Classification de transcriptions orales dans un contexte applicatif peu doté : application du TAL pour l'analyse de verbatim destinée à l'évaluation de l'acceptabilité d'une innovation

Emmanuelle Kelodjoue Nguemegne

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Informatique
Unité de recherche : Laboratoire d'Informatique de Grenoble

Classification de transcriptions orales dans un contexte applicatif peu doté : application du TAL pour l'analyse de verbatim destinée à l'évaluation de l'acceptabilité d'une innovation

Classification of interview transcripts in a low resource context: application of NLP for verbatim analysis for the evaluation of the acceptability of an innovation.

Présentée par :

EMMANUELLE KELODJOUE NGUEMEGNE

Direction de thèse :

Jérôme GOULIAN Maître de conférences, Université Grenoble Alpes	Directeur de thèse
Didier SCHWAB Maître de conférences HDR, Université Grenoble Alpes	Co-directeur de thèse

Rapporteurs :

IRIS ESHKOL-TARAVELLA
Professeur des Universités, UNIVERSITE PARIS 10 - NANTERRE
MATHIEU LAFOURCADE
Maître de conférences HDR, UNIVERSITE DE MONTPELLIER

Thèse soutenue publiquement le **5 octobre 2022**, devant le jury composé de :

DIDIER SCHWAB Maître de conférences HDR, UNIVERSITE GRENOBLE ALPES	Co-directeur de thèse
JERÔME GOULIAN Maître de conférences, UNIVERSITE GRENOBLE ALPES	Directeur de thèse
IRIS ESHKOL-TARAVELLA Professeur des Universités, UNIVERSITE PARIS 10 - NANTERRE	Rapporteure
MATHIEU LAFOURCADE Maître de conférences HDR, UNIVERSITE DE MONTPELLIER	Rapporteur
CATHERINE BERRUT Professeur des Universités, UNIVERSITE GRENOBLE ALPES	Présidente
ANNE VILNAT Professeur des Universités, UNIVERSITE PARIS- SACLAY	Examinatrice

Invités :

JEAN CAELEN

PASCAL PIZELLE

Résumé

Ce travail de thèse a pour objectif de répondre à une demande initiée par la société Ixiade dans l'optique d'automatiser une partie de ses méthodes d'analyse de contenu via les techniques du TALN. Dans cette thèse, nous nous intéressons à des données issues de transcriptions d'entretiens et des données en ligne qui ont été collectées dans le cadre d'études d'acceptabilité des innovations.

L'originalité de cette thèse est d'utiliser des méthodes d'amplification des données et des modèles de type Transformer sur de la classification de données de l'oral transcrit et des données issues d'une plateforme communautaire pour la langue française. Les contributions sont les suivantes : (1) Mise en œuvre d'une méthodologie de construction de données d'apprentissage et de test dans un contexte où les données sont indisponibles ; (2) Proposition d'une méthode d'extraction et de filtrage des contenus en fonction des classes à classifier reposant sur des patrons morpho-syntaxiques ; (3) Implémentation de différentes techniques d'amplification des données textuelles pour l'oral transcrit et du contenu en ligne ; (4) Comparaison des performances de différents modèles de langue contextualisés pour la langue française sur notre tâche de classification ; (5) Examen de l'apport de l'amplification des données dans un contexte peu doté.

Dans un premier temps, nous avons construit trois corpus d'apprentissage de nature différente. Le premier a été construit en utilisant des archives d'anciennes études réalisées, le deuxième a été construit à partir d'un autre corpus et le dernier provenait de projets d'études réalisés sur une plateforme. Pour cela, nous avons mis en place une procédure spécifique au contexte de travail chez Ixiade pour l'annotation des données dans le but de construire des corpus d'apprentissage et d'évaluation.

Dans un deuxième temps, nous avons sélectionné un corpus parmi nos corpus collectés pour implémenter une méthode d'extraction et de validation des phrases extraites. La méthode d'extraction a permis de produire des résultats satisfaisants, mais non suffisants pour être utilisés dans l'objectif d'agrandir nos corpus initiaux d'apprentissage. En outre, afin de limiter le taux d'erreurs produit par cette méthode,

nous avons également utilisé une méthode de filtrage sur les extractions faites par la première méthode. Les évaluations et les résultats des méthodes de filtrage n'ont pas permis d'augmenter significativement la précision dans l'extraction des phrases en fonction de nos classes.

Dans un troisième temps, nous nous sommes focalisés sur l'amplification de données et son apport pour la tâche de classification qui nous incombait. Pour cela, nous avons comparé les résultats produits par ces méthodes combinées à des architectures de type Transformer. Ces expérimentations ont permis de montrer l'apport non négligeable de l'amplification dans notre contexte de recherche.

Globalement, ces travaux ont permis de montrer l'intérêt des méthodes d'amplification dans un cadre où les données sont non disponibles et ouvert des perspectives dans cette tâche. S'agissant du recours au modèle de type Transformer choisi dans cette thèse, les modèles développés uniquement pour le français ont montré de bonnes performances par rapport aux modèles multilingues.

Abstract

This thesis work aims to respond to a request initiated by the company Ixiade. The request was to explore Natural language processing methods in order to propose a content classification tool. Two types of data were used throughout the study : interview transcripts and online data. Both came from studies carries out to assess the acceptability of an innovation.

This research work uses data amplification methods combined with Transformer-based-models to classify transcribed oral data and online data stemming from a community platform. The contributions are as follows : (1) Proposal of a methodology to build a training corpus in a context where data are unavailable ; (2) Proposal of a method for extracting and filtering textual content according to the classes to be classified based on morphosyntactic patterns ; (3) Implementing different textual data amplification techniques for transcribed speech and online content ; (4) Comparing the performance of different contextualized language models for the French language on our classification task ; (5) Examining the contribution of data amplification in a sparse context.

Firstly, we built three different training corpora. For this, we implemented a specific procedure for annotating the data.

Secondly, we selected a corpus among our three collected corpora to implement an extraction and filtering method. The extraction method produced satisfactory results but was not sufficient to be used to expand our initial training corpus. Furthermore, to limit the error rate produced by this method, we also used a filtering method on the extractions made by the first method. The evaluations and results of the filtering methods did not yield significant results.

Thirdly, we focused on data amplification and its contribution to the classification task we had to perform. For this purpose, we compared the results of different amplification methods combined with various transformer-based-architectures. These experiences have shown the significant contribution of amplification in our research context.

Overall, this work has shown the interest in amplification methods in a context

where data are unavailable and opened perspectives in this task. Regarding the use of the chosen transformer-based model in this thesis, the French models showed good performances compared to the multilingual model.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Contexte et demande de l'entreprise

Le travail de recherche que nous présentons dans ce document est une thèse initiée dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE), fruit d'un partenariat entre l'équipe de recherche GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole) du Laboratoire d'Informatique de Grenoble et la société Ixiade. Ixiade est une société spécialisée dans l'accompagnement de projets d'innovation d'entreprises. Cet accompagnement commence depuis la génération des idées jusqu'à la mise en place du produit sur le marché et fait appel à une multitude de compétences issues de métiers différents : le marketing, les sciences sociales, l'ingénierie, l'ergonomie, le design, la stratégie et la communication. Plusieurs méthodologies sont mises en œuvre afin de parvenir à la conception d'innovations en phase avec les futurs utilisateurs. Ixiade propose à ses clients des études qualitatives pour évaluer l'acceptabilité de leurs innovations. Par acceptabilité, nous entendons la propension des utilisateurs à accepter, voire intégrer les services ou produits innovants dans leurs usages. Ces études mobilisent la réalisation d'entretiens ou de focus groups [1] au cours desquels les potentiels utilisateurs sont interrogés sur les innovations. Le contenu recueilli à partir de ces canaux de collecte est ensuite analysé selon différentes méthodes. Cette analyse très spécifique au cadre de travail d'Ixiade consiste in fine à catégoriser les contenus textuels en fonction de trois grandes catégories (frein, motivation et condition). Ces catégories sont détaillées à la section 3.6 du chapitre 3. Pour le moment, cette analyse est manuelle et chronophage. Consciente des grosses ressources et du temps important allouer à cette tâche, Ixiade souhaite pouvoir automatiser sa

1. Groupe de discussion constitué en moyenne de 6 personnes et plus.

méthode d'analyse. Dans cette optique et après une recherche infructueuse de différents outils d'analyse qualitative pouvant répondre à son besoin, Ixiade a décidé d'initier ses travaux de recherche en traitement automatique des langues (TAL). Le TAL est une discipline qui s'intéresse à l'automatisation des processus d'analyse de données du langage. Elle a connu une forte croissance ces dernières années grâce notamment aux avancées des méthodes en intelligence artificielle. Ces méthodes ont permis de développer un très grand nombre d'applications informatiques dans des champs tels que le traitement du signal avec la reconnaissance automatique de la parole ou l'extraction d'informations avec la classification et catégorisation de documents. L'objectif était d'explorer le potentiel du TAL dans la reconnaissance des freins, des motivations et des conditions dans des contenus textuels. D'autre part, dans un souci d'évolution, Ixiade a également décidé de s'orienter vers un autre mode de collecte en ligne développant ainsi une plateforme communautaire avec une partie privée dédiée à ses études, mais également une partie publique ouverte à tous. In fine, l'entreprise souhaite se doter d'un outil capable d'analyser plus rapidement les données de ces études qualitatives.

1.2 Problématique de recherche

Cette thèse vise à proposer une solution de classification automatique issue du TAL à la demande soulevée par la société Ixiade. Néanmoins, elle pose de nombreux défis. Premièrement, les contenus à analyser sont de deux types : des données de l'oral transcrit et des données collectés en ligne. Le système de classification doit être autant performant sur la tâche de détection des freins, des motivations et conditions pour les deux types de données (données de l'oral transcrit et données collectées en ligne). Deuxièmement, nous nous situons dans un contexte où nous ne disposions pas de corpus d'apprentissage pour les deux types de données. Bien qu'il existe en TAL des travaux qui se sont focalisés sur l'analyse de données orales transcrites tels que les débats politiques (Thomas et al., 2006; Abercrombie et Batista-Navarro, 2018), les conversations téléphoniques (Cailliau et Cavet, 2010), les messages oraux (Camelin et al., 2006), des entretiens face-à-face de personnels (Parmar et al., 2018) issus de diverses industries (médical, construction, etc.) ou des entretiens face-à-face d'éducateurs et d'étudiants sur l'apprentissage personnalisé (McHugh et al., 2020), aucun travail en TAL à notre connaissance ne s'était encore penché sur la problématique de classification de contenu textuel dans le cadre de l'évaluation de l'acceptabilité d'innovations sur des corpus issus de l'oral transcrit ou collectés en ligne.

1.3 Approche et plan du manuscrit

Dans ce travail de recherche, nous proposons d'examiner l'apport de différentes méthodes du TAL pour répondre à notre tâche de classification. Dans un premier temps, nous proposons une méthode d'extraction des freins et des motivations reposant sur des patrons morpho-syntaxiques pour amplifier notre jeu données. Ensuite, nous proposons d'étudier l'apport de différentes méthodes d'amplification des données sur la tâche de classification qui nous incombe. Finalement, nous examinons et comparons différents modèles préentraînés développés pour le français sur nos données amplifiées et évaluons leur apport sur les résultats de prédictions.

Cette thèse est organisée en deux parties principales : l'état de l'art (chapitres 2, 3 et 4) et nos contributions (chapitres 5, 6, 7 et 8). Nous détaillons le contenu de chacun des chapitres ci-dessous :

État de l'art :

Chapitre 2. Ce chapitre présente le contexte général de nos travaux de recherche. Nous introduisons les termes spécifiques liés à l'innovation et explicitons la démarche de travail appliquée chez Ixiade dans le cadre d'une étude de la collecte à la présentation des résultats au client.

Chapitre 3. Ce chapitre introduit l'état de l'art de l'analyse d'opinion, les méthodes utilisées, les ressources et les challenges que le domaine pose. Ce chapitre se termine par une sous-partie où nous explicitons et détaillons à l'aide d'éléments pris dans la littérature les concepts de frein, motivation et conditions.

Chapitre 4. Ce chapitre introduit l'état de l'art des méthodes de représentations des textes en TAL en s'attardant particulièrement sur les derniers modèles de langues en général et plus spécifiquement ceux développés pour la langue française.

Contributions :

Chapitre 5. Ce chapitre s'attarde sur les méthodologies mises en place pour construire trois différents corpus pour notre tâche de classification. Une sous-partie est dédiée à la méthode d'extraction explorée pour catégoriser les freins et motivations à l'aide de patrons morphosyntaxiques.

Chapitre 6. Ce chapitre présente nos contributions pour l'amplification des données pour la langue française dans le cas où l'on dispose de peu de données. Il est divisé en deux parties : un état de l'art et une partie amplification sur nos données d'apprentissage introduites au chapitre 5.

Chapitre 7. Ce chapitre présente les résultats nos travaux de classification. Il contient les résultats des expérimentations faites.

Chapitre 8. Ce chapitre présente nos conclusions et perspectives pour la suite de notre travail.

Première partie

État de l’art

Chapitre 2

Contexte général

2.1 Introduction

Ce deuxième chapitre a pour but de présenter le contexte et le positionnement de nos travaux de recherche. Ceux-ci sont étroitement liés à l'innovation, au marketing, à la psychologie, à la sociologie et aux études d'usages. La première partie de ce chapitre s'attache à définir et à caractériser l'innovation à partir d'éléments pris dans la littérature scientifique. Dans la deuxième partie, nous apportons des éléments de réponses avancés par la littérature au problème des échecs des innovations. Ensuite, la troisième partie introduit la notion d'*acceptabilité* et son importance dans les études d'usages. Enfin, dans la dernière partie du chapitre, nous exposons les enjeux et la direction de notre travail.

2.2 Innovation : définition et caractérisation

2.2.1 Qu'est-ce qu'une innovation ?

L'achat d'un produit d'innovation n'est pas dénué de sens. Il suffit d'observer le marché actuel pour constater que chaque jour de nouveaux produits sont proposés aux consommateurs, mettant en avant les différents bénéfices qu'ils peuvent procurer à ces derniers : gain de temps, qualité de vie, confort garanti, produits éthiques ou encore écologiques. De même, de nombreuses initiatives gouvernementales comme privées se multiplient pour stimuler l'innovation (en investissant massivement dans la recherche ou dans des organismes parapublics) mais également encourager et accompagner les entreprises, les entrepreneurs ou encore les startups à innover (Soler et al.). Nous pouvons notamment citer le programme *launchpad* d'Amazon qui

accompagne et aide les entrepreneurs à lancer des marques innovantes. D'autres facteurs tels que « l'évolution des métiers, le besoin constant d'améliorer l'existant, dominer le marché et récupérer des parts de marchés non négligeables à leurs concurrents, l'obsolescence des produits liée à l'évolution technologique » (Buttard, 2018) imposent également aux entreprises d'innover constamment. Les entreprises qui se démarquent des autres sont donc celles qui réussissent à supplanter leurs concurrents en proposant des offres uniques sur le marché et en phase avec les tendances.

L'innovation peut être définie comme "quelque chose" de nouveau présentant une rupture avec le passé et le présent et qui est présenté aux individus. Ce "quelque chose" peut-être un nouveau produit de consommation ou de nouvelles méthodes de production ou organisationnelles. Néanmoins, nous retiendrons pour la suite la définition du manuel d'Oslo de l'OCDE [1] (OECD, 2005) qui nous semble plus complète. Le manuel d'Oslo définit l'innovation comme « la mise en œuvre - la commercialisation ou l'implantation, par une entreprise et pour la première fois, d'un produit (bien ou service) ou d'un procédé (de production) nouveau ou sensiblement amélioré, d'une nouvelle méthode de commercialisation ou d'une nouvelle méthode organisationnelle dans les pratiques d'une entreprise, l'organisation du lieu de travail ou les relations avec l'extérieur » (OECD, 2005).

Quatre domaines d'innovations sont ainsi distingués :

- l'innovation de produits (biens et services) ;
- l'innovation de procédés ;
- l'innovation organisationnelle ;
- l'innovation de marketing.

Ces quatre domaines sont regroupés en deux grandes composantes :

1. l'innovation technologique [2] qui va concerner les produits et les procédés ;
2. l'innovation non technologique qui va englober l'organisation et le marketing.

Dans la suite de notre rédaction, nous nous intéresserons davantage aux innovations technologiques. Ce choix repose sur le fait que nos travaux ont une finalité applicative pour les entreprises de ces secteurs.

1. Organisation de coopération et de développement économiques.
2. L'innovation technologique de produit est définie par le manuel d'Oslo comme « la mise au point et la commercialisation d'un produit plus performant dans le but de fournir au consommateur des services objectivement nouveaux ou améliorés ».

2.2.2 Typologie des innovations

Toutes les innovations n'ont pas le même impact sur la société et encore plus sur la croissance économique d'un pays. Certaines vont apporter des changements majeurs tandis que d'autres vont juste améliorer l'usage. Partant de là, deux principaux types d'innovations sont rencontrés dans la littérature : les innovations radicales et les innovations incrémentales (Garcia et Calantone, 2002; Silberzahn et al., 2007; Nelson, 2011; Buisson et Silberzahn, 2005).

- **Innovations radicales ou de rupture :** les innovations radicales incarnent une nouvelle technologie débouchant sur une nouvelle infrastructure de marché. Souvent, elles ne répondent pas à une demande reconnue, mais créent au contraire une demande auparavant non reconnue par le consommateur. Cette demande engendre de nouvelles industries avec de nouveaux concurrents, de nouvelles entreprises ou de nouveaux canaux de distribution et de nouvelles activités de marketing. Ces innovations peuvent avoir une portée tant à l'échelle mondiale (*le World Wide Web*, *le télégraphe*, *le siège enfant*[3]) qu'industrielle (*le World Wide Web*, *le Walkman de Sony*) ou encore à l'échelle de marché (*les distributeurs automatiques*) (Nelson, 2011). Elles sont également rares, car leurs coûts de développement sont plus importants et les processus longs.

- **Innovations incrémentales :** elles correspondent à des produits qui apportent de nouvelles caractéristiques, de nouveaux avantages ou des améliorations au produit ou service existant sans en modifier les caractéristiques qui le définissent (Nelson, 2011). Les clients visés restent les mêmes tout comme le produit également. Par exemple, *les smartphones (l'iPhone 13*[4] *qui est une amélioration de l'iPhone 12)*, *les automobiles*. L'innovation incrémentale est la plus répandue, car les risques paraissent minimes et les gains intéressants. Les entreprises font le choix de miser sur ce type d'innovations en raison de la présence d'un marché existant. Elles préfèrent jouer la carte de la prudence que de s'aventurer dans des développements incertains de produits nouveaux (Buisson et Silberzahn, 2005) sans disposer des garanties de succès immédiats.

3. C'est en 1964 qu'est commercialisé le premier siège auto dos à la route, en collaboration avec le constructeur automobile Volvo.

4. L'iPhone ou encore iPhone 2G est une innovation radicale en 2007 et devient incrémental ensuite avec la sortie de divers modèles.

2.3 Les risques liés à l'innovation

2.3.1 L'échec en innovation.

Dans un contexte accru d'hypercompétition, l'innovation bien qu'incontournable de nos jours n'est pas un processus sans risque. Elle comporte toujours une variable "échec" qui peut freiner son développement et qu'il est crucial de prendre en compte, et d'étudier dès lors qu'on se lance dans un processus de conception d'une innovation. L'innovation est un parcours parsemé d'incertitudes. De ce fait, le risque n'épargne aucune entreprise. Bien que les produits innovants participent à 32% du chiffre d'affaires des entreprises et à 31% de leur profit (Gotteland et Haon, 2007), le lancement de nouveaux produits a toujours constitué un facteur de risque pour les entreprises. S'il est difficile de chiffrer le *taux d'échec des innovations*[5] toutes confondues qui entrent sur le marché chaque année, Gotteland et Haon (2005) déclarent en 2005 que « 95% des nouveaux produits de grande consommation lancés sur le marché nord-américain échouent ». S'agissant du marché européen, ce taux d'échec est estimé à 90% (Gotteland et Haon, 2005, 2007). Dans ce cadre précis, nous parlerons plutôt d'échec commercial[6]. Nous pouvons citer l'exemple du produit **Juicero**[7] dont une image est présentée en 2.1.

FIGURE 2.1 – Juicero : machine à jus de fruits (Raoul, 2017).

5. Néanmoins, comme le rappelle Loeser (2019), il convient de différencier dans le terme *innovation*, ce qui se réfère aux idées brutes et aux produits commercialisables. Dans notre contexte, nous sommes plus intéressés par les produits commercialisables.

6. L'échec commercial désigne l'arrêt d'un produit ou d'un concept. Cet arrêt fait suite à la constatation de la faible rentabilité du produit et de ces volumes de ventes qui ont affectés économiquement l'entreprise (Cusin, 2008).

7. https://www.objetconnecte.com/juicero-femerture-jus-0409/

Juicero, lancé en 2013 par la start-up du même nom était une machine à jus de fruits connectée permettant d'obtenir en quelques secondes du jus frais. La start-up a donc mis en œuvre des poches permettant de conserver les fruits et légumes au frais. Ces poches devaient être placées dans la machine et celle-ci donnait un bon verre de jus de fruit. Le principe d'utilisation du Juicero était assez similaire à celui des machines à café Nespresso avec leurs capsules. Cependant, la machine était vendue à 700$. En outre, il fallait encore débourser entre 5$ et 7$ pour se procurer les fameuses poches tout en sachant qu'une seule poche produisait un seul verre de jus. Dès son lancement sur le marché, le prix de la machine suscite de vives critiques. Malgré la tentative de la start-up de baisser le prix de leur produit, les critiques s'intensifient, amenant à un échec commercial et au retrait du produit du marché. L'entreprise ne survit pas à cette débâcle et ferme ses portes en 2017 (Raoul, 2017).

2.3.2 Les raisons de l'échec

À travers l'exemple précédent, nous comprenons donc que le succès d'un produit innovant ne serait garanti du seul fait qu'il présente une ou des caractéristiques techniques inédites ou innovantes. Ainsi, les utilisateurs ne jouent pas un rôle passif dans l'adoption des produits innovants (Dupré, 2016). Toute innovation quel que soit son type doit faire l'objet d'une appropriation par les utilisateurs. En somme, une mauvaise connaissance de l'utilisateur et des usages peut causer l'arrêt brutal d'un projet d'innovation. Néanmoins, cette raison n'est pas la seule qui peut conduire à l'échec d'une innovation lancée sur le marché. D'autres peuvent être évoquées (Soler et al.) telles que les suivantes :

- une mauvaise définition du produit ou de service ;
- une connaissance insuffisante du marché ;
- un manque de culture de l'innovation ou de la stratégie ;
- une mauvaise prise en compte de l'écosystème ;
- un manque de méthodologie et de compétence ;
- des biais psychologiques et des idées reçues ;
- une erreur de ciblage dans le développement [8] ;
- une mauvaise gestion/management ;
- un mauvais timing ;
- une estimation erronée de la valeur marchande du produit.

8. le produit rencontre dans son parcours de vie sur le marché différents profils de consommateurs : les techno-enthousiastes, les technophiles, les pragmatiques, les conservateurs et les réfractaires (Rogers, 1962).

