitry V. Dylov. Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 482–489, 2019. 94

[162] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In *International Conference on Learning Representations*, 2018. 94

[163] Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. Beyond fair pay : Ethical implications of nlp crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 3758–3769, Online, June 2021. Association for Computational Linguistics. 91

[164] Max Silberztein. Les outils informatiques au service des linguistes : présentation. *Langue française*, 203(3) :7–14, 2019. 1, 2

[165] Prerana Singhal and Pushpak Bhattacharyya. Borrow a little from your rich cousin : Using embeddings and polarities of english words for multilingual sentiment classification. In *COLING*, 2016. 65

[166] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–161. Association for Computational Linguistics, 2011. 61

[167] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics, 2013. 61

[168] Miroslav Spousta, Michal Marek, and Pavel Pecina. Victor : the web-page cleaning tool. In *4th Web as Corpus Workshop (WAC4)-Can we beat Google*, pages 12–17, 2008. 37

[169] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15 :1929–1958, 2014. 72

[170] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. 117

[171] Dumais Susan T. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1) :188–230, 2004. 64

[172] K.Thirumoorthy T. Mahara Jothi. A survey on web forum crawling techniques. *International Journal of Innovative Research in Science, Engineering and Technology*, 3, March 2014. 22

[173] Jose Camacho-Collados Tommaso Pasini. A short survey on sense-annotated corpora. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, page 5759–5765, 2020. 87

[174] Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *To be Published*, 2023. 93

[175] Nikolai Tschacher. Googlescraper. `https://github.com/NikolaiT/GoogleScraper`, 2015. 38

[176] Peter D. Turney. Thumbs up or thumbs down ? : Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424. Association for Computational Linguistics, 2002. 60

[177] Unknown. Llama : Large language models, 2023. Accessed : august 2023. 93

[178] Wouter Van Atteveldt, Mariken ACG Van der Velden, and Mark Boukes. The validity of sentiment analysis : Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2) :121–140, 2021. 87

[179] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017. 36

[180] Dominique Vaufreydaz. *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Theses, Université Joseph-Fourier - Grenoble I, Jan 2002. 10

[181] David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. On the usefulness of lexical and syntactic processing in polarity classification of twitter messages. *J. Assoc. Inf. Sci. Technol.*, 66(9) :1799–1816, sep 2015. 61

[182] David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. Supervised sentiment analysis in multilingual environments. *Information Processing & Management*, 53(3) :595–607, 2017. 60

[183] Oriol Vinyals and Quoc Le. A neural conversational model, 2015. 118

[184] Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. Combining recurrent and convolutional neural networks for relation classification. *CoRR*, abs/1605.07333, 2016. 55

[185] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, Suntec, Singapore, August 2009. Association for Computational Linguistics. 65

[186] Xin Wang, Yuanchao Liu, Chengjie Sun, Baoxun Wang, and Xiaolong Wang. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *ACL*, 2015. 61, 62

[187] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people ? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. 91

[188] Janyce Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740. AAAI Press, 2000. 59

[189] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational linguistics*, 30(3) :277–308, 2004. 60

[190] Wikipedia. Web scraping. `https://fr.wikipedia.org/wiki/Web_scraping`, 2023. Consulté le : 2023-07-16. 10

[191] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers : State-of-the-art natural language processing, 2020. 117

[192] Yew-Kwong Woon, Wee-Keong Ng, and Ee-Peng Lim. Web usage mining : Algorithms and results. In *Web Mining : Applications and Techniques*, pages 373–392. IGI Global, 2005. 37

[193] M Wynne, editor. *Developing Linguistic Corpora : a Guide to Good Practice.* Oxford : Oxbow Books., 2005. 88

[194] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 592–598, July 2018. 92

[195] François Yvon. Le modèle transformer : un "couteau suisse" pour le traitement automatique des langues. *Techniques de l'Ingénieur*, 2022. 117

[196] Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8(3) :338–353, 1965. 3

[197] Yanhong Zhai and Bing Liu. Extracting web data using instance-based learning. In *Proceedings of the 6th International Conference on Web Information Systems Engineering*, WISE'05, page 318–331, Berlin, Heidelberg, 2005. Springer-Verlag. 25