2.4 Acceptabilité

2.4.1 Pourquoi évaluer le potentiel d'une innovation ?

Le développement de produits nouveaux est devenu une activité centrale pour la performance à long terme d'une entreprise. Cependant, l'innovation de produit comme de service comporte toujours une part de saut dans l'inconnu, donc de risque. L'innovation n'est pas toujours synonyme de succès malgré les nouvelles technologies ou les gains qu'elle peut apporter aux consommateurs. En conséquence, les variables « échec » et « utilisateur » sont des points cruciaux à prendre en compte dès que l'on se penche sur le processus de conception d'une innovation. Ainsi, une innovation bien que technologiquement attrayante peut se heurter à la résistance d'un marché mal analysé, non prêt à l'innovation proposée, ou encore pour qui, l'innovation ne fait quasiment pas sens. Aussi, il convient donc d'inclure dans tout processus de conception d'innovation une part consacrée à l'analyse de l'impact de l'innovation et des usages qu'elle créé. Cette partie d'analyse doit pouvoir s'appuyer sur des méthodes robustes et éprouvées qui place l'utilisateur au cœur du processus d'innovation. L'utilisateur joue ainsi un rôle prépondérant dans la validation et la diffusion des innovations dans son environnement personnel comme professionnel. Au cœur de ce processus d'innovation, plusieurs notions sont à considérer telles que l'utilisabilité, l'ergonomie, les usages et l'acceptabilité. Cette dernière notion est primordiale dans « la décision des consommateurs d'adopter ou non un produit ou service innovant » (Soler et al.).

2.4.2 L'acceptabilité : Définition

S'intéresser à l'utilisabilité, la satisfaction ou encore l'efficience d'un produit innovant sans s'assurer que celui-ci soit acceptable est une erreur fondamentale. Par exemple, « la technologie pour produire les pâtes à gâteaux toutes faites que nous retrouvons dans les supermarchés aujourd'hui existaient déjà dans les années 70. Cependant, le produit en lui-même n'était pas commercialisé en raison du contexte de l'époque. Les femmes qui sont la cible principale de ce type de produit ne travaillaient pas autant qu'aujourd'hui. Ainsi, la promesse du gain de temps et la praticité qu'apportait ce produit n'était pas entendable à l'époque ou même acceptable en raison du contexte. Aujourd'hui, cela est beaucoup plus normal, voire acceptable, car les femmes travaillent beaucoup plus [9] ». Cet exemple illustre parfaitement le constat que nous pouvons avoir de bonnes idées ou encore un concept *a priori* qui

9. https://alphonse.io/ux-et-acceptabilite-des-innovations/

a tout pour plaire, mais qui est en déphasage avec son époque.

L'acceptabilité renvoie au « degré d'intégration [10] et d'appropriation d'un objet dans un contexte d'usage » (Barcenilla et Bastien, 2009). Nous pouvons même parler d'un ensemble de conditions qui poussent un individu à adopter et à intégrer un objet à son quotidien. Comme tout processus, celui de l'acceptabilité comporte des phases : trois en l'occurrence. Partant de là, la trajectoire d'usage d'une innovation débute avec l'acceptabilité *a priori* (Pasquier, 2012), puis se poursuit avec l'acceptation pour se terminer par l'appropriation.

L'acceptabilité *a priori* est la première phase et débute au moment où l'utilisateur entend parler de l'innovation pour la première fois jusqu'à ses premiers essais réalisés avec l'innovation (Pasquier, 2012). C'est également à cette phase que l'utilisateur construit ses propres représentations de l'usage de l'objet d'innovation en fonction des premières descriptions dont il a eu connaissance.

La phase d'acceptation est la deuxième phase. La phase d'acceptation correspond à un processus psychologique résultant de l'adoption ou du rejet d'un produit (Dupré, 2016). Elle démarre dès les premières utilisations de l'innovation. C'est à cette phase que l'utilisateur utilise l'innovation. Des études d'usages peuvent être menées parallèlement pour recueillir les perceptions ou les éléments de satisfaction ou de non-satisfaction de l'utilisation de l'innovation.

La dernière phase concerne **l'appropriation** et correspond à l'usage de l'innovation réalisé par l'usager. Cette phase n'a de temps défini.

2.5 Méthodologie et objectifs applicatifs

Cette section est consacrée à la méthodologie de travail utilisée chez Ixiade dans le cadre des études d'usage qu'elle réalise.

2.5.1 Historique

Lorsque nous nous retrouvons en phase de conception d'une innovation, la question de son succès est une problématique des plus complexes. Pour anticiper ainsi tout risque d'échec d'un produit innovant et garantir ses chances de succès, plusieurs entreprises se sont spécialisées dans l'accompagnement de projets d'innovation. Ixiade, partenaire de ces travaux de recherche en est une. Elle accompagne chacun de ces clients dans son projet d'innovation en utilisant et en développant

10. *L'intégration correspond « à la manière dont le produit, ou système technique, s'insère dans la chaîne instrumentale existante et dans les activités de l'utilisateur, et comment il contribue à transformer ces activités. » (Barcenilla et Bastien, 2009).*

des méthodes qui intègrent l'expérience utilisateur. Cet accompagnement intègre différentes phases dont une phase d'étude (études d'usages, de marketing ou de sociologie). La phase d'étude d'usages essaie de répondre aux questions suivantes :

- Comment les individus s'approprient un objet ?
- Qu'est-ce qui pousse les individus à utiliser une nouvelle technologie ou service ?
- Comment prédire l'utilisation d'un objet ?

Ces questions entrent communément dans le cadre d'études d'usage et non d'opinion classique (sondages, etc.). Une étude d'usage a pour objectif principal d'évaluer les réalités d'utilisation du produit par l'utilisateur. Nous nous situons ainsi non pas dans une étude d'opinion classique, mais dans une évaluation dans un contexte d'utilisation ou de projection de l'utilisation des produits présentés. Ces études d'usage se traitent à la fois de manière qualitative (pour évaluer le degré d'acceptabilité d'un produit innovant, les comportements et attitudes) et quantitative (pour avoir des données chiffrées ou valider les hypothèses émises en phase qualitative). Les études qualitatives sont donc celles qui sont les plus prisées pour évaluer l'acceptabilité du fait qu'elles permettent notamment de s'intéresser aux freins ou aux motivations qui rendent le produit acceptable ou pas pour l'utilisateur avant usage. Dans la suite de notre travail, nous nous intéresserons davantage au cadre qualitatif, car nos travaux s'insèrent dans cet axe. Une étude qualitative va ainsi intégrer divers outils de conception, de collecte, de traitement et d'analyse.

2.5.2 Outils pour une étude qualitative

Avant toute proposition méthodologique, un diagnostic des besoins du client est établi avec lui afin de relever au cours de cet échange, ses incertitudes et ses doutes vis-à-vis de son projet d'innovation. Sont également discutés avec le client lors de cette phase de contact, les cibles auxquelles son produit innovant s'adresse ou pourrait s'adresser. Une fois le diagnostic des besoins établi, une proposition méthodologique est faite au client en fonction de l'échange ou des échanges précédents. Cette proposition méthodologique dépendra également de la maturation du projet, car certains porteurs de projets peuvent déjà avoir une bonne connaissance de leur concept, d'autres en revanche sont beaucoup plus en amont de leur projet et non qu'une vague idée de leur concept. Pour ces cas particuliers, un accompagnement spécifique peut-être proposé (session de créativité ou d'idéation [11], etc.) pour mieux formaliser leur concept (*Dans quel contexte émerge-t-il ?, À qui s'adresse-t-il ?, Le*

11. C'est un processus qui consiste à la production d'idées au travers de différentes sessions de groupe.

concept, qu'est-ce que c'est ?, À quoi sert-il ?, Comment s'utilise-t-il ?, Quels sont les principes de fonctionnement du concept ?).

Outils de conception

Toute étude qualitative repose sur un guide d'entretien ou d'animation. Un guide d'entretien ressemble à un mémento et a pour rôle de structurer l'entretien. Il permet également de garantir la traçabilité des questions posées. C'est ce guide qui est ensuite utilisé dans le cadre de l'étape de collecte d'information. Quant au guide d'animation, il reprend le déroulement d'un focus group et les questions auxquelles les participants seront appelés à répondre. Pendant la phase d'élaboration du guide, peut aussi intervenir la phase de recrutement des profils ou des futurs utilisateurs qui sont la cible du produit.

Le guide d'entretien ou d'animation contient toujours une partie introductive qui présente le concept. Celle-ci peut se faire via l'utilisation d'illustrations, de scénarios d'usages ou de vidéos. Une fois le recrutement réalisé et le guide prêt, les entretiens peuvent être réalisés. Les entretiens font partie de ce qu'on appelle les outils de collecte.

Outils de collecte

Les principaux outils de collecte pour une étude qualitative sont les entretiens face-à-face et les focus group.

Les entretiens face-à-face ou *individuels* : ils sont de type semi-directifs [12] et réalisés auprès d'utilisateurs potentiels et permettent de déterminer si l'objet technologique prend du sens. Au cours de ces entretiens, les individus peuvent tester le prototype de l'objet en cours de conception, mais cette interaction n'est pas systématique.

Les focus group ou *les tables rondes* : ils se composent d'un groupe de discussion (6 à 12 personnes en moyenne) qui a pour objectif de fournir des informations via l'interaction des membres du groupe par rapport à un sujet, un concept ou un service (Kohn et Christiaens, 2014).

Au cours de l'entretien, l'objectif est de laisser s'exprimer l'utilisateur sur les difficultés ou facilités qu'il pourrait rencontrer ou qu'il rencontre lors de l'utilisation de l'outil proposé. De plus, l'utilisateur dispose uniquement des informations nécessaires pour l'amener à se projeter dans l'utilisation de l'objet. Il pourrait même

12. Entretien semi-directif : dans ce cadre, l'interviewer possède un guide d'entretien et pose des questions ouvertes sur un sujet particulier.

prendre en main l'objet si des maquettes sont amenées durant l'entretien. En outre, les questions sont rédigées et posées de telle sorte à ne pas induire les réponses des utilisateurs (par exemple, « *Que pensez-vous de ...* », « *Que vous évoquent ...* », « *Spontanément, à quoi cela vous fait-il penser ?* »).

Outils de traitement

Les entretiens ou focus group sont toujours enregistrés. Une fois enregistré, ils doivent être transcrits et mis en forme pour être analysés. Les enregistrements des entretiens se font soit en interne par les chargés d'étude ou via l'utilisation d'un outil de transcription automatique (par exemple, *Happyscribe* [13]), soit en externe en faisant appel à un service de transcription. Dès que les transcriptions sont terminées et révisées, la phase d'analyse peut alors commencer.

Outils d'analyse

Tout processus de collecte de données mène à une étape d'analyse. En ce qui concerne l'analyse qualitative de données, elle peut s'avérer rigoureuse, longue et fastidieuse. La plupart des moyens et techniques qualitatifs ne sont pas faciles à décrire tant il en existe. Ainsi, en fonction de l'objectif de l'étude et du type des données, différentes techniques peuvent être mobilisées. Nous pouvons tout de même énumérer l'analyse thématique (Hernandez et Grau, 2002) ou encore l'analyse lexicale (Lahlou, 1994). Dans notre contexte, les retranscriptions des entretiens sont analysées en utilisant une méthode qualitative spécifique **CAUTIC** ® (Pizelle et al., 2014). Chaque verbatim [14] est ainsi associé à une classe en fonction des critères **CAUTIC** ® ou d'autres critères (frein, motivation ou condition). Ce travail est généralement long et fastidieux. Dans la suite, nous présentons la méthode de travail **CAUTIC** ®.

2.5.3 Évaluation de l'acceptabilité en fonction de critères psychologiques

La méthode de Conception Assistée par l'Usage pour les Technologies, l'Innovation et le Changement - CAUTIC ® développée par le sociologue Philippe

13. https://www.happyscribe.com/business

14. Dans notre contexte de travail, les données analysées par les chargés d'études sont appelés des verbatims. Un verbatim correspond à la « reproduction mot pour mot de données verbales » (Poland, 1995). Les mots écrits sont ainsi la réplique exacte des paroles enregistrées.

Mallein (Pizelle et al., 2014) est centrale dans le processus de conception des innovations. L'objectif de cette méthode est « d'évaluer le degré d'acceptabilité d'une innovation, en vue de formuler des voies d'amélioration à destination des porteurs de projet » (Pizelle et al., 2014). En d'autres termes, c'est une méthode d'analyse qualitative qui va permettre d'analyser pourquoi et comment les futurs utilisateurs vont accepter ou pas l'innovation dans leur vie quotidienne. Pour ce faire, CAUTIC ® s'appuie sur le discours des individus recueillis dans le cadre d'entretiens face-à-face ou de focus group. Cette méthode est structurée en quatre grands niveaux et 21 critères appelés *critères d'acceptabilité* (Pizelle et al., 2014) :

- l'**assimilation aux savoir-faire** (**Compréhension**) permet d'évaluer la possibilité pour l'utilisateur d'assimiler la technique nouvelle à ses savoir-faire techniques coutumiers ;
- l'**association ou l'Affiliation aux pratiques courantes** (**Pratiques**) évalue la possibilité pour l'utilisateur d'intégrer l'innovation dans ses pratiques habituelles ;
- l'**appropriation à l'identité socioprofessionnelle** (**Identité**) identifie la possibilité pour l'utilisateur d'approprier l'innovation à son identité privée et professionnelle ;
- l'**adaptation à l'environnement** (**Environnement**) analyse la possibilité pour l'utilisateur d'adapter l'innovation à son environnement privé ou professionnel.

Une fois les discours transcrits, ces derniers sont structurés en ces quatre niveaux et leurs sous-critères. Partant de là, il est important de pouvoir maintenant identifier ce qui constitue dans chacun de ces grands processus un *frein*, une *condition* ou une *motivation*, c'est-à-dire ce qui va venir en défaveur, faveur sous condition ou faveur de l'assimilation, l'association, l'appropriation et l'adaptation.

Pour conclure cette partie, l'analyse du discours recueilli dans le cadre d'entretiens face-à-face ou focus group est réalisée suivant deux logiques de classification. La première consiste à regrouper les éléments du discours que nous appellerons *verbatim* selon 4 niveaux principaux et leurs critères (21) tirés de la méthodologie CAUTIC ®. La deuxième consiste à reclassifier ces éléments identifiés en trois nouvelles classes (**Frein, Motivation, Motivation sous condition** ou FMC). Ces méthodes de collecte et d'analyse permettent à Ixiade de recueillir les avis des futurs ou potentiels utilisateurs sur la technologie développée par l'entreprise, d'identifier les points faibles dans le parcours utilisateur et les zones d'optimisation, etc. Par ailleurs, cette action permet également à Ixiade de s'assurer que les fonctionnalités du produit développé par l'entreprise correspondent bien aux attentes des

utilisateurs. À partir de là, Ixiade peut donc formuler des recommandations d'amélioration du produit à leurs clients sous forme de livrable. La figure 2.2 illustre le processus d'une étude.

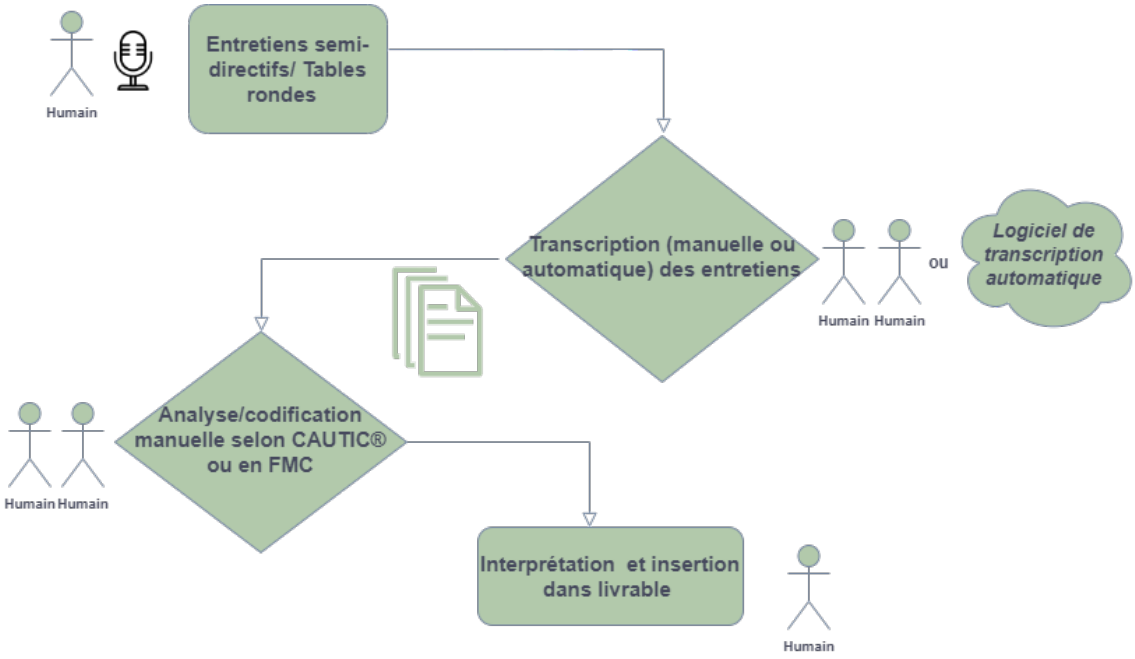


FIGURE 2.2 – Processus d'une étude.

2.6 Positionnement et enjeu

2.6.1 Constat

Les discours collectés lors des entretiens qui sont réalisés individuellement ou en groupe fournissent des informations riches et précieuses. Mais, à l'heure actuelle, la préparation et l'analyse de ces derniers sont très chronophages, car il faut non seulement retranscrire mot à mot les entretiens, mais également passer en revue l'ensemble des verbatims. Il faut également noter que plus les verbatims sont longs, plus la classification est fastidieuse. De plus, il peut y avoir un biais humain en fonction de chaque analyste. Pour un même verbatim, la classification peut varier d'une personne à l'autre en fonction de la perception de chaque chargé d'études. En outre, la quantité du contenu à analyser est appelée à croître (*via la plateforme*

Yoomaneo [15]), rendant la réalisation d'une analyse manuelle de plus en plus longue et difficile faute d'outils automatiques.

Cela ne peut donc pas être satisfaisant à long terme pour deux raisons :

— pour certaines études, les transcriptions ne sont pas forcément analysées dans leur globalité, donc, certaines informations ou certains signaux faibles peuvent ne pas être pris en compte lors de l'analyse manuelle ;

— d'un autre côté, les analyses sont stockées séparément selon les différents projets rendant la capitalisation sur les anciennes impossible : dès qu'un projet est fini, on archive les données. Or, pour l'analyse automatique, il faut un grand volume de données.

Aperçu de l'itinéraire d'analyse :

Nous nous situons dans une étude qualitative dont l'objectif est d'évaluer le degré d'acceptabilité d'un produit innovant qui a été présenté à un utilisateur dans le cadre d'un entretien. Pour analyser le discours de l'utilisateur après transcription, le chargé d'étude mène une classification selon deux niveaux spécifiques (voir paragraphe 2.5.3). Au premier niveau, il s'agit d'assigner à chaque verbatim qu'il va juger pertinent un critère selon la grille de la méthode CAUTIC ®. En fonction des critères généraux assignés, le verbatim peut être encore annoté en fonction des sous-critères du critère général sélectionné. Après cette première phase, il s'agit ensuite de déterminer l'axe (frein, motivation ou condition) sur lequel se positionne le verbatim. Les exemples suivants illustrent des verbatims recueillis dans le cadre d'entretiens menés pour différentes études.

Évaluation d'un tableau électrique

Verbatim (1) : [*La fonction différentielle intégrée est aussi intéressante.*]

15. Yoomaneo est une application destinée à celles et ceux qui souhaitent partager leurs opinions sur des projets innovants. Ces opinions sont ensuite recueillies via la plateforme et analysées par les chargés d'études. https://www.yoomaneo.com/fr

— Catégorie Cautic : Niveau 1 → **Assimilation**

  — critère 1.4 L'utilisateur peut-il trouver et choisir les fonctions qui l'intéressent dans le concept ?

— Catégorie FMC : **Motivation**

Interprétation : L'utilisateur comprend bien le produit présenté et est capable d'identifier les fonctions qui l'intéressent. Le verbatim est donc considéré comme une "motivation" (Présence d'éléments subjectifs à caractère positifs : "intéressante".)

Évaluation d'un tableau électrique

Verbatim (2) : [*je pense qu'elle est plausible, mais il ne faut pas que ce soit... un... un produit, euh... Comment ? Il faut que ce soit un produit qui soit ouvert. Il ne faut pas qu'il soit... dédié à X... Enfin, qu'il ne puisse être intégré que dans une solution, une architecture X, mais il faut qu'il soit ouvert. L'ouverture est primordiale, sinon... sinon les gens seront... Oui, les gens qui ne feront que du Schneider, ils seront intéressés, mais les autres... Il y a tout un tas d'autres... Je pense que c'est quelque chose qui peut être innovant, donc... Mais il faut qu'il soit ouvert, au moins en communication.*]

— Catégorie Cautic : Niveau 4 → **Adaptation**

  — critère 4.4 Le concept est-il adapté aux types d'organisation et à leur capacité d'évolution / aux manières de vivre de l'utilisateur et à leur évolution ?

— Catégorie FMC : **Condition**

Interprétation : L'utilisateur juge que le concept est inclusif pour un certain type de clientèle et qu'il serait souhaitable que le produit soit le plus ouvert possible pour toucher le maximun de personnes.

Évaluation d'un peigne d'alimentation électrique d'un tableau électrique

Verbatim (3) : [*Actuellement il faut une vis, mais là, les clips plastiques ça va pêter.*]

— Catégorie Cautic : Niveau 2 → **Association**

— critère 2.3 La comparaison avec les pratiques existantes valide-t-elle et rend-elle crédibles les pratiques nouvelles proposées ?

— Catégorie FMC : **Frein**

Interprétation : L'utilisateur fait comprendre que l'utilisation de clips plastiques pour fixer le peigne n'est pas adapté et solide. Des vis seraient plus adaptées.

On constate ainsi que les verbatims sont classifiés en fonction d'une première grille de critères et dans une deuxième phase, on cherche à connaître l'orientation du verbatim vis-à-vis de son premier niveau de classification.

2.6.2 Positionnement

L'analyse du discours et particulièrement des opinions est une tâche assez connue en traitement automatique des langues. Elle s'est accrue en parallèle de l'essor d'Internet et des médias sociaux. En TAL, l'analyse d'opinion est un problème très complexe. Elle englobe plusieurs tâches selon le domaine et le cadre applicatif dont la détection de la polarité. Cette tâche consiste habituellement à attribuer une polarité aux mots (positive, neutre ou négative). Dans le cas de nos travaux, nous souhaitons effectivement analyser les opinions présentes dans les textes transcrits, mais que d'attribuer une polarité, nous souhaitons attribuer à nos opinions une orientation en frein, en motivation ou en condition. Rappelons toutefois que les verbatims qui sont l'objet de notre tâche de classification sont codés manuellement selon deux processus. Notre tâche de classification sera précisément réalisée pour le second niveau de classification (*qui consiste à déterminer si une phrase émise dans un discours relatif à un objet d'innovation est un frein ou une motivation à l'acceptabilité de l'objet en fonction des critères CAUTIC* ®) comme observé à la figure 2.3. Cette tâche étant pour l'heure réalisée manuellement, il devient impératif de réfléchir à une autre méthode bien plus rapide et optimisée pour analyser les avis des personnes interviewées dans le cadre d'une étude.

Une **analyse automatisée** de ces verbatims permettrait non seulement **une réduction des temps d'analyse**, mais également **le traitement rapide et efficace des données actuelles** (issues des transcriptions d'entretiens ou de focus group) comme **des nouvelles données** (contenu en ligne de la plateforme Yoomaneo). Cette analyse automatique s'appuierait ainsi sur les données déjà annotées des anciennes études pour mieux classifier les données des nouvelles études.

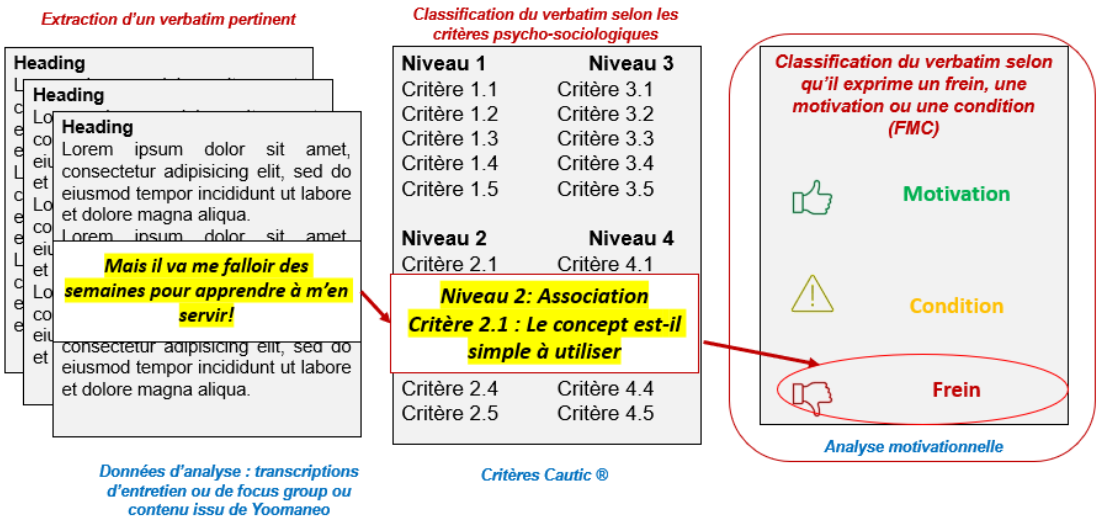


FIGURE 2.3 – Processus d'analyse à semi-automatiser (encadré en rouge).

Partant de ce postulat, de nombreux travaux en TAL se sont effectivement intéressés à l'analyse d'opinion dans une variété de domaines : finance (Moore et Rayson, 2017; Gaillat et al., 2018), politique (Thomas et al., 2006; Abercrombie et Batista-Navarro, 2018), articles de presse (Rahab et al., 2019), domaine médical (Grabar et al., 2019; Yadav et al., 2018), de commentaires de films (Maas et al., 2011), de produits Amazon (Haque et al., 2018; Rain, 2013). Les méthodes supervisées sont celles qui sont les plus utilisées dont les méthodes d'apprentissage profond. Bien que les systèmes présentés dans ces travaux donnent de bonne performance selon le domaine et pour la tâche de détection de la polarité, il reste beaucoup à faire sur la détection des freins, motivations et conditions.

Le travail qui sera présenté dans la suite s'axera sur l'analyse des opinions, perceptions des locuteurs dans le cadre d'études qualitatives visant à évaluer l'acceptabilité d'un produit innovant.

— En analyse de sentiment classique, nous avons :
  $\Longrightarrow$ Entrée = document/phrase, Sortie = sentiment (soit positif/négatif, soit une valeur sur une échelle)

— En analyse de sentiment basée sur l'aspect où la polarité est attribuée à chaque aspect évoqué dans une phrase d'opinion, nous avons :
  $\Longrightarrow$ Par exemple, sur un commentaire de restaurant [drinks=good, food=bad, service=good] => il faut donc déterminer quels aspects sont évoqués dans le document et prédire un jugement pour chacun.

Pour faire le parallèle avec notre tâche, les critères Cautic pourraient être considérés comme des "aspects" qu'il faut détecter pour chaque phrase et ensuite prédire une classification (motivation/condition/frein) pour chacun. Toutefois, dans ces travaux, notre tâche consistera à réaliser une classification générale et non fine au niveau de l'aspect. La partie concernant les critères d'acceptabilité de la méthode Cautic ® fait l'objet de travaux de recherche [16] distincts.

2.7 Conclusion

Dans cette partie, nous avons introduit le concept d'*innovation* en le définissant et en présentant les différents types d'innovations rencontrées. Nous avons également souligné que les innovations incrémentales sont celles qui sont les plus privilégiées par les entreprises. En outre, nous avons également spécifié qu'une innovation bien que technologiquement novatrice peut ne pas rencontrer de succès à son lancement sur le marché en raison de multiples facteurs comme une mauvaise définition du produit ou encore une connaissance insuffisante du marché. De ce fait, il est important d'évaluer l'impact d'une innovation avant de la développer ou la lancer sur le marché. Cet impact est mesuré au travers de d'études d'acceptabilité qui permettent de faire ressortir après analyse les freins, les motivations ou les conditions à l'acceptabilité du produit, c'est-à-dire l'adoption de l'innovation avant son usage réel par l'usager. L'analyse actuelle du matériau textuel recueilli au travers des outils de collecte mentionnés à la section 2.5.2 concerne essentiellement les transcriptions d'entretiens semi-directifs et des focus group. De plus, elle est réalisée manuellement. Cette dernière est déjà très fastidieuse et le sera encore plus avec les données issues de la plateforme Yoomaneo. Notre but est de proposer un outil qui permet de classifier les données provenant des transcriptions que celles provenant de la plateforme Yoomaneo. L'outil serait une aide précieuse pour les chargés d'études en termes de temps de travail. Cette thèse s'intéresse donc aux énoncés textuels sur des innovations issus de transcriptions d'entretiens ou de tables rondes et des contenus en ligne issus de la plateforme Yoomaneo. Dans le chapitre suivant, nous nous intéressons à l'état de l'art de l'analyse d'opinion avant d'introduire dans ce même chapitre les notions de frein, motivation et motivation sous condition.

16. http://www.theses.fr/s191792

Chapitre 3

Analyse des opinions par la détection des freins, motivations et conditions

Ce troisième chapitre s'intéresse aux concepts clés de la subjectivité, de l'opinion et du sentiment afin de mieux caractériser les classes de notre tâche de classification. Nous commençons par présenter en section 3.1 un bref historique de l'analyse d'opinion. En section 3.2, nous définissons l'analyse d'opinion et présentons les différentes tâches et campagnes qu'englobe l'analyse d'opinion. Ensuite, dans la section 3.3, nous caractérisons l'opinion en présentant les éléments qu'elle contient. La section 3.4 présente une synthèse des approches dans le domaine de l'analyse de l'opinion. Ensuite, la section 3.5 présente les challenges auxquels l'analyse d'opinion fait face. La section 3.6 est entièrement consacrée aux notions de frein, motivation et condition. Nous terminons le chapitre par une synthèse.

3.1 Historique et application

3.1.1 Historique

Les premiers sondages d'opinion naissent dans les années 1930 à 1940. Ils sont popularisés par le sociologue et statisticien George Gallup et repris ensuite par d'autres pays démocratiques dans le cadre des élections. L'intérêt pour les enquêtes par sondage d'opinion s'accroit sur la base qu'il existe une opinion publique distincte de l'opinion de la foule (Boullier et Lohard, 2012). Ainsi, la demande toujours plus pressante des organismes publics et privés pour obtenir des informations sur les caractéristiques de la population et leurs idées a donné aux enquêtes d'opinions une place de choix. Tout cela contribue ainsi à la création d'instituts de sondage tels que

l'IFOP (Institut français d'opinion publique) en 1938 par Jean Stoetzel. Les sondages d'opinion, fondés sur des enquêtes statistiques se démocratisent et prennent place dans le champ médiatique et politique.

Bien avant même l'avènement du Web 2.0, ce que pensent les gens a toujours été une information importante dans les processus de décision. Avec le boom d'Internet et du Web, les échanges en ligne se sont multipliés et réciproquement, les opinions des individus sont devenues facilement accessibles. Ces informations ou opinions des individus sont maintenant disponibles sous forme de commentaires, d'avis, de discussions à travers les différents médias et réseaux sociaux créés (blogs, microblogs, etc.) et portent aussi bien sur la politique que les personnalités, les produits, les biens et services, les marques, le sport, etc. En effet, selon la plateforme Digimind [1], 72% des consommateurs connectés dans le monde utilisent les médias sociaux (voir figure 3.1) au moins une fois par jour.

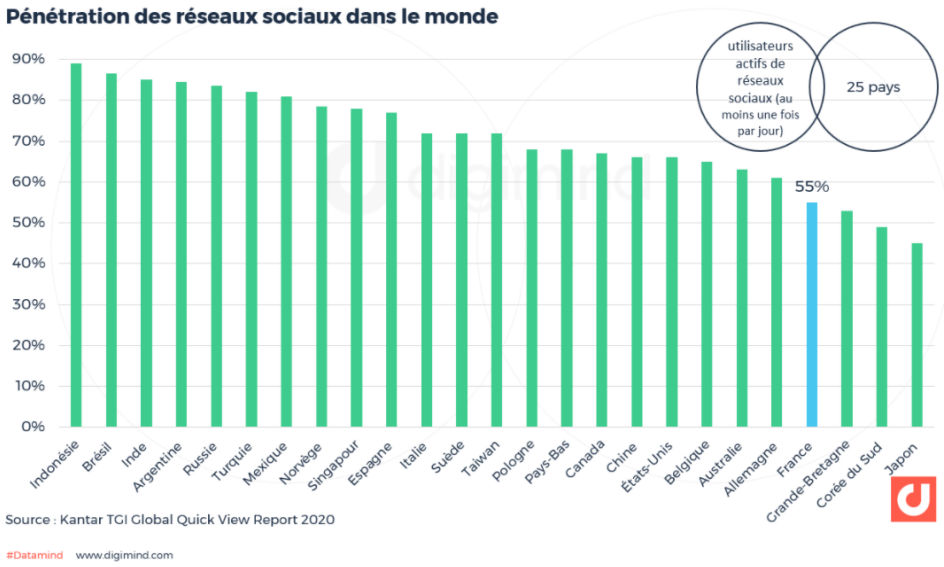


FIGURE 3.1 – Taux d'utilisation des réseaux sociaux dans le monde. [2]

Plus spécifiquement, 71% des utilisateurs connectés dans le monde utilisent YouTube [3], 67,5% utilisent Facebook, 52% Instagram, 32% Twitter, 22% Pinterest comme le montre la figure 3.2.

1. https://www.digimind.com/
2. https://www.kantar.com/fr/inspirations/publicite-medias-et-rp/
3. YouTube est plus un média social qu'un réseau et le 2ème moteur de recherche le plus utilisé dans le monde derrière Google.

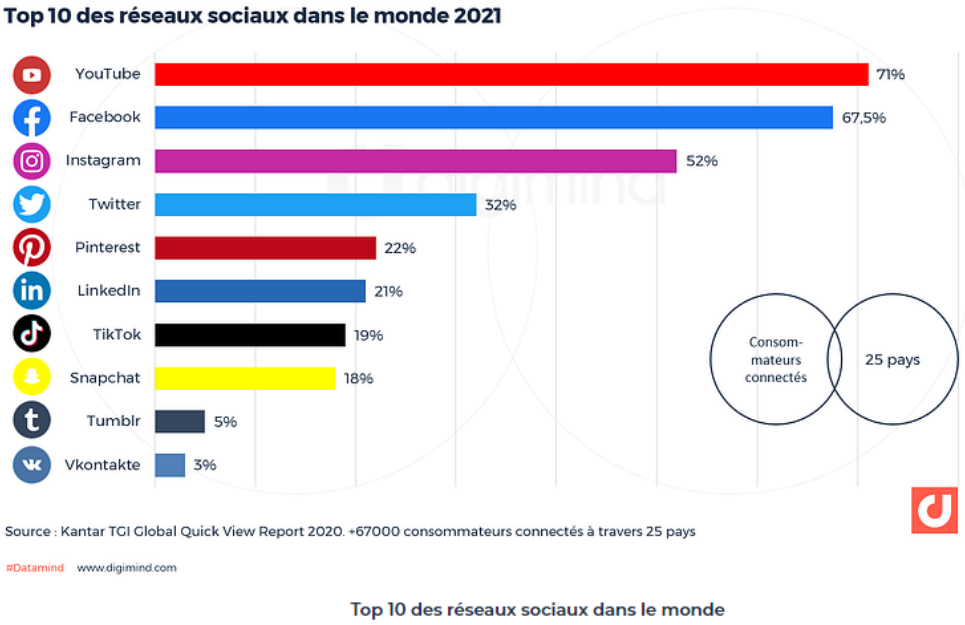


FIGURE 3.2 – Top 10 des réseaux sociaux dans le monde. [4]

L'émergence de ces multiples plateformes d'expressions a donné lieu à de nouvelles technologies d'analyses, alternatives aux sondages. C'est donc dans ce sillage de la croissance des messages et discussions en ligne qu'apparaît l'analyse d'opinion. Les sondages d'opinion et l'analyse d'opinion deviennent ainsi deux domaines similaires dans leurs buts et dans leurs applications qui proviennent essentiellement du domaine du marketing. Toutefois, ils diffèrent dans leurs approches de collecte d'opinions. Les sondages s'appuient sur des opinions issues de réponses à des questionnaires posés directement aux individus dans le cadre d'une enquête et sur un sujet bien identifié. L'échantillon des répondants est également et préalablement défini. La fouille d'opinion quant à elle s'alimente par des opinions émises spontanément sur Internet à travers les différents réseaux sociaux notamment le réseau Twitter qui est l'un des réseaux sociaux les plus utilisés pour la collecte d'opinion dans le cadre de campagnes d'évaluation de systèmes en analyse d'opinion.

Dans les années 2000, l'*analyse d'opinion* a ainsi bénéficié d'un énorme élan de recherche (Bakshi et al., 2016) lié non seulement à l'essor des méthodes d'apprentissage automatique dans le traitement du langage naturel et la recherche d'information, mais également de la disponibilité d'un ensemble de données sur lesquels

4. https://www.kantar.com/fr/inspirations/publicite-medias-et-rp/

les algorithmes d'apprentissage automatique pouvaient être entraînés.

3.1.2 Définition

L'analyse d'opinion est un domaine qui s'intéresse au traitement automatique des opinions, des sentiments et de la subjectivité exprimée dans les textes (Liu, 2012; Gillot, 2010). Il existe différents noms dans la littérature pour parler de l'*analyse d'opinion,* (Liu, 2012) comme *analyse de sentiment*, *fouille d'opinion*, *extraction d'opinion*, *analyse d'opinion*, etc. Toutefois, tous ces termes renvoient et représentent le même domaine d'étude. Même si le terme *analyse de sentiment* est premièrement apparu dans Nasukawa et Yi (2003) et le terme *analyse d'opinion* mentionné en premier dans Dave et al. (2003), ils sont utilisés de manière interchangeable dans la littérature. Dans la suite de notre rédaction, nous utiliserons uniquement le terme *analyse d'opinion*.

3.1.3 Domaine d'application

L'analyse d'opinion a un grand potentiel dans les applications. Elle est régulièrement utilisée dans tout processus décisionnel (élection, achat d'un bien, e-réputation).

— **Domaine commercial** :

Lorsque les consommateurs font des achats en ligne, nombreux sont ceux qui font leur choix en fonction des commentaires existants des autres clients sur les produits qu'ils souhaitent acquérir. Selon un rapport[5], 84% des personnes font confiance aux critiques en ligne et 90% des consommateurs lisent des avis en ligne avant de se rendre dans un lieu (restaurant, hôtel, entreprise, etc). Li et al. (2021b) proposent une méthode d'extraction de paires d'émotions et de causes (Emotion-Cause pair). Cette méthode extrait d'abord les émotions et les causes, et ensuite filtre les causes émotionnelles. Leur travail permet d'extraire non seulement les émotions ressenties par les consommateurs, mais aussi les raisons de ces dernières, de sorte qu'un commerçant peut tout à fait selon ses besoins ajuster sa stratégie de vente en connaissant la cause des sentiments exprimés par le consommateur. Kauffmann et al. (2019) montrent que l'application de l'analyse de sentiment peut aider dans le processus de décision commerciale à travers trois niveaux. Dans un premier temps, les préférences des clients sont analysées en attribuant des scores de sentiment, puis les parties positives comme négatives des commentaires sont séparées et finalement

5. https://www.netoffensive.blog/e-reputation/statistiques/

les principales caractéristiques du produit qui suscitent les sentiments négatifs ou positifs chez les clients sont extraites. Ce processus à pour objectif d'aider les responsables marketing dans leur processus décisionnel.

D'autres travaux ont également démontré les avantages d'avoir recours à l'analyse d'opinion (Islam et al., 2019; Jabbar et al., 2019). L'analyse d'opinion permet non seulement de classifier les avis au sujet d'un produit comme dans le cas des critiques de cinéma ; d'hôtels ou de restaurants, mais également d'être utilisée en sous composante d'autres technologies telles que dans le cas du résumé automatique d'opinion (Wang et Liu, 2015; Zhu et Penn, 2006). Il est également possible de réaliser une analyse d'opinion plus fine que la simple classification en polarité. Les commerçants peuvent ainsi utiliser un système d'analyse de sentiment au niveau des commentaires des clients afin de savoir quel attribut (par exemple, le prix ou le style) du produit intéresse plus le client. De même, la qualité des articles et du service après-vente des entreprises peut être améliorée au travers de l'analyse des remarques et des critiques des clients. Les commentaires des consommateurs sur un produit, un service ou encore une marque. peuvent également être récupérés pour faire de la veille.

— **Domaine de la finance** :

Dans le domaine de la finance, l'analyse des sentiments sur les actualités des marchés financiers fournit des informations significatives sur différentes échelles de temps pour la gestion des risques (Boullier et Lohard, 2012).

— **Domaine de la santé** :

L'analyse d'opinion est aussi appliquée au domaine de la santé. Barhoumi (2020) expliquent que diverses études ont été menées pour mesurer l'impact de certaines maladies sur les personnes touchées et leur entourage citant les travaux de Foufi et al. (2019) sur les maladies chroniques ou encore Gabarron et al. (2019) pour le diabète. L'analyse d'opinion peut également aider à mieux comprendre les problèmes des patients, à suivre les dossiers médicaux et à évaluer la réaction des patients (Chintalapudi et al., 2021).

Par ailleurs, il existe également de nombreux logiciels et plateformes pour l'analyse d'opinion. Semantria [6] est un outil d'analyse textuelle s'appuyant sur du traitement automatique du langage. IntenCheck API [7] est un outil basé sur un dictionnaire et principalement utilisé pour l'analyse d'opinion et l'analyse de texte. Il peut classer la polarité du texte en positif, négatif et neutre. KNIME Analytics Plat-

6. https://www.lexalytics.com/semantria/excel
7. https://www.intencheck.com/text-analytics-api/

form [8] (Lin et Luo, 2020) est un logiciel open source d'analyse de données. Cet outil a recours des modèles d'apprentissage pour la tâche de détection de polarité des textes (Minanovic et al., 2014). L'analyse d'opinion s'applique ainsi dans plusieurs domaines et intervient afin de développer des outils pour extraire, identifier, synthétiser ou encore comparer les opinions.

3.2 L'analyse d'opinion

3.2.1 La subjectivité

Les jugements d'évaluation et de valeur font partie intégrante de l'acte subjectif. La subjectivité est ainsi inhérente à la communication entre humains et est présente partout où il faut nouer des relations sociales (Jackiewicz, 2016). La notion de subjectivité a été introduite par les travaux de Benveniste (1966) et plus tard par Kerbrat et al. (1980). Benveniste (1966) définit la subjectivité comme « [...] la capacité du locuteur à se poser comme sujet dans et par le langage. Elle se définit, non par le sentiment que chacun éprouve d'être lui-même [...], mais comme l'unité psychique qui transcende la totalité des expériences vécues qu'elle assemble, et qui assure la permanence de la conscience. Or nous tenons que cette subjectivité [...] n'est que l'émergence dans l'être d'une propriété fondamentale du langage. Est ego qui dit ego. [...] la subjectivité se détermine par le statut linguistique de la personne.» - Comme l'explique Vernier (2011) dans ces travaux, cette définition de Benveniste oriente les linguistiques vers l'étude des marqueurs de subjectivité, c'est-à-dire les éléments, traces laissés volontairement ou involontairement par l'énonciateur dans son propre énoncé.

Deux systèmes d'énonciation [9] (Vernier, 2011) sont distingués : le récit et le discours. Le premier tend à l'objectivité et l'autre admet la subjectivité en plaçant l'énonciateur dans son énoncé mais dans la plupart des textes, ces deux systèmes s'entremêlent.

Kerbrat et al. (1980) distingue deux niveaux de subjectivité :

- le premier niveau de subjectivité n'implique pas l'expression d'une évaluation. Les traces de l'énonciateur peuvent être présentes de manière implicite comme explicite (usage des *embrayeurs*). Les embrayeurs ou marqueurs de personnes (*je, nous, etc.*) et spatio-temporels (*ici,)* sont des éléments qui révèlent la présence explicite de l'énonciateur dans son énoncé ;

8. http://www.greenxf.com/soft/207097.html

9. « L'énonciation est l'activité langagière exercée par celui qui parle au moment où il parle. » (Anscombre et Ducrot, 1976)

- le deuxième niveau de subjectivité est caractérisé par les traces d'une évaluation. Elle se caractérise par la présence de modalisateurs et d'unités linguistiques évaluatives (Vernier, 2011).

3.2.2 Tâches en analyse d'opinion

L'analyse d'opinion englobe plusieurs tâches appliquées à différents domaines. Parmi ces tâches, nous distinguons :

— la détection de la subjectivité : il s'agit de déterminer si un document véhicule une opinion (subjectif) ou ne présente que des faits (objectif) (Pang et al., 2002; Wiebe et al., 2004; Ounis et al., 2008; Ding et al., 2008) ;
⟶ *Le président s'exprimera ce soir à la télévision.* => objectif

— la détection de la polarité : elle est l'une des tâches les plus populaires en analyse d'opinion. Elle consiste à déterminer la polarité ou l'axiologie de l'opinion : positive, négative ou neutre (Pang et al., 2002; Turney, 2002; Mulki et al., 2017) ;
⟶ *Cet appareil photo est laid.* => polarité négative

— la détection de l'intensité de la polarité : il s'agit de déterminer l'intensité de la polarité de l'énoncé textuel (Mulder et al., 2004) à l'aide de marqueurs d'intensification ou d'un lexique polarisé. Est-ce qu'il est très positif ou très négatif ;
⟶ *Cet appareil photo est très laid.* => polarité très négative

— la détection des émotions (Lu et al., 2006; Vidrascu, 2007) : plusieurs typologies d'émotions existent dont celle de Ekman (1992) à savoir : *joie, colère, tristesse, dégoût, surprise et peur* ou encore celle de Plutchik et Kellerman (1986) avec ses 8 émotions de bases ;
⟶ *Je suis contente car j'ai eu mon permis.* => joie

— l'identification de la cible de l'opinion (Lark et al., 2015)
⟶ *Nous devons valoriser les* ***énergies renouvelables.***

— l'identification de la polarité de l'aspect Lark et al. (2015) citant (Brun et al., 2014; Kiritchenko et al., 2014) ⟶ *Les* ***plats*** *de ce restaurant sont délicieux.*

=> polarité postive

— l'identification du détenteur/l'émetteur de l'opinion (Gangemi et al., 2014)
⟶ *Selon le* ***directeur des ressources humaines****, les stagiaires ne sont pas compétents cette année.*

3.2.3 Campagnes d'évaluation

Les campagnes d'évaluation pour l'analyse d'opinion consistent à évaluer différents systèmes sur des tâches spécifiques. Ces évaluations sont rendues possibles grâce à la création de la première campagne d'évaluation nommée SensEval en 1998 (Kilgarri, 1998) centrée sur la tâche de désambiguïsation lexicale. Cette campagne a été renommée en SemEval deux ans plus tard et propose des tâches allant de la détection des entités nommées à l'analyse de sentiment. On retrouve les tâches liées à l'analyse de sentiment dans les éditions 2007 (Strapparava et Mihalcea, 2007), 2013 (Chawla et al., 2013), 2014 (Daval-Frerot et al., 2018), 2015 (Ghosh et al., 2015a), 2017 (Rosenthal et al., 2017) et 2018 (Saias, 2014). On peut également citer la campagne TREC (Text Retrieval Conference) de 2007 (Macdonald et al., 2007) et de 2008 (Ounis et al., 2008).