[198] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. A survey on aspect-based sentiment analysis : Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20, 2022. 87

[199] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Chi-Moon Lau. A c-lstm neural network for text classification. *CoRR*, abs/1511.08630, 2015. 69

[200] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems (NIPS)*, 2004. 60

[201] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. Attention-based lstm network for cross-lingual sentiment classification. In *EMNLP*, 2016. 65

[202] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *CoRR*, abs/1911.02685, 2019. 96

Acronymes

ML Machine Learning

AA Active Learning

TALN Traitement Automatique du Langage Naturel

DOM Document Object Model

ANN Artificial Neural Network

IE Information Extraction

LSE Linguist's Search Engine

WCM Web Content Mining

DNN Deep Neural Network

DBN Deep Belief Network

SSA Subjectivity Sentiment Analysis

ABSA Aspect Based Sentiment Analysis

NER Named Entity Recognition

HDR Habilitation à Diriger des Recherches

LSTM Long Short-Term Memory

CNN Convolutional Neural Network

RNN Recurrent Neural Network

Bi-CNN-LSTM Bidirectional Convolutional Long-Short Term Memory

Bi-LSTM-CNN-CRF Bidirectional Long Short-Term Memory - Convolutional Neural Network - Conditional Random Field

CRF Conditional Random Field

NLP Natural Language Processing

LLM Large Language Model

ConvNet Convolutional Network

PoS Part-of-Speech

Titre : Contributions pour la création des corpus et les modèles d'apprentissage profond pour les données textuelles multilingues

Résumé : Cette Habilitation à Diriger des Recherches (HDR) synthétise près d'une décennie de travaux en TALN. Elle met un accent particulier sur la création de corpus multilingues et thématiques, spécialement conçus pour l'analyse des sentiments et des aspects. Les méthodologies et outils développés pour créer ces datasets multilingues, sans bruit et issus d'avis d'utilisateurs, servent de base solide pour les expérimentations subséquentes. L'utilisation de réseaux neuronaux convolutifs hiérarchiques (ConvNet) et de réseaux neuronaux récurrents (RNN) permet de relever les défis de la prédiction de la polarité et de la classification thématique. La performance est améliorée grâce à l'architecture Bi-CNN-LSTM, qui combine des convolutions et une mémoire à long terme, atteignant une précision allant de 90% à 100% selon les expérimentations, et ce, sur des corpus multilingues non annotés. Des techniques d'apprentissage profond intégrées, telles que l'apprentissage par transfert et l'apprentissage actif au sein d'une architecture combinée Bidirectional Long Short-Term Memory - Convolutional Neural Network - Conditional Random Field (Bi-LSTM-CNN-CRF), sont utilisées pour l'annotation d'aspects, améliorant ainsi les performances des modèles, notamment dans des contextes où les données ou les langues sont sous-représentées. Les travaux présentés dans cette habilitation contribuent aux méthodes et aux pratiques en TALN, en s'appuyant sur des jeux de données sur mesure et des architectures de modèles sophistiquées pour surmonter des défis complexes en annotation sémantique et en analyse multilingue.

Title: Contributions to corpus creation and deep learning models for multilingual textual data

Abstract: This HDR synthesizes nearly a decade of research work in Natural Language Processing (NLP). It places a particular emphasis on the creation of multilingual and thematic corpora, specifically designed for sentiment and aspect analysis. The methodologies and tools developed for generating noiseless, multilingual datasets, sourced from user reviews, serve as a solid foundation for subsequent experiments. The use of hierarchical Convolutional Neural Networks CNN and Recurrent Neural Networks (RNN) addresses the challenge of polarity prediction and thematic classification. This performance is further enhanced by the Bi-CNN-LSTM architecture, which combines convolutions with long-term memory, achieving an accuracy ranging from 90% to 100% depending on the experiments, and this on non-annotated multilingual corpora. Integrated deep learning techniques, such as transfer learning and active learning within a combined Bi-LSTM-CNN-CRF architecture, are employed for aspect annotation, thus improving the model's performance, especially in contexts where data or languages are underrepresented. In summary, this habilitation contributes to the methods and practices in NLP, relying on tailor-made datasets and sophisticated model architectures to overcome complex challenges in semantic annotation and multilingual analysis.