Sur le plan francophone, on peut citer la campagne d'évaluation Défi Fouille de Texte (DEFT) avec une première édition en 2005 (Azé et Roche, 2005). Les éditions 2007 (Torres-Moreno et al., 2007), 2015 (Hamon et al., 2015) et 2017 (Benamara et al., 2017a) ont proposé des tâches dans le domaine de la fouille d'opinion. On distingue deux principales tâches pour l'évaluation des systèmes en analyse de sentiment :

1. la tâche de classification des tweets selon leur polarité. Cette tâche consiste à classer le tweet selon l'opinion/sentiment/émotion exprimé par son auteur en objectif, positif, négatif ou mixte (si le tweet contient à la fois des opinions positives et des opinions négatives).

2. la tâche de classification du sentiment relatif à un aspect. Il s'agit d'attribuer une polarité (positive, négative ou neutre) à chaque aspect évoqué dans une phrase d'opinion. On procède d'abord à l'extraction des aspects et ensuite une polarité est attribuée à chaque aspect extrait.

3.3 Caractéristiques générales de l'opinion

3.3.1 Modalisation de l'opinion

Liu (2012) dans ses travaux sur l'analyse d'opinion définit l'opinion comme un quintuple d'éléments où :

- e est la cible ou le sujet de l'opinion ;
- a est un aspect de e ;
- s est le sentiment que h exprime envers e ou a ;
- h est l'émetteur ;
- t est la date à laquelle l'opinion est exprimée.

s peut être catégorisé de trois manières différentes (Karoui, 2017) :

— **le type sémantique** du sentiment exprimé. Par exemple, dans la phrase *"Cette émission m'a ennuyé "*, l'auteur exprime un sentiment d'ennui alors que dans *J'ai adoré cette émission*, l'auteur exprime un jugement d'évaluation.

— **la polarité** du sentiment exprimé qui peut être positive ou négative.

— **la valence** du sentiment qui indique l'intensité ou degré de l'opinion.

Liu (2012) fait également savoir que cette définition ne couvre pas toutes les facettes possibles de la signification sémantique d'une opinion qui peut être arbitrairement complexe. Par exemple, elle ne couvre pas la situation décrite dans *"Le viseur et l'objectif sont trop proches"*, qui exprime une opinion sur la distance entre deux objets. Elle ne couvre pas non plus le contexte de l'opinion, par exemple *"Cette voiture est trop petite pour une personne de grande taille"*, ne dit pas que la voiture est trop petite pour tout le monde. *"Une personne de grande taille"* est le contexte ici. Si la description de l'opinion proposée au travers de ce quintuple semble facile à comprendre, son implémentation est plus complexe pour un programme informatique (Liu, 2012).

3.3.1.1 Les facettes de l'opinion

Les opinions peuvent être exprimées sur des choses, des entités, des êtres, des évènements. Les opinions sont considérées comme des expressions subjectives que le locuteur utilise pour évaluer ou juger. Lorsque l'on parle d'évaluation, il s'agit d'une évaluation qui porte sur une entité ou un concept. Cette évaluation peut se manifester par l'utilisation de jugements, de sentiments ou via l'émission des conseils ou recommandations.

Prenons le cas des phrases suivantes que nous allons essayer de caractériser avec les différentes typologies d'opinions énoncées plus haut :

(1) *J'ai trouvé la vidéo claire et les animations plutôt bonnes.*
(2) *L'auto-diagnostic est appréciable.*
(3) *Plutôt sceptique par rapport au lien entre utilisation du produit & remise sur mutuelle.*
(4) *Je suis jalouse de sa maison.*
(5) *Je ne pourrais pas venir ce soir, je crois qu'il va pleuvoir.*

Les phrases 1 à 3 illustrent parfaitement la définition de l'opinion évaluative mentionnée ci-dessus. L'auteur y exprime son opinion via des adjectifs de type positifs *(claire, bonne, appréciable)*. Dans la phrase 4, il s'agit de l'expression d'un sentiment, néanmoins, il ne traduit pas une évaluation, mais plutôt une émotion et apparaît indépendamment d'une opinion portée sur une entité. De même, certaines expressions qui relèvent de l'opinion peuvent ne pas constituer des évaluations. Pour la phrase 5, l'auteur émet une hypothèse sans évaluer la météo.

3.3.1.2 Catégorisation sémantique de l'opinion

Asher et al. (2008); Péry-Woodley et al. (2009) proposent également une catégorisation sémantique de l'opinion. Chaque expression d'opinion peut appartenir à l'une des catégories suivantes : le reportage, le jugement, le sentiment-appréciation et le conseil. Cette taxinomie est utilisée dans les travaux de Chardon (2013) sur la constitution d'une chaîne de traitement pour l'analyse discursive.

1. **Le reportage**
   Les expressions de reportage relatent ou introduisent les opinions des autres ou les siennes.
   $\longrightarrow$ *J'affirme que ce restaurant est bon.* Elle comporte plusieurs sous-classes.

   - Informer/soutenir : les expressions appartenant à cette classe véhiculent une opinion que l'auteur estime vraie.
   $\longrightarrow$ *Le prévenu affirme ne pas être coupable.*
   - Dire/remarquer : entre dans cette classe les expressions de l'auteur qui n'apportent pas son point de vue.
   $\longrightarrow$ *Il remarque que l'écrivain signe de la main de gauche.*

- Penser/supposer : les expressions renvoient à ce que pense ou considère l'auteur.
⟶ *J'estime que cela est faux.*

2. **Le jugement**
Les expressions de jugement expriment des évaluations. Ces expressions possèdent très souvent une polarité et une intensité.
⟶ *Ce tableau est magnifique.* Ces expressions sont regroupées en différentes sous-classes.

- Blâmer/louer : les expressions appartenant à cette classe indiquent un jugement sur la responsabilité de quelqu'un ; ces expressions ont une polarité, mais ne possèdent pas d'intensité.
⟶ *Je condamne les actions des habitants de la ville.*
- Évaluer : les expressions de cette classe jugent de manière positive ou négative un objet ou une personne.
⟶ *Ce tableau est magnifique.*

3. **Le sentiment-appréciation**
Les expressions de sentiment-appréciation expriment des émotions ou sentiments ressentis. Elles sont regroupées selon leurs catégories : les sentiments positifs, négatifs ou neutres. Chaque catégorie est subdivisée en différentes classes.
⟶ *J'ai adoré l'émission*.

Les sentiments positifs peuvent contenir les classes **Apaisement** (*rassurant*), **Divertissement** (*réjouissant*), etc.
Les sentiments négatifs se composent des classes : **Colère** (*énervant*), **Peur** (*angoissant*), etc.
En ce qui concerne les sentiments neutres, on distingue les classes **Étonnement** (*sidérant*) et **Émotion** (*émouvant*).

4. **Le conseil**
Les expressions de conseil incitent à faire ou à penser quelque chose.

Les termes et expressions de cette catégorie sont regroupés entre les classes suivantes :

- Recommander : les expressions appartenant à cette classe véhiculent une opinion bonne ou mauvaise et essayent de convaincre avec force.
⟶ *Je plaide en faveur de la gratuité des transports le week-end.*
- Suggérer : entrent dans cette classe les expressions de l'émetteur pour suggérer ou spéculer sans être absolument certain.
⟶ *Je suggère que nous déplacions la réunion en fin de semaine prochaine.*
- Espérer : les expressions renvoient que les désirs et souhaits de l'émetteur seront satisfaits.
⟶ *J'espère que mon train sera à l'heure.*

Comme nous pouvons l'observer, l'opinion revêt plusieurs variantes. Elle est émise par un locuteur/émetteur (personne, une institution, etc.) pour évaluer, juger, réagir sur un sujet qui lui-même peut être de nature différente (objet, personne, etc.). Dans les différentes taxinomies énumérées ci-dessus, les catégories jugement et appréciation sont des opinions évaluatives sur un sujet. Cette évaluation est généralement positionnée un axe dit *polarisé* (positif au négatif en passant par le neutre). Dans le cadre de nos travaux, nous nous intéressons uniquement aux opinions évaluatives non sur une échelle polarisée, mais sur un nouvel axe à trois dimensions FMC.

3.3.1.3 Les modalités évaluatives

Les opinions évaluatives ont été également l'objet des études de Charaudeau (1992) qui caractérise différentes modalités énonciatives qui servent à exprimer une évaluation en français. Charaudeau (1992) propose une liste de modalités énonciatives pour la langue française parmi lesquelles l'opinion, l'accord/désaccord, le jugement et l'appréciation. Vernier (2011) reprend les modalités de Charaudeau (1992) et propose de les répartir en deux types de modalités : les modalités d'évaluations axiologiques et les modalités d'évaluations logiques.

Les modalités logiques regroupent l'opinion et l'accord/désaccord.

— L'**opinion** est une modalité logique qui permet d'exprimer une évaluation, mais n'actualise pas d'axiologie positive ou négative en elle-même (Jarukan,

2014). Ici, l'énonciateur évalue son propos tout en exposant son point de vue. Ce point de vue peut être plus ou moins certain. Elle possède deux variantes : l'une concerne la **conviction** qui suppose un doute sur la vérité et le besoin de l'énonciateur d'exprimer sa certitude qu'il a de la vérité et l'autre la **supposition** qui implique que l'énonciateur n'a pas la certitude totale sur ce qu'il exprime. Cette certitude peut varier de la certitude forte au pressentiment. Les expressions suivantes illustrent une opinion.
⟶ *Je suis persuadé, je suis certain, je me doute que, j'imagine que*

— L'**accord/désaccord**. « L'accord ou le désaccord présuppose qu'il a été adressé à l'énonciateur une demande de dire s'il adhère ou non à la vérité d'un propos ou d'un acte tenu par un autre (que cette demande ait été faite réellement ou non). L'énonciateur répond en exprimant qu'il adhère ou non au propos ou à l'acte tenu. Du même coup, il contribue à la validation de la vérité de celui-ci. Entre l'accord et le désaccord des nuances peuvent être spécifiées : accord total, accord approximatif, rectificatif » (Charaudeau, 1992). L'énonciateur a le choix d'aller dans le sens cette croyance ou de s'opposer. Les expressions suivantes illustrent un accord ou un désaccord.
⟶ *Je suis globalement d'accord que, oui, mais..., je ne suis pas d'accord*

Les modalités axiologiques regroupent le jugement et l'appréciation.

— Le **jugement** est une modalité axiologique à travers laquelle l'énonciateur juge si quelque chose est bon ou mauvais en déclarant son approbation ou réprobation. Le jugement porte sur les domaines de l'*éthique* et la *morale* ou l'*intellect*.
⟶ *Je vous félicite pour votre travail ! ; Ton attitude n'est pas correcte*

— L' **appréciation** est la modalité évaluative la plus subjective. Charaudeau (1992) propose la définition suivante pour l'appréciation : « il est présupposé un fait à propos duquel l'énonciateur dit quel est son sentiment. L'énonciateur évalue donc, non plus la vérité du propos, mais sa valeur, en révélant ses propres préférences. Cette évaluation est nécessairement polarisée. »
⟶ *Je trouve positif que, je trouve bien, j'apprécie que versus je trouve négatif que, je trouve dommage que*

3.3.2 Niveaux de granularité en analyse d'opinion

En section 3.1.1, nous avons mentionné que l'analyse d'opinion a bénéficié de la disponibilité d'une multitude d'informations accessibles sur les forums, les sites

d'e-commerce sous la forme de commentaires ou encore sur Twitter sous la forme detTweets. Néanmoins, la longueur de ces énoncés varie en fonction de sa source et des règles imposées par la source. Par exemple, le nombre de caractères sur Facebook est limité à 8 000 [10] tandis que sur Twitter il est de 280. Un autre point à prendre en considération est l'émetteur. En effet, certains émetteurs rédigent des commentaires concis comme d'autres détaillent en profondeur leur opinion ajoutant même des sources extérieures. L'opinion peut ainsi être analysée selon différents niveaux de granularité. Le premier niveau se situe à l'échelle du document entier, le deuxième niveau concerne la phrase, le troisième niveau est porté sur un segment ou groupe de mots, et le dernier niveau se focalise sur un seul mot. Ces quatre niveaux sont détaillés dans la suite.

3.3.2.1 L'opinion au niveau du document

L'analyse d'opinion au niveau du document (Farra et al., 2010; Paroubek et al., 2018) est un problème de classification qui peut être subdivisé en deux sous-tâches :

— La tâche de détection de la subjectivité (section 3.2.2)

— La tâche de détection de la polarité (section 3.2.2)

L'hypothèse avancée est que l'opinion exprimée dans le document porte sur une seule entité (i.e un commentaire sur un film). Cette analyse n'est pas applicable aux documents dans lesquels différentes entités sont comparées en même temps. Aussi, si différentes polarités sont exprimées, la plus pertinente doit être choisie.

3.3.2.2 L'opinion au niveau de la phrase

Tout comme l'analyse au niveau du document, l'analyse au niveau de la phrase est un problème de classification qui s'attèle à distinguer les phrases objectives des phrases subjectives. Les phrases subjectives expriment des opinions et peuvent être classées en fonction de leur polarité. L'hypothèse formulée est que chaque phrase dans le texte exprime une opinion unique sur une entité unique (Yessenalina et al., 2010; Shoukry et Rafea, 2012; Liu, 2012).
⟶ *J'aime bien les chaussures de Reebok.* L'orientation de la phrase est de nature positive.

10. https://www.blogdumoderateur.com/taille-limite-posts-reseaux-sociaux/

3.3.2.3 L'opinion au niveau de l'aspect

L'analyse au niveau de l'aspect est plus détaillée que l'analyse au niveau du document et de la phrase parce qu'elle se focalise sur ce que les gens aiment ou n'aiment pas réellement. Ainsi, cette analyse consiste à associer une polarité étant donné un aspect (Pontiki et al., 2015, 2016; Poria et al., 2020; Chinsha et Joseph, 2014; Guo et al., 2021). Cette analyse peut être précédée par la tâche d'identification des aspects. Par exemple,
⟶ *Personnel sympathique, mais la note est chère.*
Il s'agira dans la phrase précédente d'associer une polarité positive à "personnel" et une polarité négative à "note". Alors que la plupart des méthodes proposées dans la littérature concerne le niveau du document ou de la phrase, il existe un intérêt croissant pour le niveau de la cible/entité ou de l'aspect. Cependant, la tâche est encore beaucoup plus complexe, car il faut non seulement détecter le sentiment, mais également le lier à une entité ou à un aspect qui doit préalablement être détecté.

3.3.2.4 L'opinion au niveau du mot

L'analyse d'opinion au niveau du mot consiste à déterminer la polarité du mot. Cette analyse est généralement réalisée dans le cadre de la constitution de lexiques polarisés. Un lexique polarisé se compose d'un ensemble de couples (w, s) ou le w est le mot et s le score de polarité ou la polarité. Les mots que l'on retrouve généralement dans des lexiques polarisés sont les verbes, les adjectifs, les noms ou encore les adverbes (Barhoumi, 2020). L'analyse d'opinion est essentiellement étudiée selon les quatre niveaux mentionnés précédemment (document, phrase, aspect et mot). Toutefois, ces niveaux ne sont pas les seuls. Un grand nombre de travaux ont exploré le problème de l'analyse d'opinion en utilisant d'autres niveaux (Ravi et Ravi, 2015) tels que le niveau de la proposition, du concept, du sens, des syntagmes.

3.4 Les approches en analyse d'opinion

Dans cette section, nous passons en revue les différentes méthodes existantes pour traiter le problème de l'analyse d'opinion. Les approches en analyse d'opinion reposent sur trois grandes familles de méthodes (voir figure 3.3) : **les méthodes symboliques** qui s'appuient sur des règles (*les règles sont généralement implémentées sous la forme d'expressions régulières*) ou des lexiques polarisés (*il s'agit de dictionnaires*) et **les méthodes statistiques** dont le cœur est l'apprentissage automatique. Ces dernières reposent sur la construction d'un modèle à partir d'un corpus

annoté. La dernière famille de méthode concerne les méthodes hybrides qui sont des combinaisons des méthodes symboliques et numériques. In fine quelle que soit la méthode choisie, le modèle choisi repose sur le corpus utilisé ou à partir duquel il a été construit. Le reste de cette section présente tout à tour ces méthodes.

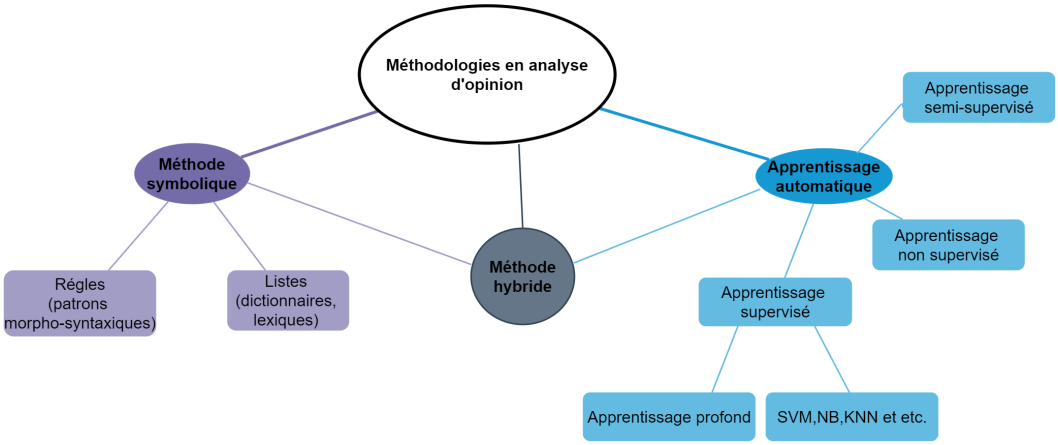


FIGURE 3.3 – Approches en analyse d'opinion.

3.4.1 Acquisition des données

Les données en TAL peuvent ne pas être directement disponibles ou même si elles sont disponibles, il faut les agréger. Si les données ne sont pas disponibles, elles peuvent être directement récupérées depuis différentes plateformes comme les réseaux sociaux (i.e Twitter), les blogs, les sites d'e-commerce, etc. Pour cela, certains réseaux sociaux comme Twitter propose une API [11] pour directement collecter les données. Dans d'autres cas, les données peuvent être directement accessibles : il s'agit généralement de données mises à disposition pour la communauté scientifique, par exemple les données à utiliser pour une campagne d'évaluation (section 3.2.3) ou spontanément par des chercheurs. Ces données peuvent être encodées sous divers formats tels que le format XML [12] ou sous forme tabulaire (par exemple, le format CSV [13].

11. En anglais : « Application Programming Interface », traduit par Interface de programmation d'application qui est un ensemble de méthodes, fonctions ou protocoles pour créer et intégrer des logiciels d'application.
12. En anglais : « Extensible Markup Language » traduit par « langage de balisage extensible en français »
13. En anglais, « Comma-seperated values ».

3.4.2 Le prétraitement des données

Une fois les données collectées de différentes sources, elles ont besoin d'être prétraitées. Les étapes de prétraitements les plus utilisées en analyse d'opinion sont : le nettoyage, la tokenisation, la suppression des mots vides, la racinisation, l'étiquetage morpho-syntaxique, etc.

— la tokenisation : elle consiste à segmenter un texte en une série de tokens individuels (mots ou phrases) en séparant la ponctuation. Le tokeniseur, l'outil qui effectue la segmentation est complètement dépendant de la langue du corpus sur lequel on travaille ;

— le nettoyage : il varie selon l'origine des données et inclut par exemple la suppression d'urls, émoji ;

— la lemmatisation : elle consiste à obtenir la forme canonique ou lemme des mots. Par exemple, *bruits -> bruit* ;

— la racinisation : elle consiste à découper le mot afin de ne conserver que la racine du mot. Le but de la racinisation est de regrouper toutes les variantes d'un mot sous une seule forme. Par exemple, *manger -> mang* ;

— l'étiquetage morpho-syntaxique : Elle consiste à associer à chaque token, sa partie du discours (verbe, nom, déterminant, etc.) ;

— d'autres opérations : la suppression des mots vides, de la ponctuation, des nombres, des symboles, le passage en minuscule, etc.

3.4.3 L'approche symbolique

Les approches symboliques s'appuient soit sur un ensemble de règles, soit sur des lexiques polarisés pour déterminer la polarité d'un énoncé textuel.

a) Les méthodes à base de règles

Les méthodes à base de règles requièrent la mobilisation de connaissances d'experts et reposent sur l'implémentation d'expressions régulières (Fielstein et al., 2004) pour l'écriture de patrons syntaxiques. Ces experts doivent également être capables de représenter les informations sous toutes les formes possibles. En effet, ils doivent pouvoir connaître toutes les multiples formes de présentation de l'information à traiter. Toutefois, bien qu'elles produisent des résultats de qualité, elles se révèlent très coûteuses en temps et ne sont pas généralisables à d'autres domaines, ce qui est un inconvénient majeur.

b) Les méthodes fondées sur des lexiques

Les lexiques polarisés sont généralement utilisés pour déterminer l'orientation des documents par identification et agrégation des mots polarisés présents dans le document. L'identification des mots polarisés est réalisée en se basant sur un lexique polarisé. Concernant l'agrégation, une technique réside dans le comptage du nombre de mots positifs et négatifs. Si le nombre de mots positifs est plus important que les mots négatifs alors l'énoncé est déclaré positif. Si, c'est l'inverse l'énoncé est déclaré négatif. Pour les lexiques mentionnant le score de polarité des mots, la polarité est déterminée par le calcul d'un score de polarité qui est obtenu en additionnant le score de tous mots les mots positifs et négatifs identifiés dans l'énoncé. Certaines méthodes symboliques prennent également en compte la négation ou l'intensification (Mulki et al., 2017; Molina-González et al., 2013). Une grande majorité des recherches s'est focalisée sur l'usage des adjectifs (Taboada et al., 2011; Hatzivassiloglou et McKeown, 1997; Wiebe et al., 2000) comme indicateurs de l'orientation d'un texte. Les lexiques utilisés sont habituellement construits manuellement. Par exemple, Abdulla et al. (2013) construisent manuellement un lexique de sentiments pour la langue arabe de 4815 mots sous la supervision de linguistes natifs de la langue. Ensuite, un système calcule le nombre de mots positifs et négatifs dans le document afin de générer la polarité globale ou de référence du texte, semi-automatiquement (Abdaoui et al., 2017) ou automatiquement (Barhoumi, 2020; Mourad et Darwish, 2013). Pour la méthode automatique, un dictionnaire initial est utilisé. Ce dictionnaire contient une liste de mots appelés « *seed* » Nous présentons dans la suite une liste de lexiques polarisés pour le français (Kelodjoue, 2019) :

- la ressource Feel [14] (Abdaoui et al., 2017) développée par le LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) contient 14 129 termes pour le français. Chaque terme est associé à deux polarités (positive, négative) et à six émotions (joie, colère, peur, tristesse, dégoût et surprise). Le lexique a été obtenu en traduisant et en étendant aux synonymes le lexique anglais NRC EmoLex. Ensuite, chacune des entrées a été validée par un annotateur humain ;

- la ressource JeuxDeMots [15] (Lafourcade et al., 2015) où l'enrichissement de la liste de mots polarisés se fait de façon contributive. Les joueurs/visiteurs sont amenés à indiquer la polarité et l'émotion des expressions affichées par des jeux. Ils peuvent choisir entre trois polarités (neutre, positif et négatif) et 21 émotions. Ils peuvent également rajouter des nouvelles émotions lorsqu'ils jugent que l'émotion

14. http://www.lirmm.fr/~abdaoui/FEEL
15. http:www.jeuxdemots.org/jdm-accueil.php

affichée par rapport à un terme n'est pas présente parmi les 21 choix. La ressource contient plus de 824 434 termes ;

- Polarimots [16] (Gala et Brun, 2012) est un lexique de polarité construit en utilisant la deuxième version de POLYMOTS (Rey et Gala, 2011), une ressource lexicale regroupant 19 009 mots en 2 069 familles morpho-phonologiques. La deuxième version de cette ressource contient une description plus fine de quelques familles de mots en clusters sémantiques. À ce jour, la ressource contient 7 483 noms, verbes, adjectifs et adverbes français dont la polarité (positive, négative ou neutre) a été annotée semi-automatiquement. 3 247 mots ont été ajoutés manuellement et 4 236 mots ont été créés automatiquement en propageant les polarités ;

- la ressource Affect (Augustyn et al., 2006) contient environ 1 300 termes français décrits par leur polarité (positive et négative) et plus de 45 catégories hiérarchiques émotionnelles. Il a été construit automatiquement et inclut d'autres informations telles que l'intensité et le niveau de langue (commun, littéraire, etc.) ;

- le lexique CASOAR [17] (Benamara et al., 2014) contient 2830 mots ou expressions d'opinion classés en 4 catégories sémantiques (reportage, jugement, sentiment-appréciation et conseil).

Bien que la méthode symbolique ne nécessite pas l'usage de données étiquetées et d'étapes d'apprentissage, la construction des ressources et des règles nécessitent du temps et sont très dépendants du domaine. Les règles préétablies et les lexiques génériques ne sont pas fiables pour d'autres domaines. Des lexiques spécifiques au domaine cible seraient une alternative.

3.4.4 L'approche statistique

L'approche statistique repose essentiellement sur des techniques d'apprentissage automatique et/ou profond. Ces techniques peuvent être utilisées soit pour une tâche de classification binaire (deux classes), soit multiclasse (plus de deux classes). Ainsi, deux étapes primordiales sont distinguées quelle que soit la technique utilisée : une phase d'apprentissage et une phase de prédiction. La phase d'apprentissage est l'étape au cours de laquelle le modèle est construit à partir d'un corpus d'apprentissage. La phase de prédiction permet d'évaluer la performance du modèle préalablement construit sur un nouveau jeu de données. Parmi les techniques d'apprentissage les plus répandus, nous citerons l'apprentissage supervisé

16. http://polarimots.lif.univ-mrs.fr/
17. https://projetcasoar.wordpress.com/

(AS), l'apprentissage non supervisé (ANS) et l'apprentissage par renforcement [18].

Nous nous focaliserons ici sur l'AS. L'AS peut s'appuyer sur différents algorithmes tels que la régression linéaire, les arbres de décision, les séparateurs à vaste marge (SVM), les réseaux de neurones (Barhoumi et al., 2018; Gaillat et al., 2018; Moore et Rayson, 2017; Cliche, 2017; Ali et al., 2018; Kooli et Pigneul, 2018; Mulki et al., 2017), etc.

En apprentissage supervisé, les systèmes sont entraînés sur des données où chaque échantillon est associé à une catégorie. De ce fait, il est préférable de disposer d'un grand nombre d'exemples pour chaque classe. À l'inverse, l'apprentissage non supervisé (ANS) ne requiert pas des données d'entraînement étiquetées. Les algorithmes en ANS opèrent à partir d'exemples non annotés.

3.4.5 L'approche hybride

L'approche hybride combine les méthodes fondées sur les lexiques et les méthodes statistiques (Hedar et Doss, 2013; Maurel et al., 2008). Refaee et Rieser (2016) utilisent la régression logistique pour la prédiction de l'intensité de la polarité dans les tweets. La régression logistique a été utilisée pour prédire les scores qui sont ajustés en appliquant les règles extraites du lexique.

3.5 Les challenges en analyse de sentiment

3.5.1 Les opérateurs d'opinions

La détermination de la polarité est une tâche primordiale en analyse de sentiment. Néanmoins, certains éléments linguistiques agissent directement sur la polarité des opinions en altérant les caractéristiques de l'opinion dans sa portée. Parmi ces éléments linguistiques, nous pouvons distinguer les opérateurs de négation, les intensifieurs et les modalités (voir figure 3.4).

- Les opérateurs de négation

La négation a été étudiée en linguistique pour l'anglais dans les travaux de Horn (1989); Giannakidou (2011); de Swart (2010) et le français dans les travaux de Moeschler (1992); Muller (1991b); Corblin et Tovena (2003); Abeillé et Godard (2010); Muller (1991a).

18. L'apprentissage par renforcement est un type d'apprentissage au cours duquel le système apprend à prendre des décisions optimales dans un contexte spécifique.

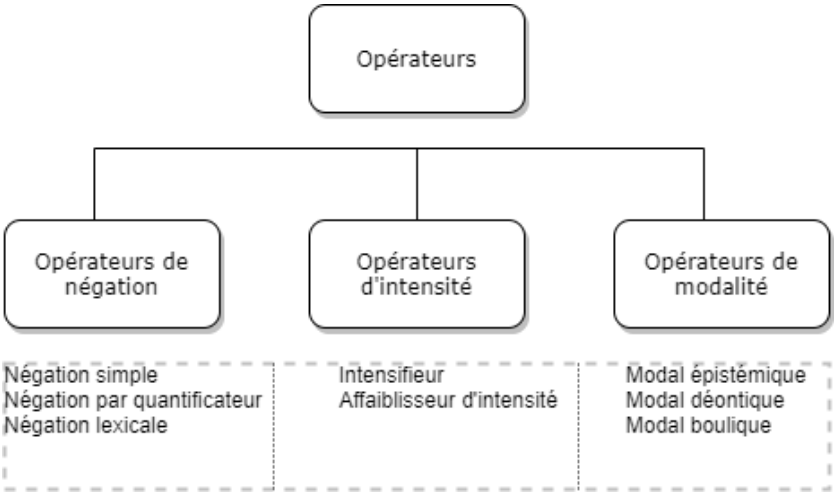


FIGURE 3.4 – Opérateurs d'opinion (Chardon, 2013).

La phrase (1) véhicule une opinion plus intense que la seconde et est portée par le terme *excellent*.

(1) ⟶ *Cet employé est excellent.*

La phrase (2) est influencée par la présence d'un élément de négation (ne...pas).

(2) ⟶ *Cet employé* ***n****'est* ***pas*** *excellent.*

La portée de la négation peut également varier comme dans la phrase (3). (3) ⟶ *Tous les employés* ***ne*** *sont pas excellents.*

Les opérateurs de négation sont des éléments linguistiques qui influent la polarité ou encore la valence des opinions sur lesquelles ils portent. En TAL, divers systèmes ont été proposés pour traiter le problème de la négation. Pour les méthodes fondées sur l'utilisation de ressources et de règles et réalisées pour la langue anglaise, citons le système NegEx de Chapman et al. (2001). NegEx utilise les expressions régulières pour détecter les marqueurs de négations et localiser les termes qui se trouvent sous leur portée à droite comme à gauche du marqueur. Pour les méthodes d'apprentissage, Pang et al. (2002) utilisent des classifieurs statistiques comme un SVM et un naïf bayes pour une tâche de détection de la négation. D'autres systèmes ont été également proposés. Pour une lecture détaillée, le lecteur pourra consulter les travaux de Dalloux (2017); Wiegand et al. (2010) qui proposent un état de l'art complet sur cette problématique de la négation.

- Les intensifieurs
Les marqueurs d'intensifieurs modifient la valeur initiale d'un terme en l'intensifiant ou en l'atténuant (Boubel, 2012, 2001; Bernhard et Ligozat, 2011). Ceux-ci sont généralement des adverbes à l'exemple de *très, moins, moyennement, etc*. Prenons les deux exemples suivants.

⟶ (1) *Luc est* ***très*** *bon.*

⟶ (2) *Luc est* ***moins*** *bon.*

Dans la phrase (1) l'adverbe *très* amplifie le terme *bon* et pour la phrase (2) le *moins* atténue le terme bon. Les deux phrases précédentes ont certes une orientation positive mais la valence sur la phrase (1) est plus forte que sur la (2).

- Les verbes de modalité (Chardon, 2013) comme *croire, devoir, etc.* qui influencent la force d'une expression et son degré de certitude.

⟶*J'espère que ce restaurant est bon.*

3.5.2 La dépendance au domaine

Le domaine peut également impacter la polarité et la valence. En effet, une expression subjective peut être positive dans un domaine et ne plus avoir la même polarité dans un autre domaine. Il existe des mots intrinsèquement positifs *délicieux* comme des mots intrinsèquement négatifs *mauvais*. D'autres ont une valeur neutre comme *court*. Toutefois, il existe également des mots (polysémiques comme homonymes) qui changent d'orientation selon le contexte dans lequel ils sont employés (Marchand, 2013, 2015). Par exemple, le terme *navet* qui est un légume est utilisé dans le domaine du cinéma pour signifier qu'un film est nul.

3.5.3 Les expressions figuratives

Les algorithmes actuels en traitement automatique du langage naturel ont du mal à détecter les subtilités du langage humain. Parmi ces subtilités, nous pouvons citer le langage figuratif. L'analyse du langage figuratif est un sujet difficile auquel le TAL doit faire face. Contrairement au langage littéral, le langage figuratif « détourne le sens propre pour lui conférer un sens dit figuré ou imagé, comme l'ironie, le sarcasme, l'humour, la métaphore ou encore les jeux de mots » (Benamara et al., 2017b). C'est un sujet de recherche en constante évolution notamment en raison de son importance pour l'évaluation et la performance des systèmes en analyse d'opinions (Benamara et al., 2017b; Ghosh et al., 2015b; Maynard et Greenwood, 2014).

Le langage figuratif intègre les différentes notions suivantes :

- l' **ironie** qui est une figure rhétorique par laquelle on dit le contraire de ce que l'on veut exprimer. Deux types d'ironie sont distingués : l'ironie conversationnelle et l'ironie textuelle. Par exemple, *Quelle belle réussite ! (de quelqu'un qui vient d'échouer)* ;
- le **sarcasme** qui est une façon de dire exagérément ce que l'on pense. Il permet d'exprimer la raillerie et donc la dévalorisation. Par exemple, *Quel chef-d'œuvre, le style rivalise avec Victor Hugo. (un critique pourrait écrire ceci à-propos d'un livre).*

Pour un état de l'art plus complet, les lecteurs peuvent se référer aux travaux de Karoui et al. (2019).

3.6 Définition et caractérisation des freins, des motivations et des conditions

Dans la section 2.3.2, nous avons souligné qu'un grand nombre d'entreprise fait face à un taux élevé d'échec avant ou après commercialisation de leurs nouveaux produits. Cet échec tire son origine de plusieurs causes notamment de la mauvaise connaissance ou de la connaissance insuffisante des potentiels utilisateurs. Dans l'optique de limiter les risques d'échec d'un produit nouveau avant son lancement effectif sur le marché, nous avons relevé en section 2.6.1 l'importance de réaliser des études d'usages pour mesurer l'acceptabilité des nouveaux produits par les potentiels utilisateurs. Ces études reposent sur une utilisation combinée d'outils de collecte spécifiques et de méthodes qualitatives propres à la société Ixiade mais issues des sciences sociales, du marketing et la sociologie des usages. Ces études comportent une phase de collecte des données réalisée au travers d'entretiens semi-directifs ou tables rondes avec des individus. Par ailleurs, les entretiens sont enregistrés, ensuite transcrits et enfin analysés. Cette analyse consiste à extraire manuellement les verbatims pertinents de cet amas de données textuelles et à assigner dans un premier temps à chacun des verbatims un des 21 critères de la méthode CAUTIC® et ensuite à les classifier selon un axe frein-motivation-condition (FMC).

Dans ce travail, nous nous focalisons sur la classification de verbatim en FMC. La présence d'un de ces trois éléments peut informer le chargé d'études sur les raisons qui pourraient pousser ou ne pas pousser un individu à adopter le concept ou l'innovation présentée. En TAL, le mot opinion a été le sujet de nombreux travaux (voir section 3.3 et 3.4) mais les freins, motivations et les motivations sous condition n'ont pas à notre connaissance été étudiés. Les freins et motivations ont fait

l'objet de travaux dans le domaine de la psychologie, du marketing et de la philosophie. Dans ce qui suit, nous nous intéresserons à ces travaux et nous proposons une définition pour ces notions pour notre domaine de travail.

3.6.1 Frein

Commençons par reprendre la définition générale du terme frein que propose l'encyclopédie du marketing (Bathelot). Ce dernier définit le frein comme :

« Frein : un élément matériel ou psychologique qui gêne ou empêche la décision d'achat ou d'utilisation d'un produit ou service ».

Nous pouvons ainsi avoir comme synonyme du terme frein, les termes *obstacle, empêchement, entrave et limitation*. Dans la littérature, particulièrement anglophone, le terme frein (*obstacle* en anglais) se substitue souvent au terme barrière (*barrier* en anglais) que l'on retrouve dans la majeure partie des travaux qui se sont penchés sur les raisons des échecs des innovations. Dans la suite de notre rédaction, nous utiliserons le terme frein à la place de barrière. L'étude des freins à l'adoption des innovations a fait l'objet de nombreux travaux (Ram et Sheth, 1989; Chemingui et Ben lallouna, 2013; Laukkanen et al., 2008; Kuisma et al., 2007; Mani et Chouk, 2017). Ces travaux ont essayé de déterminer ou d'expliquer les causes de la résistance des consommateurs aux innovations. Dans la suite de cette section, nous en détaillons quelques uns.

Les consommateurs dans les pays industrialisés manifestent de l'intérêt pour les innovations ; néanmoins une grande majorité des innovations échouent sur le marché (Ram, 1987). Ram et Sheth (1989) prennent l'exemple du *Videotex* qui offrait la possibilité de faire du shopping depuis chez soi. À son lancement, il a rencontré de fortes résistances de la part d'une majorité de consommateurs à cause des changements qu'il a créés dans les pratiques de shopping des consommateurs. Les consommateurs ne pouvaient plus interagir avec le personnel des magasins afin de recevoir des informations utiles. Ceux qui aimaient faire du shopping avec la famille ou les ami(e)s ont été privés de ces interactions sociales. Les consommateurs ont dû apprendre à utiliser ce type d'innovation. Cet exemple montre clairement que des changements aux *statu quo* peuvent amener des consommateurs à résister à une innovation. D'un autre côté, une innovation peut entrer en conflit avec les systèmes de croyance des consommateurs. Par exemple, une bonne partie des consommateurs américains pensent que les produits venant des pays du tiers monde sont de moins bonne qualité. D'autres pensent qu'acheter des produits étrangers est

antipatriotique. Par exemple, la campagne autour du slogan « consommer français » montre que les individus favorables à cette démarche sont susceptibles de résister à toute innovation provenant d'un pays étranger.

Dans cette perspective, Ram (1987) propose ainsi la théorie de la résistance. La théorie de la *résistance à l'innovation* désigne la résistance ou l'opposition manifestée par les consommateurs envers une innovation, soit parce qu'elle amène des changements potentiels par rapport à un *statu quo*, soit parce qu'elle se heurte à leurs croyances. Cette résistance s'accompagne d'une non-adoption de l'innovation. Ram et Sheth (1989) expliquent que les consommateurs font face à différents freins qui les empêchent d'adopter des innovations. Ils les regroupent en freins fonctionnels et psychologiques (Joachim et al., 2018).

Les freins fonctionnels sont de trois catégories : l'utilisation, la valeur et le risque.

1. l'utilisation intervient lorsque l'innovation est incompatible avec les pratiques et les habitudes existantes de l'individu. En outre, l'innovation peut également nécessiter plus d'efforts dans la manipulation. Kuisma et al. (2007) ont montré que certains non-utilisateurs des services bancaires en ligne les trouvent difficiles et lents à utiliser. Pour d'autres encore, le problème de l'utilisation se trouve au niveau des fonctions que présente l'innovation et qui peuvent être inadéquates (forme du clavier ou dispositif d'affichage) ;

2. la valeur représente l'aspect financier. De manière générale, les potentiels utilisateurs d'une innovation vont comparer les fonctions que présente l'innovation par rapport à son prix et également à ses substituts présents sur le marché. Si le prix proposé ne correspond pas à ce que l'innovation offre, ils vont considérer que l'acquisition n'est pas rentable, voire intéressante pour eux ;

3. le risque fait référence au degré de risque qu'une innovation implique. Le doute étant inhérent aux innovations, ces dernières impliquent très souvent des risques. Le risque peut être physique, dans ce cas l'utilisateur pourrait envisager ne pas acquérir une innovation de peur de se blesser lui-même ou de blesser l'autre, voire endommager ses biens. Il peut être aussi social, l'individu craindra d'être jugé ou d'être isolé par la société, et également économique, l'individu aura peur de perdre de l'argent ou ne ressentira pas tout simplement le besoin de dépenser de l'argent dans l'acquisition de l'innovation. Le risque peut être aussi fonctionnel, ici l'inquiétude sera liée au bon fonctionnement de l'objet.

Les freins psychologiques sont répartis en deux catégories qui sont la tradition et l'image.

1. si une innovation est incompatible avec les valeurs coutumes d'un individu ou de sa famille, il y a très peu de chance qu'elle soit acceptée. De même, si l'innovation implique des changements dans la routine de vie de l'individu, elle peut être également un obstacle. Les innovations qui semblent se rapprocher des normes traditionnelles sont plus acceptées par les individus que celles qui dévient (Herbig et Day, 1992) ;
2. l'image est liée à l'origine du produit. Il s'agit ici de la marque ou l'entreprise impliquée dans la fabrication du produit ou le pays dans lequel le produit est fabriqué.

La théorie de la résistance à l'innovation a ainsi motivé plusieurs travaux particulièrement dans le domaine des services bancaires en ligne comme ceux de Kuisma et al. (2007) et Laukkanen et al. (2008) qui décrivent des facteurs qui contribuent à la réticence des consommateurs à utiliser des services bancaires mobiles en ligne. Kleijnen et al. (2009) examinent la résistance des consommateurs en expliquant ses « principaux composants » à savoir : le rejet, le report et l'opposition. Dans la même perspective, Zhou (2011) s'intéresse aux facteurs affectant l'intention d'utilisation des services bancaires mobiles en ligne comme la confiance initiale et l'utilité perçue. Chemingui et Ben lallouna (2013) concluent dans leurs travaux que la tradition a un impact négatif et significatif sur l'intention d'utiliser les services financiers mobiles. Ils identifient également 4 dimensions motivationnelles : la compatibilité, la testabilité, le plaisir perçu et la qualité du système qui ont tous un impact significatif et positif sur l'intention d'utiliser un service mobile financier.

Si nous reprenons tout ce qui a été décrit plus haut et en nous appuyant sur les travaux de Ram et Sheth (1989), nous définissons les freins comme des attitudes psychologiques ou morales qui amènent l'individu à résister à une innovation même si celle-ci est considérée nécessaire et désirable. Les freins sont présents dans le discours des individus interviewés dans le cadre d'une étude. Elles traduisent les raisons pour lesquelles l'individu ne se projette pas dans une acceptabilité (section 2.4.2) de l'innovation ou les raisons qui retarderaient cette acceptabilité. Ces raisons sont habituellement liées au changement qu'apporte l'innovation ou le conflit que l'innovation créé avec les systèmes de croyance de l'individu. Dans notre contexte de travail, nous nous intéressons ainsi à la détection de ces freins dans le discours des individus interrogés lors d'entretiens semi-directifs ou tables rondes.

En s'appuyant sur les travaux de Vernier (2011), Kerbrat et al. (1980) et de Ja-

rukan (2014), et des travaux présentés dans cette section, nous définirons un frein de la manière suivante. Un frein peut être considéré comme une expression subjective (dont une opinion) que l'émetteur utilise pour évaluer, juger, apprécier un objet en révélant les raisons pour lesquelles il ne se projette pas dans l'usage ou l'adoption de l'objet (produit) en question. Ces raisons peuvent concerner l'objet dans son entièreté ou simplement un élément de l'objet.

Exemples d'expression de freins

Les verbatims suivants tirés de notre corpus d'apprentissage sont des exemples auxquels la classe *frein* a été attribuée.
(1) *Ce qui me dérange, c'est le tout connecté, avec les ondes.*
(2) *Ce module qui est un gadget, mon problème, c'est qu'il soit raccordé à la puissance car si j'ai un problème dessus, je dois couper la puissance.*
(3) *Le but, c'est toujours de gagner du temps. Donc il faut qu'il y ait une personne avec un IPAD accroché au cou et qui regarde en permanence... C'est anti-social ça.*
(4) *Viser un public jeune avec De Dietrich cela paraît difficile. Ce sont des produits chers.*
(5) *Donc ce n'est pas adaptable à de l'ancien, dans ces cas-là il faut tout refaire.*

3.6.2 Les motivations

L'une des difficultés majeures de la psychologie de la motivation est l'absence de consensus sur sa définition. Pour tenter de résoudre cette confusion terminologique, Kleijnen et al. (2009) compilent 102 déclarations à partir de sources diverses définissant ou critiquant le concept. Huitt (2001) citant Kleijnen et al. (2009) déclare que les définitions suivantes de la motivation reflètent le consensus général selon lequel la motivation est un état ou une condition interne (parfois décrite comme un besoin, un désir ou une envie) qui sert à activer ou à dynamiser le comportement et à lui donner une direction :

« Motivation : état ou condition interne qui active le comportement et lui donne une direction » ;
« Motivation : désir ou envie qui dynamise et oriente le comportement vers un but » ;
« Motivation : influence des besoins et des désirs sur l'intensité et la direction du comportement ».

L'élaboration d'une liste exhaustive des motivations humaines a été l'un des axes de travail sur la motivation. Cependant, il n'existe pas de consensus sur une telle liste. Au fil des années, plusieurs suggestions ont été présentées dans la littérature. Dans les années 1900, MacDougall (1933) propose 12 instincts innés (i.e le dégoût, la fuite, la curiosité, la reproduction, etc.) qui poussent l'individu à atteindre 12 buts spécifiques avec leurs énergies motivationnelles. Talevich et al. (2017) présentent dans leurs travaux une taxonomie hiérarchique structurée empiriquement et théoriquement de 161 motivations tirées d'une analyse documentaire de différents travaux (Schank et Abelson, 2013; Fiske et al., 2009; Bernhard et Ligozat, 2011; Kenrick et al., 2010; Murray, 1938; Bugental, 2000). L'éventail des motivations qui ont été proposées est large ainsi que la variabilité de leur structure. Les travaux énoncés ci-dessus montrent clairement qu'il manque un consensus clair sur une typologie générale des motivations humaines et de leurs structures.

La théorie de la motivation, initialement présentée par Richard Deci en 1975 et enrichie par Deci et Ryan (1985, 2002) avec la théorie de *l'auto-détermination* se distingue des autres théories en identifiant deux catégories de motivation : la motivation intrinsèque (internes à la personne) et la motivation extrinsèque (extérieures à la personne). Une personne motivée intrinsèquement fait quelque chose non pas à cause d'un évènement externe ou une raison extérieure, mais tout simplement par plaisir ou envie. La motivation extrinsèque quant à elle caractérise une action provoquée par un évènement extérieur. les personnes extrinsèquement motivées auront tendance à faire des choses dans le but d'obtenir un résultat.

Ils introduisent également le concept de l'amotivation. Lorsqu'il est amotivé, le comportement d'une personne manque d'intentionnalité et d'un sens de causalité personnelle. L'amotivation résulte du fait de ne pas valoriser une activité ou de ne pas se sentir capable de le faire, ou encore de ne pas croire qu'elle produira un résultat souhaité. La figure 3.5 illustre la taxonomie des types de motivation proposée par Deci et Ryan (1985).

Dans le cadre de la théorie de *l'auto-détermination*, deux sous-théories, appelées respectivement la *théorie de l'évaluation cognitive* et la *théorie de l'intégration organique* en français, ont été introduites pour détailler les différentes formes de motivation intrinsèque et extrinsèque et pour préciser les facteurs dans les contextes sociaux qui favorisent ou entravent ces deux types de motivation (Deci et Ryan, 1985). Ainsi, il existerait des facteurs ou forces qui inciteraient un individu à adopter un comportement.

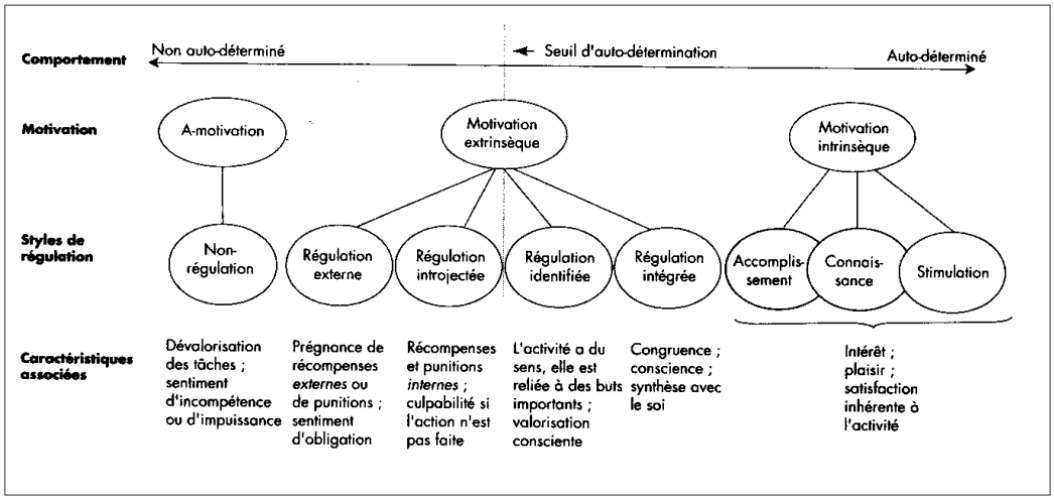


FIGURE 3.5 – Taxinomie des motivations en français (Sarrazin et Trouilloud, 2006).

Pour définir une motivation, nous reprenons certains éléments mentionnés dans la définition apportée pour qualifier un frein. Une motivation correspond à une expression subjective (dont une opinion) que l'émetteur utilise pour évaluer, juger, apprécier un objet en révélant les raisons pour lesquelles il se projette dans l'usage ou l'adoption de l'objet (produit) en question. Ces raisons peuvent concerner l'objet dans son entièreté ou simplement un élément de l'objet. L'émetteur peut être présent explicitement (usage de pronoms personnels, adjectifs possessifs, etc.) comme implicitement (pas de traces de l'émetteur) dans son énoncé.

Exemples d'expression de motivation

Les verbatims suivants tirés de notre corpus d'apprentissage sont des exemples auxquels la classe *motivation* a été attribuée.

(1) *Sur les stations de pompage, c'est très intéressant, car milieu pollué, corrosif ?*

(2) *Ça, c'est bien pensé. La présentation couleur, photographique est belle, attrayante. L'évocation photographique dans la démo là-bas est intéressante. La double entrée : le manuel et l'informatique, c'est vraiment intéressant. Voilà, je n'ai pas expérimenté. Techniquement, ça me paraît très bien fait et abordable en dehors d'un professionnel, c'est ça qui est intéressant aussi.*

(3) *Après, c'est sûr que si sur un sac comme ça on arrive à avoir... le sous vide tout le temps, peut-être que ça sera plus facile au niveau de la pression.*

(4) *Aujourd'hui on rentre chez les gens avec ça. C'est ça ? Sur des appareils que tout le monde connaît, on n'est pas devant une armoire électrique en train de tamponner ! Là on est sur des tablettes, des téléphones, que tout le monde utilise. On se rapproche plus des outils de la vie quotidienne.*

3.6.3 Les motivations sous conditions

Les motivations sous conditions ou tout simplement les conditions sont une troisième catégorie de classification. De manière générale, une condition est un « État, situation, fait dont l'existence est indispensable pour qu'un autre état, un autre fait existe [19].» Pour notre contexte, les conditions désignent les opinions dans lesquelles l'émetteur utilise, évalue, juge, apprécie un objet en révélant les raisons pour lesquelles il se projetterait dans l'usage ou l'adoption de l'objet (produit) en question. Ces raisons sont généralement des conditions ou des exigences qu'il énonce dans son discours. Dans la suite, nous utiliserons le terme *condition*.

Exemples d'expression de condition

Les verbatims suivants tirés de notre corpus d'apprentissage sont des exemples auxquels la classe *condition* a été attribuée.

(1) *Dans le cadre d'un diagnostic plus large, si l'équipement est proposé par le prestataire et que le matériau est plus performant, ça peut être utile. On n'a pas toujours la performance de la paroi.*

(2) *Il faudrait qu'il soit... après je pense que ça demande beaucoup, qu'il soit capable de dire : voilà, tant que le courant ne monte pas au-dessus de cette valeur, je ne disjoncte pas. Donc ce ne serait plus vraiment de la notion de choisir un appareil en fonction d'une puissance de moteur, mais en fonction d'un courant d'appel de moteur.*

(3) *Je tiens à ce qu'on ait les 2 : comme en local et à distance. Au niveau des fonctions : il faudrait pouvoir choisir si on ne veut que la mesure d'énergie et*

19. https://dictionnaire.lerobert.com/definition/condition

n'avoir que ça.

(4) *il faudrait qu'au sein du logiciel, qu'on ait déjà des boîtes de dialogue... déjà réalisées, dans une bibliothèque constructeur associée à chaque appareil, où ça nous remonte directement les informations et qu'on n'ait plus qu'à... à les exploiter, oui. Ça, c'est important, oui.*

(5) *Si on peut trouver l'ampoule dans le commerce c'est bien mais sinon c'est contraignant d'acheter que le bloc. Je ne sais pas si on peut acheter une ampoule d'une autre marque, ça serait mieux, pour pas avoir l'impression d'être verrouillée par une marque.*

3.7 Traitement automatique des freins, motivations et conditions

Lorsque les motivations ou les freins sont étudiés en marketing ou en psychologie, les méthodes proposées diffèrent les unes des autres, car elles sont très dépendantes du domaine ou de l'objet d'étude. Certains travaux auront recours à des méthodes quantitatives reposant sur des questionnaires en ligne et fournissant des données statistiques (Joiret et al., 2019; Chemingui et Ben lallouna, 2013). D'autres par contre s'appuieront sur des méthodes qualitatives reposant sur des entretiens semi-directifs ou tables rondes combinés à des grilles d'analyse ou de codification pour l'analyse des verbatims transcrits (Pilon-Caron, 2015; Goreux et Jeandrain). Toutefois, les grilles d'analyse ou de codification seront toujours différentes les unes des autres. On peut citer la grille d'analyse développée pour l'étude des freins et des motivations à l'utilisation des services Airbnb (Pilon-Caron, 2015) ou encore l'analyse des freins et motivations des consommateurs à la pratique du shopping dans les univers virtuels (Goreux et Jeandrain). Ainsi, pour les méthodes qualitatives, les verbatims transcrits et jugés pertinents sont assignés à un segment de la grille. D'autres méthodes d'analyse peuvent être utilisées telles qu'une analyse thématique.

Néanmoins, nous constatons que certaines variables motivationnelles ou de freins reviennent assez souvent dans les différentes études de l'état de l'art notamment le prix, la culture, l'utilisation, etc. Par contre, il n'existe pas encore, à notre connaissance des travaux concernant les *motivations sous condition*. Les travaux mobilisés en psychologie et en marketing et énoncés dans cette section nous ont permis de proposer des définitions aux notions de freins, motivations et conditions dans le cadre de notre travail. Notre objectif étant l'identification des freins, motivations et

conditions dans des entretiens/focus group et à long terme dans du contenu en ligne, nous décidons de traiter le problème comme une tâche de classification.

3.8 Conclusion

Dans ce chapitre, nous avons présenté l'état de l'art de l'analyse d'opinion en partant de son origine jusqu'à l'exposition des challenges rencontrés dans ce domaine. Par la suite, nous avons introduit les notions de frein, motivation et condition en nous appuyant sur la littérature scientifique en psychologie et marketing. Finalement, nous avons conclu en soulignant que nous traiterons le problème de l'identification des FMC comme une tâche de classification. L'objectif de notre travail est de proposer une approche pour la détection automatique de freins, motivations et conditions dans divers types de données. Ces données sont réparties en deux groupes : d'une part des données orales transcrites issues d'entretiens semi-directifs et tables rondes, et d'autre part des données issues de la plateforme de collecte Yoomaneo. Pour cela, nous avons introduit au premier chapitre le contexte industriel dans lequel le présent travail s'inscrit et, dans le deuxième chapitre le domaine de l'analyse d'opinion. Nous avons également proposé des définitions pour nos classes principales avec des exemples issus de notre corpus d'entraînement. Dans le prochain chapitre, nous introduisons les principales méthodes de représentations des textes en nous attardant sur celles utilisées dans ce travail de recherche.

Chapitre 4

Méthodes de représentation

Les précédents chapitres ont introduit le contexte de recherche de notre travail et l'état de l'art en analyse de sentiment. Comme nous l'avons mentionné au chapitre 3, l'analyse de sentiment est une tâche de classification qui consiste à attribuer à des documents ou des textes une polarité précise à l'aide de différentes méthodes. Ces méthodes peuvent être supervisées comme non supervisées. Dans le présent chapitre, nous détaillons de la section 4.2 à 4.4 les différents types de représentations des documents.

4.1 Introduction

Pour analyser des données textuelles, plusieurs étapes peuvent intervenir comme le montre la figure 4.1 :

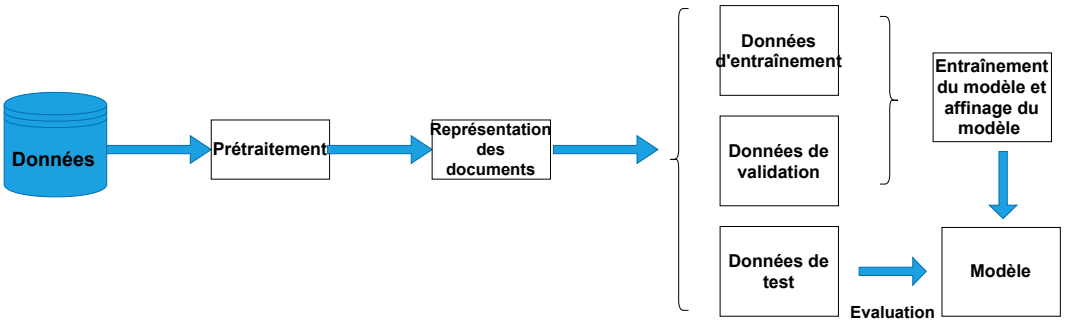


FIGURE 4.1 – Processus d'apprentissage automatique (Nzali, 2017; Mercadier, 2020).

1. une phase de **prétraitement** qui consiste à supprimer les mots vides, la ponctuation, les caractères spéciaux, à lemmatiser les mots, etc ;
2. une phase de **représentation des documents** ou de *feature engineering* qui consiste à transformer les textes en des valeurs numériques ;
3. une phase d'**entraînement** et **construction du modèle** qui consiste à entraîner et affiner différents modèles à partir d'un corpus d'apprentissage et de validation ;
4. une phase d'**évaluation du modèle** qui consiste à évaluer les performances du modèle sélectionné à l'étape précédente sur un nouveau jeu de données (test).

La phase de prétraitement à fait l'objet d'une sous-section au chapitre 3 en section 3.4.1. Nous détaillons les principales méthodes de représentations des données dans ce qui suit.

4.2 Représentations traditionnelles

La phrase présentée en table 4.1 est compréhensible par un locuteur français. Par contre, les algorithmes d'apprentissage ont besoin d'une entrée sous la forme de représentations vectorielles. La phase de représentation des documents est ainsi primordiale, car elle permet de transformer les données textuelles en un format numérique interprétable par les algorithmes d'apprentissage.

Example 1 : *Je dirais que c'est une appli innovante qui va tout à fait dans les développements futurs avec la possibilité de connectivité à distance et que c'est génial !*

TABLE 4.1 – Exemple d'une phrase prise dans notre corpus d'apprentissage.

Cette section détaille deux des méthodes traditionnelles les plus utilisées pour représenter un texte : la méthode sac-de mots et le TF-IDF.

4.2.1 Représentation par sac de mots

La **représentation en sac de mots** est une méthode pionnière de la représentation des documents. C'est une méthode de représentation des documents très populaire et très utilisée en TAL. Elle est également très simple à mettre en place. Cette méthode consiste à compter les occurrences d'un token dans un texte. Le but est

de trouver tous les mots discriminants du texte en fonction de leur présence ou absence. Pour mieux la comprendre nous utilisons les exemples présents dans la table 4.2.

Document 1 : *Ce design est super intéressant.*
Document 2 : *Je trouve ce design chouette.*
Document 3 : *Ce design est moche.*

TABLE 4.2 – Exemples de documents.

Comme évoqué plus haut, ces phrases doivent être représentées sous la forme de vecteurs. Dans un premier temps, nous appliquons une phase de prétraitement. Après cette phase, nous obtenons un dictionnaire unique de tous les mots des documents de notre exemple représenté par $W = \{w_1....w_n\}$. Ainsi, pour chaque document, nous construisons un vecteur de type $d = \{d_1....d_n\}$ qui correspond au nombre d'occurrences des mots présents dans le document. Si le terme est présent on met 1, sinon 0. Le vocabulaire obtenu constitué des mots suivants : *design, être, super, intéressant, chouette, moche*. Ensuite, pour chaque document, nous obtenons la table présentée en 4.3.

	Doc 1	Doc 2	Doc 3
design	1	1	1
être	1	0	1
super	1	0	0
intéressant	1	0	0
trouver	0	1	0
chouette	0	1	0
moche	0	0	1

TABLE 4.3 – Exemple de représentation des mots dans différents documents en utilisant la méthode - Sac de mots.

Cette méthode présente également quelques inconvénients :

— la dimensionnalité de la représentation d'un document dépend de la taille du vocabulaire ;

— une grande majorité des valeurs présentes dans la matrice de représentation ont des zéros ce qui résulte en une matrice clairsemée ;

— aucune information sur la grammaire des phrases ou l'ordre des mots dans les documents n'est retenue.

4.2.2 Représentation par TF-IDF

Le **TF-IDF** (*Term Frequency – Inverse Document Frequency*) est une mesure statistique qui reflète l'importance d'un terme pour un document dans une collection de documents. Il fait appel à deux notions : le **TF** (*Term Frequency*) et l'**IDF** (*Inverse document frequency*). Pour un document d, le **TF-IDF** est calculé comme suit :

$$\textbf{TF-IDF}(w_i, d) = tf(w_i, d) * idf(w_i) \tag{4.1}$$

avec $\text{tf}(w_i, d)$ qui correspond au nombre d'occurrences du terme w_i dans le document d et $\text{idf}(w_i)$ la fréquence inverse de document du terme w_i calculée comme suit :

$$\textbf{IDF}(w_i) = \log \frac{|\mathcal{D}|}{\text{df}(w_i)}, \tag{4.2}$$

avec |D| correspondant au nombre total de documents dans une collection et $\text{df}(w_i)$ qui correspond au nombre de documents contenant le terme (w_i). L'**IDF** d'un terme est élevée lorsque ce dernier apparait peu fréquemment dans la collection de documents. Le terme est ainsi considéré comme rare. Par contre, si elle est basse, cela signifie que le mot apparait très fréquemment. C'est le cas des mots tels que *le, de, etc.* Son utilisation permet ainsi de donner plus d'importance aux mots rares. La table 4.4 affiche les valeurs **TF-IDF** de chaque mot pour chaque document en utilisant les exemples de la table 4.2.

	Doc 1	Doc 2	Doc 3
design	0.345205	0.385372	0.425441
être	0.444514	0.000000	0.547832
super	0.584483	0.000000	0.000000
intéressant	0.584483	0.000000	0.000000
trouver	0.000000	0.652491	0.000000
chouette	0.000000	0.652491	0.000000
moche	0.000000	0.000000	0.720333

TABLE 4.4 – Exemple de représentation des mots dans différents documents en utilisant la méthode - TF-IDF.

Tout comme la représentation en sac de mots, le **TF-IDF** présente l'inconvénient de créer des matrices très clairsemées.

4.3 Représentation continue des mots : « Plongement lexical »

4.3.1 Modèles pré-entraînés

Avant d'introduire les différentes méthodes de représentation continues des mots, il est important d'introduire la notion de **modèles pré-entraînés**. Le recours aux modèles pré-entraînés a permis ces dernières années d'améliorer les performances des classifieurs pour différentes tâches comme l'analyse de sentiment. Un modèle pré-entraîné est un modèle générique conçu pour résoudre une tâche spécifique et qui peut être réutilisé sur une autre tâche. La fabrication de modèles requiert de grandes quantités de données et de grandes capacités de calculs. L'usage des modèles pré-entraînés permet d'utiliser directement les poids et l'architecture du modèle obtenu après entraînement de celui-ci pour résoudre n'importe quelle tâche. Cela s'appelle de l'apprentissage par transfert, car nous allons tout simplement transférer au moment de l'utilisation ce que le modèle a appris sur notre propre tâche. Par exemple, un modèle précédemment entraîné sur une tache de détection peut très bien être utilisé ensuite pour une tâche similaire. À ce jour, plusieurs modèles pré-entraînés sont directement accessibles et disponibles via des librairies telles que *Keras* [1], *spacy* [2] ou encore ceux disponibles sur le concentrateur *Hugging Face* [3].

La représentation vectorielle continue d'un mot désigne une représentation apprise pour un texte où les mots qui ont la même distribution ont une représentation similaire. Ainsi, chaque mot est représenté par un vecteur de nombres réels, d'une dizaine ou centaines de dimensions qui contraste avec les milliers ou millions de dimensions requises pour la représentation en sac de mots. Les représentations continues de mots ont été proposées pour pallier les problèmes rencontrés avec les représentations de type sac de mots ou TF-IDF. Ces problèmes particulièrement complexes à gérer concernaient :

— la taille des vecteurs qui était importante ;

— les représentations obtenues qui étaient particulièrement creuses, c'est-à-dire qu'
elles contenaient beaucoup de zéros ou valeurs nulles ;

1. https://keras.io/
2. https://spacy.io/usage/models
3. https://huggingface.co/docs/transformers/index

— la taille des matrices qui était également assez importante.

Les modèles de représentation continue de mots ont été créées en se basant sur la linguistique distributionnelle (Harris, 1954; Firth, 1957) qui se résume à l'idée suivante : les mots utilisés dans les mêmes contextes auront généralement des sens similaires. Les méthodes en représentation vectorielle ou continue de mots apprennent une représentation vectorielle pour un vocabulaire prédéfini de taille fixe à partir d'un corpus de texte. Ce processus d'apprentissage est soit lié à un réseau de neurones pour une tâche spécifique telle que la classification de documents, soit à un processus non supervisé utilisant les statistiques du document. Pour la classification de documents, les représentations vectorielles de mots ont essentiellement été utilisées comme entrée des modèles neuronaux. Dans les sous-sections suivantes, nous détaillerons trois des méthodes phares de représentation continue des mots.

4.3.2 Word2vec

Word2vec (Mikolov et al., 2013a,b) est un algorithme reposant sur des réseaux de neurones peu profond à trois couches [4] utilisé pour apprendre les représentations vectorielles des mots d'un texte. Il prend en entrée un large corpus de texte et produit à la sortie un ensemble de vecteurs dans un espace vectoriel. Chaque vecteur de cet espace est une représentation vectorielle de chaque mot unique du vocabulaire du corpus qui a été apprise selon deux méthodes. Dans la suite, nous présentons succinctement les deux architectures de Word2vec.

- Le modèle **CBOW (Mikolov et al., 2013a) « Continuous Bag of Words »** dont une figure est donnée en 4.2 a l'objectif de prédire un mot cible en fonction du contexte. Le contexte représente ici la fenêtre de voisinage [5] (dont la taille est paramétrable) du mot, c'est-à-dire les mots qui l'entourent. L'architecture CBOW prend en entrée le contexte de mots et est entraîné à prédire le mot correspondant. L'algorithme parcourt ainsi fenêtre après fenêtre l'ensemble du corpus pour apprendre la représentation des mots. Plus spécifiquement, l'entrée du modèle est un vecteur *one hot* [6] et la sortie est également un vecteur qui correspond à une probabilité d'occurrence pour chaque mot du corpus. Le mot du vocabulaire ayant la plus grande probabilité est prédit.

4. La première couche est la couche d'entrée, la couche du milieu qui est la couche cachée et la dernière couche qui est la couche de sortie.

5. La fenêtre de voisinage correspond au nombre de mots environnants à prendre en compte pour l'entraînement du modèle.

6. Vecteur de valeur binaire de 1 ou 0.

Cette probabilité est obtenue en utilisant une fonction d'activation (la fonction softmax) sur la couche de sortie.

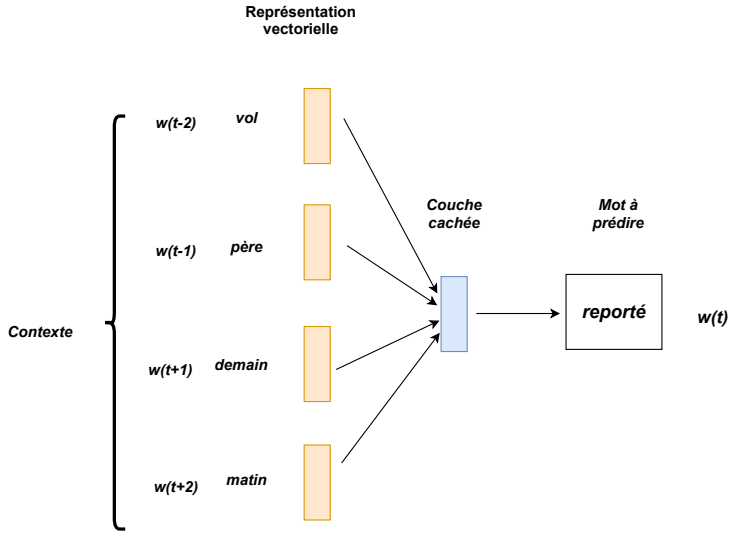


FIGURE 4.2 – Architecture du modèle CBOW (Mikolov et al., 2013a) avec l'exemple *« Le **vol** de mon **père** est **reporté** à **demain matin** »*.

Le but du modèle est donc de minimiser la fonction d'erreur de log probabilité comme observé sur l'équation 4.3.

Équation CBOW (Mikolov et al., 2013c)

$$= \sum_{t=1}^{T} \log \mathbb{P}\left(w_t | w_{t-m}, ...w_{t-1}, w_{t+1}, ...w_{t+m}\right) \tag{4.3}$$

- Le modèle **Skip-gram** (Mikolov et al., 2013a,b) fait les choses inversement en devinant à l'avance les mots voisins à partir d'un mot cible. Le mot cible est donné en entrée du réseau et les mots du contexte en sont les prédictions. La couche de sortie, au lieu de produire un seul terme, en produit plusieurs selon la taille fixée. Étant donné le terme *reporté* noté w_t qui est notre entrée, le modèle Skip-gram (voir figure 4.3) doit prédire les mots de son voisinage. Cette probabilité d'observer un contexte de mot c_t résulte du produit scalaire entre le vecteur de mot et le vecteur de contexte auquel on applique une fonction softmax au niveau de la couche de sortie. Plus spécifiquement, étant donné un grand corpus d'apprentissage représenté comme une séquence de

mots $w_1, ..., w_m$, l'objectif du modèle est de minimiser la fonction d'erreur de log probabilité dont la formule est donnée à l'équation 4.4.

Équation Skip-gram (Mikolov et al., 2013b)

$$= \sum_{t=1}^{T} \sum_{c \in C_t} \log \mathbb{P}\left(w_c | w_t\right) \tag{4.4}$$

où C_t représente le contexte des mots entourant le mot d'entrée w_t.

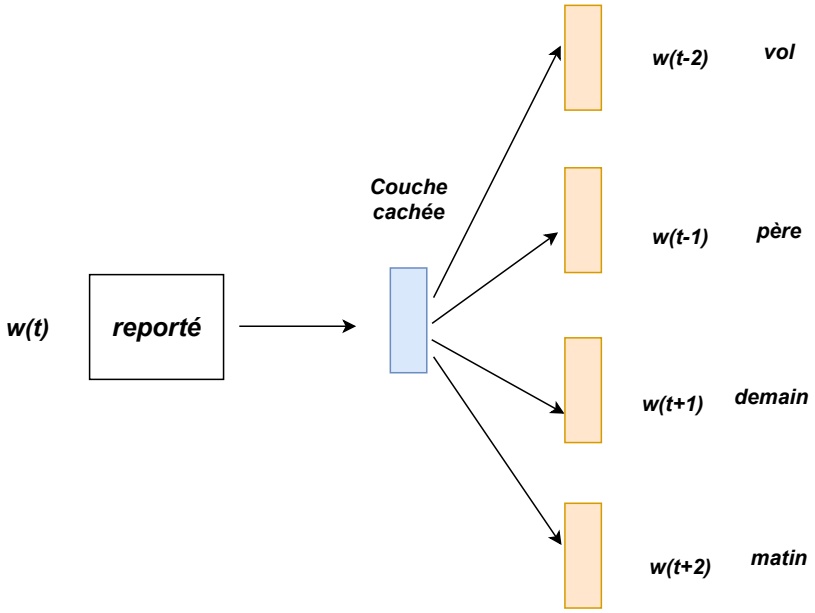


FIGURE 4.3 – Architecture du modèle Skip-gram (Mikolov et al., 2013a) avec l'exemple « *Le* ***vol*** *de mon* ***père*** *est* ***reporté*** *à* ***demain matin*** ».

Les deux architectures de Word2vec [7] ont été entraînées sur environ 100 milliards de mots provenant de nouvelles de Google [8]. Word2vec a marqué un grand pas dans le domaine des représentations vectorielles en TAL [9]. Selon les auteurs, le modèle CBOW est plus rapide, tandis que le skip-gram donne de bien meilleurs résultats et particulièrement pour les mots peu fréquents. Néanmoins, plusieurs variantes de Word2vec ont par la suite été proposées dont le modèle *Glove* (Pennington et al., 2014) que nous présentons dans la sous-section suivante.

7. https://github.com/tmikolov/word2vec/blob/master/word2vec.c
8. Ces corpus sont malheureusement non disponibles.
9. Au 20 janvier 2022, l'article de Mikolov et al. (2013a) a été cité 31562 fois.

4.3.3 Glove

Glove (Pennington et al., 2014) qui vient de *Global Vectors* est un modèle de représentation distribuée des mots. Glove est un modèle de régression log-bilinéaire pour l'apprentissage non supervisé des représentations vectorielles des mots en alliant des méthodes de comptage et des méthodes neuronales reposant sur des fenêtres de contexte.

Au niveau du fonctionnement, l'algorithme Glove comprend trois phases importantes :

1. Définition d'une **matrice de cooccurrence des mots** notée X. Chaque entrée X_{ij} de cette matrice représente le nombre de fois que le mot i apparaît dans le contexte du mot j. Autrement dit la matrice de cette cooccurrence permet d'observer la fréquence d'apparition d'une paire de mots. Pour chaque terme, le modèle recherche les termes contextuels dans une zone définie par une taille de fenêtre avant le terme et une taille de fenêtre après le terme. Les mots les plus éloignés reçoivent moins de poids.
2. Définition d'une **contrainte souple** (Pennington et al., 2014; Selivanov, 2020) pour chaque paire de telle manière que le vecteur du mot principal w_i et le vecteur du mot de contexte w_j sont des biais scalaires (b_i , b_j) pour le mot principal et de contexte.

$$\left(w_T^i w_j + b_i + b_j\right) = \log X_{ij} \tag{4.5}$$

3. Définition d'une **fonction de coût** (Pennington et al., 2014; Selivanov, 2020) :

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f\left(X_{ij}\right) \left(w_T^i w_j + b_i + b_j + \log X_{ij}\right)^2 \tag{4.6}$$

où V est la taille du vocabulaire. La fonction de pondération utilisée par les auteurs est la suivante :

$$f\left(X_{ij}\right) = \begin{cases} \left(X_{ij}/x_{max}\right)^{\alpha} & Si \quad x < x_{max} \\ 1 & sinon \end{cases} \tag{4.7}$$

Glove a été proposé avec l'intuition que la cooccurrence des mots dans un corpus peut révéler des informations sur leurs significations comme le fait qu'ils peuvent être similaires ou opposés. De plus, les auteurs de Glove soutiennent que les modèles CBOW et Skip-gram n'exploitent pas pleinement les informations statistiques relatives aux cooccurrences des mots, mais plutôt parcourent le corpus de

manière itérative, examinent quelques mots à la fois et essaient de prédire le mot du milieu à partir des mots qui l'entourent ou les mots qui l'entourent à partir du mot du milieu. Cette démarche ne leur permet pas de tirer parti de la grande quantité de répétitions dans les données comme Glove.

4.3.4 FastText

FastText (Joulin et al., 2016) est une extension de Word2vec fondée sur le modèle Skip-gram où chaque mot est représenté comme un ensemble de n-grammes de caractères. Étant donné que le modèle Skip-gram ne prend pas en compte les informations relatives aux sous-mots, FastText a été proposé pour pallier ce problème. Chaque mot w est représenté par un ensemble de caractères n-grammes. Des symboles spéciaux < et > sont ajoutés au début et à la fin de chaque mot permettant de distinguer les préfixes des suffixes des autres séquences de caractères. Le mot w est également inclus dans l'ensemble de ces n-grammes afin d'apprendre une représentation de chaque mot en plus des n-grammes de caractères. En prenant le mot *nocturne* où n=3, il sera représenté par les n-grammes de caractères : < *no, noc, oct, ctu, tur, urn, rne, ne* > et <nocturne>. De cette manière, toutes sortes de n-grammes sont considérées.

Supposons que l'on a un dictionnaire de n-grammes de taille G, étant donné un mot w, G_w correspond à l'ensemble des n-grammes qui apparaissent dans w. Un vecteur de représentation Z_g est ainsi associé à chaque n-gramme g. Un mot est donc représenté par la somme des représentations vectorielles de ses n-grammes selon la fonction suivante :

$$s(w,c) = \sum_{g \in G_w} Z_g^T V_c. \tag{4.8}$$

FastText [10] permet ainsi de partager les représentations entre les mots, ce qui permet d'apprendre des représentations fiables pour les mots rares. Il a été entraîné sur des données de Wikipédia en 157 langues : arabe, tchèque, allemand, anglais, espagnol, français, italien, roumain, russe, etc.

4.3.5 Conclusion

Les modèles tels que Word2vec et ses extensions (FastText et Glove) ont montré qu'il était possible d'apprendre efficacement des représentations vectorielles distributionnelles des mots de petites dimensions, c'est-à-dire de montrer qu'une paire

10. https ://fasttext.cc/docs/en/crawl-vectors.html

de mots comme « France » et « Paris » avait la même relation que « Allemagne » et « Berlin », « avait » et « a » a la même relation (syntaxique) qu' « était » et « est ». Ces modèles ont été mis à la disposition de la communauté scientifique. Il est ainsi possible de télécharger et d'utiliser une liste de mots et leurs vecteurs générés par ces modèles dans des tâches en TAL. Entraîner ses propres représentations reste une autre option et permet également d'obtenir des vecteurs spécifiques à son corpus et à la tâche que l'on souhaite réaliser.

4.4 Représentations contextualisées des mots

Les représentations continues de mots (voir section 4.3.1) ou plongement de mots ont longtemps été utilisées comme un composant standard des architectures en TAL. Ceci s'explique par leur capacité à capturer les informations syntaxiques et sémantiques des mots sous forme de vecteurs à partir de corpus non étiquetés. Néanmoins, et malgré leur impact, ces représentations (Mikolov et al., 2013b, 2017; Pennington et al., 2014; Joulin et al., 2016) n'ont pas permis de traiter certains phénomènes lexicaux telle que la polysémie. En effet, ces représentations sont statiques : chaque mot a un seul vecteur qu'importe le contexte dans lequel il est utilisé (Ethayarajh, 2019). Des travaux datant de 2018 sur les modèles de langue neuronaux (Devlin et al., 2018; Peters et al., 2018) ont ainsi introduit de nouveaux types de représentations dites *contextualisées*, c'est-à-dire des vecteurs de mots sensibles au contexte dans lequel les mots apparaissent. Ces approches ont initialement été fondées sur des réseaux de neurones récurrents [11] (Howard et Ruder, 2018) mais ont peu à peu intégré des modèles de type *Transformer* (Devlin et al., 2018). Remplacer les représentations continues par ces représentations contextualisées a permis ainsi d'obtenir des améliorations significatives sur un éventail diversifié de tâches en TAL pour la langue anglaise, allant de l'extraction de relations à la classification de documents (Peng et al., 2019; Laskar et al., 2020). Il existe plusieurs types de modèles de langues pré-entraînés que nous nous présentons succinctement dans

11. Un réseau de neurone récurrent (Hopfield, 1982; Rumelhart et al., 1986; LeCun et al., 2015) prend en entrée une séquence $[x_1, x_1, ..., x_T]$ pour produire une séquence de sortie $[y_1, y_1, ..., y_T]$ tout en maintenant un état caché ou *Hidden state* $[h_1, h_1, ..., h_T]$ qui représente sa mémoire du contenu de la séquence à chaque instant. Cet état caché est donné en entrée à la prochaine prédiction ainsi que l'entrée suivante. Cela permet au réseau de garder une forme de mémoire interne de ce qui a été vu précédemment dans la séquence. Contrairement à un réseau de neurone profond traditionnel qui utilise différents paramètres à chaque étape, le RNN partage les mêmes paramètres à toutes les étapes. Le RNN effectue ainsi la même tâche à chaque étape, mais avec des entrées différentes. Cela réduit considérablement le nombre total de paramètres qu'il doit apprendre. Par contre, il est moins bon à apprendre à relier les informations entre elles et gérer les dépendances à long terme.

la suite, mais nous nous attarderons plus particulièrement sur l'architecture BERT et ses deux variantes en langue Française FlauBert (Le et al., 2020) et CamemBert (Martin et al., 2019, 2020a).

- **CoVe** (McCann et al., 2017) ou vecteurs de mots contextualisés est un type de plongements de mots appris par un encodeur de type LSTM [12] entraîné sur une tâche de traduction anglais-allemand. Les vecteurs contextuels (CoVe) sont des vecteurs appris par-dessus d'autres vecteurs de mots originaux, qui peuvent être des vecteurs de type GloVe, Word2Vec ou FastText. Ces derniers sont ensuite utilisés pour fournir des plongements de mots contextualisés pour diverses tâches en TAL.
- **ELMo** ou *Embeddings from Language Models* (Peters et al., 2018) est un modèle de langue pré-entraîné qui a été formé sur un grand corpus anglais d'un milliard de mots, avec un fichier de vocabulaire d'environ 800 000 mots. Par la suite, des modèles ELMo ont également été formés pour d'autres langues, mais ils se sont limités à des langues disposant de nombreuses ressources, comme l'allemand et le japonais. C'est un modèle qui remédie au problème de la polysémie et introduit une composante contextuelle.
- **ULMFIT** ou *Universal Language Model Fine-tuning for text Classification* (Howard et Ruder, 2018) est une architecture et une méthode d'apprentissage par transfert. Elle repose sur une architecture de type AWD-LSTM [13]) constitué d'un plongement, de trois couches LSTM et un ensemble de couches linéaires finales. L'entraînement consiste en trois étapes : 1) **pré-entraînement d'un modèle de langage** sur un corpus issu de Wikipédia non labellisé, 2) **affinage du modèle** de langue sur une tâche de classification et 3) **usage du modèle affiné** pour initialiser un modèle de classification. Cette méthode est très avantageuse lorsque l'on dispose d'une grande quantité de données non étiquetées.
- **GPT-1** ou *Generative Pretrained Transformer 1* a été introduit par la société OpenAI 16 et présenté par Radford et al. (2018). Il s'agit d'un modèle de langage génératif construit en utilisant des données non étiquetées puis affiné avec des exemples spécifiques sur certaines tâches de TAL telles que la classification, l'analyse de sentiment, etc. Le modèle a été appris sur le corpus *BooKsCorpus* [14] Il est monodirectionel et conçu pour la génération de texte.

12. Les réseaux de neurones à mémoire court-terme et long terme ou LSTMs sont une extension des RNN capables d'apprendre des dépendances à long terme.
13. Un AWD-LSTM est un LSTM régulier sans mécanisme d'attention (voir 4.4.1
14. https://yknzhu.wixsite.com/mbweb

Avant d'introduire le modèle de langue BERT, nous allons d'abord présenter l'architecture Transformer.

4.4.1 Architecture Transformer

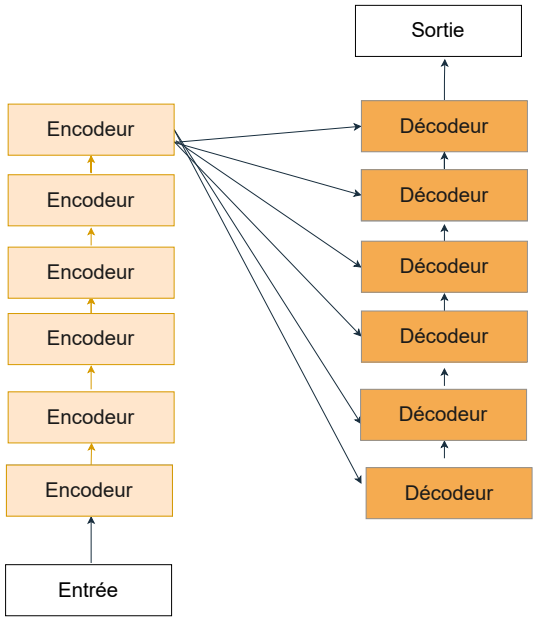


FIGURE 4.4 – Illustration simplifiée de l'architecture *Transformer* (Vaswani et al., 2017).

L'architecture Transformer (Vaswani et al., 2017) est un réseau de neurones de type *seq2seq* [15] qui a été principalement développé pour la tâche de traduction automatique. L'architecture est représentée en figure 4.4. Sa particularité réside dans le fait qu'il n'utilise que le mécanisme d'attention. Le mécanisme d'attention a été introduit par Bahdanau et al. (2014) comme un moyen de contourner le problème de la gestion des longues séquences au niveau de l'encodeur pour un modèle seq2seq. Au lieu de compresser l'information en un seul vecteur de longueur fixe, l'encodeur via le mécanisme d'attention examine d'autres positions dans la séquence d'entrée qui contiennent des informations lui permettant d'encoder le mot cible. Cela évite

15. Un modèle seq2seq (Sutskever et al., 2014; Cho et al., 2014) ou de séquence à séquence est un modèle qui prend en entrée une séquence (mots, lettres, caractéristiques d'une image, etc.) et renvoie une séquence de sortie. Ce type de modèle est généralement utilisé dans le domaine de la traduction automatique neuronale où une séquence est une série ou suite de mots, traités les uns après les autres.

au modèle d'encoder toute la phrase dans un seul vecteur de longueur fixe et lui permet de se concentrer uniquement sur les informations pertinentes pour la génération du mot cible suivant. Le *Transformer* est constitué de deux parties : un bloc de 6 encodeurs et un bloc de 6 décodeurs.

Encodeur

Le bloc d'encodeurs est construit en empilant 6 encodeurs. La sortie d'un encodeur correspond à l'entrée du suivant sauf pour le premier encodeur qui prend en entrée une séquence de plongements de mots représentant l'entrée. Chaque encodeur est constitué deux de couches intermédiaires qui sont toutes les deux des réseaux de neurones : une couche *Self-attention* ou d'auto-attention et un *Feed-Forward Neural Network* (FFNN) ou réseau de neurone de type propagation avant [16]. Les entrées de l'encodeur passent d'abord par la couche d'auto-attention, une couche qui aide l'encodeur à regarder les autres mots de la phrase d'entrée lorsqu'il encode un mot spécifique. Les sorties de la couche d'auto-attention sont ensuite transmises à un réseau neuronal de type propagation avant. Le FFNN est appliqué de manière indépendante à chaque position de la séquence d'entrée. Un autre détail dans l'architecture de l'encodeur est que chaque couche (auto-attention, FFNN) dans chaque encodeur possède une connexion résiduelle [17] autour d'elle et est suivi d'une couche de normalisation.

Décodeur

Le décodeur possède également deux couches, mais entre elles se trouve une autre couche d'attention de type *encoder-decoder attention* qui aide le décodeur à réaliser le mécanisme d'attention. L'entrée d'un décodeur comprend la sortie du précédent décodeur et la sortie du 6ème encodeur. Le dernier décodeur est généralement connecté à un réseau de neurones linéaire couplé à une fonction softmax. Ce réseau a pour objectif d'identifier les mots du vocabulaire qui correspondent à la sortie du dernier décodeur. La figure 4.5 présente l'architecture de l'encodeur et du décodeur.

16. Un réseau de neurone de type propagation avant est un type de réseau neuronal artificiel composé de plusieurs couches au sein desquelles l'information est transmise dans un seul sens de la couche d'entrée vers la couche de sortie.

17. La connexion résiduelle consiste à « sommer les représentations d'une couche d'un réseau de neurones avec les représentations d'une couche précédente (Frej, 2021)». Cela permet d'augmenter les performances du réseau.

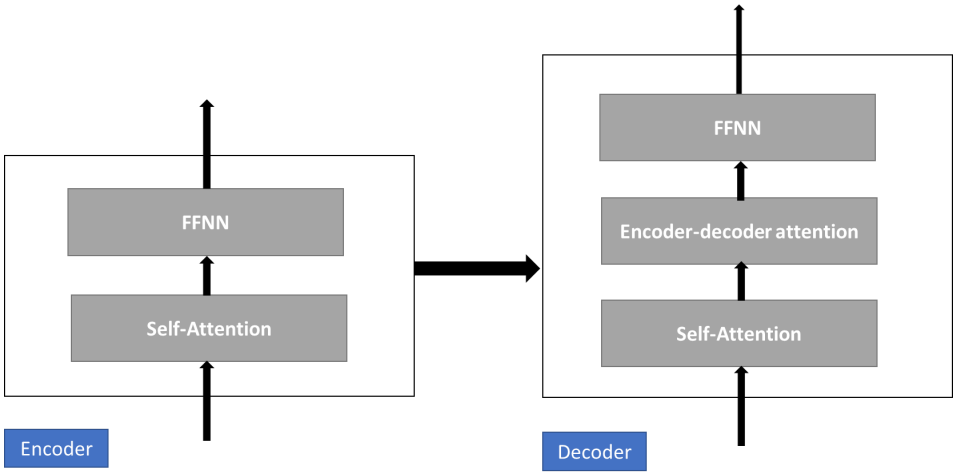


FIGURE 4.5 – Illustration simplifiée de l'encodeur et du décodeur (Vaswani et al., 2017).

4.4.2 BERT

BERT est l'acronyme de *Bidirectional Encoder Representations From Transformers* qui est un modèle de langue neuronal qui repose sur une architecture de type Transformer comprenant un mécanisme d'attention pour produire des représentations contextualisées de mots. Il utilise un encodeur qui lit la séquence d'entrée et une couche de classification pour la prédiction des mots masqués. BERT a été conçu pour apprendre des représentations bidirectionnelles profondes à partir de textes non étiquetés en prenant obligatoirement en compte les contextes de gauche et de droite conjointement, et ce, dans toutes les couches du Transformer. De ce fait, BERT peut être affiné [18] avec une seule couche de sortie pour créer des modèles pour une variété de tâches telles que la réponse aux questions, l'inférence en langage naturel et pleines d'autres sans une modification substantielle de son architecture. BERT est donc simple et empiriquement puissant. BERT a été réalisé en s'appuyant sur un corpus de 3,3 milliards de mots provenant de *BooksCorpus* (800 Millions de mots) (Zhu et al., 2015) et du Wikipédia Anglais (2.5 Milliards de mots). S'agissant de Wikipedia, seul les passages de texte ont été extraits. Les listes, tables et titres n'ont pas été pris en compte. Deux tailles de modèles ont été propo-

18. En anglais : « *Finetuning* ». Dans certains articles le terme *adapter finement* ou l'anglicisme *finetuner* sont utilisés. Dans le cadre de cette thèse, nous utiliserons les termes *affinage* ou *affiner*.

sées lors de la première publication des travaux [19] : une version $BERT_{BASE}$ et une $BERT_{LARGE}$. En plus de ces deux modèles, une variante multilingue a également été proposé : $mBERT_{BASE}$. mBERT [20] (Kenton et Toutanova, 2019) a été pré-entraîné sur des données multilingues (104 langues [21]) issues de Wikipédia.

BERT a appris à produire des représentations (plongements) contextualisées des mots du vocabulaire issus de corpus non labellisés selon deux tâches auto supervisées simultanément : *la modélisation du langage masqué* et *la prédiction de la phrase suivante*.

Tâche 1 : Modélisation du langage masqué (MLM)

Contrairement aux modèles de langues classiques (Mikolov et al., 2013b; Pennington et al., 2014; Joulin et al., 2016) qui prédisent le mot voisin à partir de son entourage, BERT le fait de manière bidirectionnelle. BERT prend en compte simultanément les contextes droit et gauche du mot cible de la séquence d'entrée. Cette caractéristique permet ainsi au modèle de capturer le contexte d'un mot par rapport à d'autres mots dans les deux sens, gauche et droite, et de produire des représentations davantage contextualisées que des modèles tels que ELMo (Peters et al., 2018) ou GPT-1. Pendant la tâche de modélisation du langage masqué, un pourcentage (15%) des tokens (obtenus avec l'algorithme WordPiece [22]) de chaque séquence d'entrée du réseau est masqué de façon aléatoire. Ces mots sont supprimés dans la séquence et remplacés par le token <MASK>. Après cette étape, le modèle essaie de prédire ces mots en regardant les mots non masqués appartenant à la séquence. Devlin et al. (2018) mentionnent dans leur article que les mots masqués ne sont pas toujours remplacés par le token<MASK>. Un générateur de données d'entraînement choisit plutôt 15% des tokens au hasard pour la phase de prédiction. En substance, si le *i*ème token est choisi, ce *i*ème token est remplacé 80% du temps par un token <MASK>, 10% du temps par un token aléatoire et 10% du temps, il reste inchangé. Cette procédure est représentée à la figure 4.6.

Tâche 2 : Prédiction de la phrase suivante (NSP)

Durant la procédure NSP, le modèle reçoit des paires de phrases en entrée et apprend

19. Toutefois, d'autres modèles beaucoup moins grands, ont été mis par la suite à disposition par les auteurs. Ces modèles sont accessibles à cette adresse : https://github.com/google-research/bert.
20. https://huggingface.co/bert-base-multilingual-cased
21. https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages
22. WordPiece est un algorithme de segmentation des mots en sous mots utilisé en TAL (Wu et al., 2016; Devlin et al., 2018). Il peut être utilisé pour différentes langues.

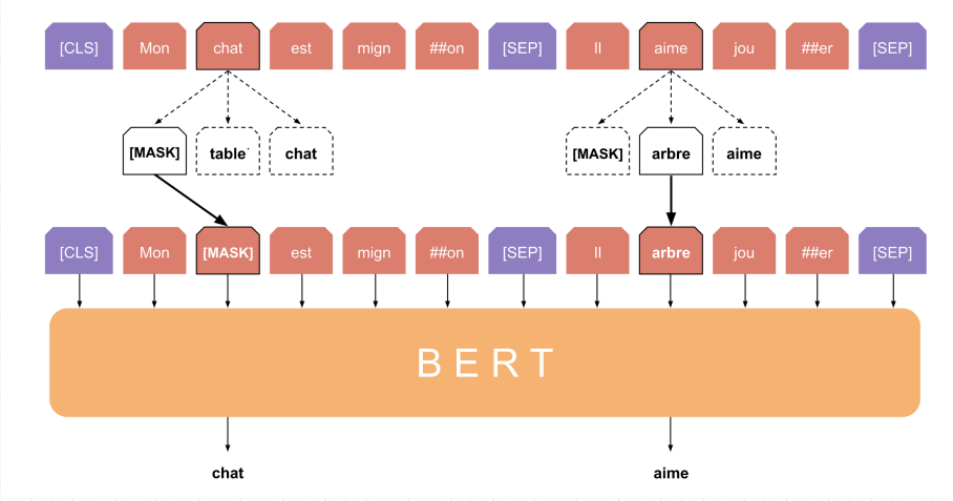


FIGURE 4.6 – Représentation de l'entrée de BERT (El Boukkouri, 2020).

à prédire si la deuxième phrase de la paire est la phrase suivante dans le document original. Pendant l'apprentissage, 50% des entrées sont une paire dans laquelle la deuxième phrase est la phrase suivante dans le document original et labellisée *IsNext*, tandis que dans les 50% restants, une phrase aléatoire du corpus est choisie comme deuxième phrase et labellisée (*NotNext*). Ceci permet à BERT de comprendre le contexte du langage au travers de plusieurs phrases. BERT a été proposé pour produire des plongements contextualisés pour la langue anglaise. Des variants de BERT [23] pour d'autres langues [24] ont été proposés pour le français comme FlauBERT (Le et al., 2020), CamemBert (Martin et al., 2019, 2020a). Pour un inventaire plus général, le lecteur peut se référer à Subramanyam Kalyan et al. (2021).

Après le pré-entraînement, BERT peut être utilisé pour d'autres tâches très spécifiques [25] en TAL. Plus spécifiquement, BERT peut être utilisé comme générateur de représentations contextualisées conjointement avec un autre modèle sur une tâche plus spécifique. Cette phase est généralement très simple, moins coûteuse en

23. BERT n'est pas le seul modèle de langue à disposer de variant. Il existe également des variants pour le modèle ELMo en portugais, japonais, allemand et basque.

24. On peut également citer AraBERT pour la langue Arabe (Antoun et al., 2020), BETO pour l'espagnol (Canete et al., 2020). Des modèles par rapport au domaine ont également été proposés comme BioBERT (Lee et al., 2020) ou ClinicalBERT (Alsentzer et al., 2019) pour le domaine médical.

25. Devlin et al. (2018) montrent que BERT obtient des résultats de pointe sur 11 tâches : la classification de texte, la réponse aux questions, l'inférence, la détection de paraphrases, l'analyse syntaxique et étiquetage morphosyntaxique, la désambiguïsation lexicale et, etc.

temps et ne nécessite pas de modification profonde de l'architecture. Seules les entrées et la sortie ont besoin d'être changées.

4.4.2.1 FlauBERT

FlauBERT a la même architecture que BERT et a été appris sur un corpus de 71GB (*après prétraitement*) issus de 24 sous-corpus de types divers (des textes monolingues des campagnes d'évaluation WMT19 (Li et al., 2019), des textes en français de la collection OPUS (Tiedemann, 2012) et du projet Wikimedia [26]. Contrairement à BERT qui a été entraîné sur deux tâches, FlauBERT a été entraîné uniquement sur une seule tâche : la modélisation du langage masqué (MLM) tout en conservant le format d'entrée consistant en une paire de phrases. Un vocabulaire de 50K unités sous-lexicales est construit en utilisant l'algorithme *Byte Pair Encoding* [27]. Avant l'application du BPE, le corpus d'entraînement est prétraité et tokenisé en utilisant le tokenizeur français Moses (Koehn et al., 2007). Deux modèles de langues ont été entraînés : $FlauBERT_{BASE}$ et $FlauBERT_{LARGE}$. Le tableau 4.5 donne une comparaison du modèle FlauBERT avec BERT (version anglaise et multilingue).

	$BERT_{BASE}$ / $mBERT_{BASE}$	$FlauBERT_{BASE}$/ $FlauBERT_{LARGE}$
Langue	Anglais/Multilingue	Français
Données d'apprentissage	13 GB / Wikipedia(104 langues)	71 GB
Objectifs de pré-entraînement	NSP et MLM	MLM
Nombre total de paramètres	110 M	138M/373M
Tokenisation	WordPiece 30K / 110 000K	BPE 50K
Masque	Statique + sous-mots	Dynamique + sous-mots

TABLE 4.5 – Comparaison des modèles BERT et FlauBERT (Devlin et al., 2018; Le et al., 2020).

26. https://meta.wikimedia.org/w/index.php?title=Data_dumps&oldid=19312805

27. *Byte Pair Encoding* est un algorithme de segmentation en sous-mots qui code les mots rares et inconnus comme des séquences d'unités de sous-mots. Au lieu d'encoder le terme *playing* ou chacune de ses lettres séparément, BPE pourra les encoder "play" d'un côté et "ing" de l'autre. Cette logique permet ainsi de construire des blocs qui pourront être utilisés par d'autres mots également (Sennrich et al., 2015b).

4.4.2.2 CamemBERT

CamemBERT est un modèle de langue pré-entraîné pour le français fondé sur RoBERTa (Liu et al., 2019). À ce jour, il existe 6 modèles [28] de CamemBERT pré-entraînés sur divers corpus. Le tableau 4.6 présente une comparaison générale entre RoBERTa, FlauBERT et CamemBERT. Le premier modèle de CamemBERT a été pré-entraîné sur la version française du corpus OSCAR (Suárez et al., 2019) qui représente 138 GB de texte après filtrage et nettoyage. Les deux modèles (BASE et LARGE) de CamemBERT ont été pré-entraînés en utilisant l'ensemble des 135 GB du corpus CCNet (Wenzek et al., 2019) qui est également extrait de Common Crawl. Les trois autres derniers ont été pré-entraînés sur les versions réduites (4 GB) d'OSCAR, CCNet et d'un autre corpus issu de Wikipédia.

	$RoBERTa_{BASE}$	$CamemBERT_{BASE}$	$FlauBERT_{BASE}$/ $FlauBERT_{LARGE}$
Langue	Anglais	Français	Français
Données d'apprentissage	160 GB	138 GB	71 GB
Objectifs de pré-entraînement	MLM	MLM	MLM
Nombre total de paramètres	125 M	110 M	138M/373M
Tokenisation	BPE 50K	SentencePiece 32K	BPE 50K
Masque	Dynamique + sous-mots	Dynamique + mot entier	Dynamique + sous-mots

TABLE 4.6 – Comparaison des modèles, RoBERTa, CamemBERT et FlauBERT (Le et al., 2020; Devlin et al., 2018; Martin et al., 2019, 2020a).

RoBERTa est un modèle développé à partir de BERT en modifiant certains hyperparamètres pendant la phrase de pré-entrainement [29]. Contrairement à RoBERTa et FlauBERT, CamemBERT utilise l'algorithme SentencePiece [30] pour la tokenisation et a recours à un modèle de masque du mot entier [31]. Tout comme RoBERTa, les tokens sont masqués dynamiquement pour tout le corpus durant la phase de prétraitement. CamemBERT est entraîné sur la tâche de modélisation du langage masqué (MLM). Le modèle de masquage du mot entier est utilisé par CamemBERT en sélectionnant de manière aléatoire 15% des mots de la séquence. Pour chaque 15%,

28. Les modèles sont accessibles à cette adresse : https://camembert-model.fr/.

29. RoBERTa a été pré-entraîné uniquement sur la tâche MLM et sur 160 GB de données. Le vocabulaire pris en compte était de 50K sous-mots (30K pour BERT). L'algorithme de segmentation utilisée était le BPE (WordPiece pour BERT).

30. SentencePiece est une extension de BPE et de WordPiece qui ne nécessite pas de pré-tokenisation au niveau du mot ou du token supprimant le besoin d'avoir recours à des tokenizers spécifiques à la langue (Zouari).

31. Une version mise à jour de BERT a montré que le masquage des mots au lieu des sous mots a amélioré la performance. Les modèles sont accessibles à l'adresse suivante : https://github.com/google-research/bert/blob/master/README.md.

tous les sous-mots de la séquence sont considérés comme des candidats au remplacement. Dans ces 15%, 80% sont remplacés par un token <MASK>, 10% par un token aléatoire et les 10% restants restent inchangés. Le modèle est ensuite entraîné pour prédire les mots masqués. La tâche de NSP, tout comme, pour FlauBERT n'est pas utilisée.

4.4.3 Synthèse

BERT a fait progresser l'état de l'art sur plusieurs tâches en TAL, notamment sur la tâche de classification des textes portant sur le **Stanford Sentiment Treebank** (Socher et al., 2013) ou la tâche de question-réponses sur le **Stanford Question Answering Dataset** (Rajpurkar et al., 2016). De même, les bons résultats obtenus par ses variants en langue française (CamemBERT et FlauBERT [32]) pour la tâche de classification des textes sur le référentiel d'évaluation FLUE [33] ont montré l'importance de développer des modèles monolingues. Étant donné les bonnes performances qu'ont obtenues FlauBERT et CamemBERT sur la classification de textes, nous pensons qu'ils seraient également très performants sur la classification de verbatim en FMC par rapport à des méthodes fondées sur des règles. Nous présentons les résultats obtenus par les modèles monolingues en langue française sur différentes tâches de Flue au tableau 4.7.

Tâche	Classification			Paraphrase	NLI	Constituants		Dépendances		Désambiguïsation	
Section	Livres	DVD	Musique							Noms	Verbes
Mesure	Acc.	Acc.	Acc.	Acc.	Acc.	F_1	POS	UAS	LAS	F_1	F_1
État de l'art ant.	91.25[c]	89.55[c]	93.40[c]	66.2[d]	80.1/**85.2**[e]	87.4[a]		89.19[b]	85.86[b]	-	43.0[h]
Sans pré-entr.	-	-	-			83.9	97.5	88.92	85.11	50.0	-
FastText	-	-	-			83.6	97.7	86.32	82.04	49.4	34.9
mBERT	86.15[c]	86.9[c]	86.65[c]	89.3[d]	76.9[f]	87.5	98.1	89.5	85.86	56.5	44.9
CamemBERT	**93.40**	**92.70**	94.15	89.8	81.2	88.4	**98.2**	91.37	88.13	56.1	**51.1**
FlauBERT$_{BASE}$	**93.40**	92.50	**94.30**	**89.9**	81.3	**89.1**	98.1	**91.56**	**88.35**	54.9/**57.9**[g]	47.4

FIGURE 4.7 – Résultats finaux sur les tâches de FLUE (Le et al., 2020).

32. CamemBERT et FlauBERT surpassent le modèle multilingue de BERT *mBERT* avec respectivement 94.30 et 94.15 par rapport à 86.65 pour la classification sur un corpus de commentaires.

33. FLUE est la version française de GLUE qui est une collection de ressources pour l'apprentissage, l'évaluation et l'analyse des systèmes de compréhension du langage naturel.

4.5 Conclusion

Dans ce chapitre, nous nous sommes intéressés aux différentes représentations utilisées dans le TAL pour traiter les données textuelles. Nous nous sommes particulièrement intéressés aux modèles de langues basés sur les architectures Transformer notamment le modèle BERT et ces deux variantes en français FlauBERT et CamemBERT que nous avons utilisés dans cette thèse. Dans le prochain chapitre, nous présentons le matériel textuel utilisé pour la tâche de classification.

Deuxième partie

Contributions

Chapitre 5

Données de travail et protocole d'annotation

Durant les précédents chapitres, nous avons présenté les différents domaines sur lesquels s'appuie ce travail de recherche. Pour ce chapitre, nous proposons d'exposer la démarche méthodologique que nous avons appliquée pour constituer un corpus d'apprentissage. La section 5.1 introduit le chapitre en présentant de manière générale les corpus et la procédure de recueil suivie. Le matériel linguistique utilisé afin de mener ce travail de recherche est ensuite détaillé dans les sections 5.2 à 5.5. Dans ces différentes sections, nous présentons successivement les données utilisées, le processus de collecte mis en œuvre, le protocole d'annotation, les caractéristiques linguistiques du corpus et les expériences préliminaires menées sur le premier corpus de verbatim annoté.

5.1 Recueil des données d'apprentissage

Notre travail de recherche a mobilisé trois types de corpus qui ont été constitués à différentes étapes.

- un premier corpus constitué de **verbatim** issu de **transcriptions** et de **prises de notes** issues d'**entretiens semi-directifs** et de **focus groups** ;
- un deuxième corpus constitué de **commentaires** sur divers produits tels que des livres, des films et de la musique provenant du site Amazon ;
- un dernier corpus constitué de **commentaires** ou **posts** sur des **produits innovants** provenant d'études menées sur la plateforme Yoomaneo [1].

1. https://www.yoomaneo.com/fr

En apprentissage automatique, quelle que soit la tâche envisagée, le corpus constitue une ressource cruciale, essentielle et centrale pour l'entraînement de modèles. En TALN, plusieurs corpus d'apprentissage et de test sont généralement mis à la disposition de la communauté scientifique pour diverses tâches (détection d'opinions, résumé automatique, etc.). Nous pouvons citer les corpus disponibles via la librairie *dataset* [2] de la plateforme *Huggingface* [3] ou encore les corpus accessibles depuis le site zenodo [4]. Néanmoins, il n'existe pas encore de corpus annoté en FMC disponible. Le matériau linguistique historique utilisé dans les études d'usages chez Ixiade étaient essentiellement de l'oral transcrit, c'est-à-dire des transcriptions d'entretiens semi-directifs et de *focus groups*. Ces données étaient compilées dans des archives de la société et n'avaient jamais jusqu'à l'initiation de ces travaux de recherche fait l'objet d'un assemblage. De ce fait, une procédure de constitution d'un corpus d'apprentissage a été mise en œuvre. Par ailleurs, nous avons également précisé au chapitre introductif qu'un autre mode de collecte de données était également amené à être utilisé. Il s'agit de la collecte réalisée par le biais de la plateforme communautaire *Yoomaneo*. Dans ce cas particulier, nous ne sommes plus sur de l'oral transcrit, mais sur du contenu saisi en ligne dans le cadre d'études. Dans les prochaines sections de ce chapitre, nous détaillons les méthodologies employées pour constituer des corpus d'apprentissage pour ces deux types de données.

5.2 Corpus de transcriptions et de prises de notes

Le premier corpus sur lequel nous nous sommes penchés est le corpus de transcriptions et de prises de notes issues de précédentes études. De manière spécifique, l'objectif était de collecter un corpus de verbatim puis de le soumettre à l'évaluation d'experts (*des chargés d'études*) afin d'obtenir un corpus d'apprentissage suffisamment important pour construire des modèles de classification pour notre tâche. Le processus de collecte dont un schéma est présenté en figure 5.1 a nécessité une **phase de compilation et d'identification** [5] d'anciennes études pertinentes pour la tâche, une **phase de regroupement des textes** et une **phase d'annotation**. Ces trois phases sont détaillées dans les sections suivantes.

2. https://huggingface.co/docs/datasets/
3. https://huggingface.co/
4. https://zenodo.org/record/4498086#.Yg9wC-jMLIU
5. Cette phase a été réalisée par une doctorante d'Ixiade qui avait préalablement commencé à travailler sur un processus de collecte de verbatim pour ses propres travaux de recherche.

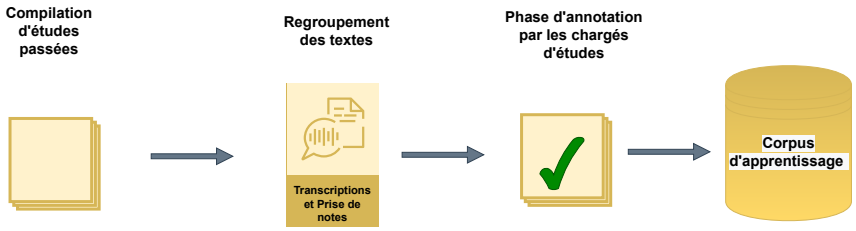


FIGURE 5.1 – Procédure générale d'agrégation du corpus d'apprentissage.

5.2.1 Compilation et identification d'études passées

Pour cette sous-section, nous détaillons la méthode appliquée pour sélectionner les études à partir desquelles nous avons extrait le matériau nécessaire pour constituer le corpus d'apprentissage de notre travail. Le corpus que nous souhaitons constituer sera principalement composé d'unités phrastiques appelés verbatim (voir 2.5.2). Globalement, le verbatim renvoie aux propos tenus par les individus principalement lors d'une étude qualitative, mais il peut également provenir des questions ouvertes lors d'une étude quantitative.

Les verbatims récupérés pour former ce premier corpus d'apprentissage proviennent d'anciens projets d'études qui ont été réalisés par la société Ixiade entre 2005 et 2018. Les informations générales comme détaillées de chacun de ces projets d'études avaient été préalablement listées par une doctorante de l'équipe travaillant sur la méthode Cautic®. Les informations générales intégraient tous les projets réalisés par Ixiade (environ 482 projets), l'année et le statut du projet (*archivé*, *en cours*, *accessibilité sur le serveur*). Les informations plus détaillées portaient sur une portion d'environ 33 projets dont 17 (voir le tableau 5.1) avaient été sélectionnés pour être exploitables [6]. Celles-ci incluaient des informations telles que l'année de réalisation du projet d'étude, le nom de l'analyste ayant réalisé l'étude, le nom du client, la langue de l'étude, l'exploitabilité de l'étude, le domaine (voir le tableau 5.2) de l'étude [7], le fichier d'analyse [8], le nombre d'entretiens réalisés, le nombre d'enre-

6. Pour être sélectionné comme exploitable, le dossier du projet d'études devait contenir le fichier d'analyse ou de *mise à plat* des transcriptions ou prises de notes de l'ensemble des entretiens ou *focus groups* réalisés pour l'étude en question. En effet, c'est dans ce fichier qu'est contenu l'ensemble des verbatims codés en FMC pour l'étude.

7. Les études ont été regroupées en deux catégories : la catégorie *Électricité* ou *Non Électricité*. Les études appartenant à la catégorie *Électricité* étaient du domaine de l'électricité. Les études appartenant à la catégorie *Non Électricité* concernaient des domaines tels que la médecine, la pâtisserie, l'électroménager, etc.

8. L'analyse ou encore la *mise à plat* est une étape d'un projet d'études au cours duquel le chargé d'études extrait manuellement dans les transcriptions les verbatims qu'il juge pertinents. Ces

gistrements présents, le nombre de transcriptions [9], de *focus groups*, le nombre de mots, le nombre de personnes interrogées [10], le nombre de prises de notes [11], le type de verbatim (*discours* [12] et *illustratif* [13]), le nombre de verbatims et les commentaires [14]. 17 études sur les 33 présélectionnées ont été jugées exploitables et utilisées pour recueillir les verbatims pour le corpus d'apprentissage.

Langue	Français
Études exploitables	17
Entretiens semi-directif	Nombre d'entretiens : 112 Nombre d'enregistrements : 90 Nombre de transcriptions : 55 Nombre de mots total : 303 134
focus groups	Nombre de *focus groups* : 29 Nombre d'enregistrements : 12 Nombre de prises de notes : 21 Nombre de mots total : 130 947

TABLE 5.1 – Synthèse récapitulative des études exploitables (Ixiade).

Domaines	Total
Electricité	8
Electroménager	1
Médécine	4
Automatisme	1
Pâtisserie	2
Gérontologie	1
Total	**17**

TABLE 5.2 – Répartition des études exploitables par domaine.

verbatims vont être triés en fonction des catégories auxquels ils appartiennent (Cautic ou FMC).

9. Certains enregistrements peuvent ne pas avoir été transcrits.
10. Si mentionné dans le fichier d'analyse.
11. La prise de note est une transcription écrite et résumée des tables rondes.
12. Type de verbatim présent dans le fichier d'analyse. Un verbatim présent dans le fichier d'analyse peut ne pas figurer dans le livrable final. Pour rappel, certains verbatims du fait de leur longueur peuvent être scindés. De ce fait, juste une portion est incluse dans le livrable.
13. C'est le verbatim qui est choisi pour figurer dans le livrable. Il peut être une portion d'un verbatim ou le verbatim tout entier.
14. Le commentaire peut contenir des précisions sur le lieu des études. (Par exemple, le nom du pays ou la ville)

Étant donné qu'il s'agit de conversations issues d'entretiens semi-directifs ou de tables rondes, la durée moyenne des dialogues est généralement très longue. En effet, un entretien dure en moyenne trente minutes pour les plus courts à plus d'une heure et demie pour les plus longs. S'agissant des tables rondes, la durée est d'environ deux heures à trois heures et demie.

5.2.2 Phase de regroupement des verbatims

L'objectif de l'étape précédente était de permettre d'identifier les études potentiellement utilisables pour la constitution du corpus d'apprentissage. L'objectif de cette phase est de collecter dans les fichiers d'analyse et les livrables clients, les verbatims déjà codés au cours du projet par les différents chargés d'études en critères CAUTIC® et en FMC.

La construction du corpus s'est ainsi faite à partir de l'extraction manuelle des verbatims (*déjà codés en critères CAUTIC® et FMC*) dans les différents fichiers d'analyse et de livrables disponibles dans chaque projet exploitable vers un fichier final de compilation. Ainsi, nous disposions de 8 projets du domaine de l'électricité et 9 du domaine général comme présenté au tableau 5.2. Il est aussi important de souligner que l'analyse des verbatims se présentait dans le fichier d'analyse en deux niveaux : CAUTIC® et ensuite FMC. Cependant, tous les verbatims codés en CAUTIC® n'étaient pas codés en FMC. C'est ainsi que nous nous sommes retrouvés avec une grande portion de verbatim codée en CAUTIC® mais pas nécessairement en FMC réduisant ainsi la taille des verbatims à notre disposition. Ainsi, 4367 verbatims [15] ont été collectés et seulement 1944 ont été codés en FMC. La répartition des verbatims collectés est détaillée à la sous-section 5.2.2.1.

Au niveau du **fichier d'analyse**, il se compose :

- d'un tableau Excel comprenant diverses informations relatives au numéro de l'entretien, la catégorie CAUTIC®, la catégorie FMC [16], le verbatim illustratif, et la synthèse de l'analyse.

Plus spécifiquement, les verbatims dont la catégorie (CAUTIC® ou FMC) est clairement mentionnée, sont manuellement extraits des fichiers de mise à plat et li-

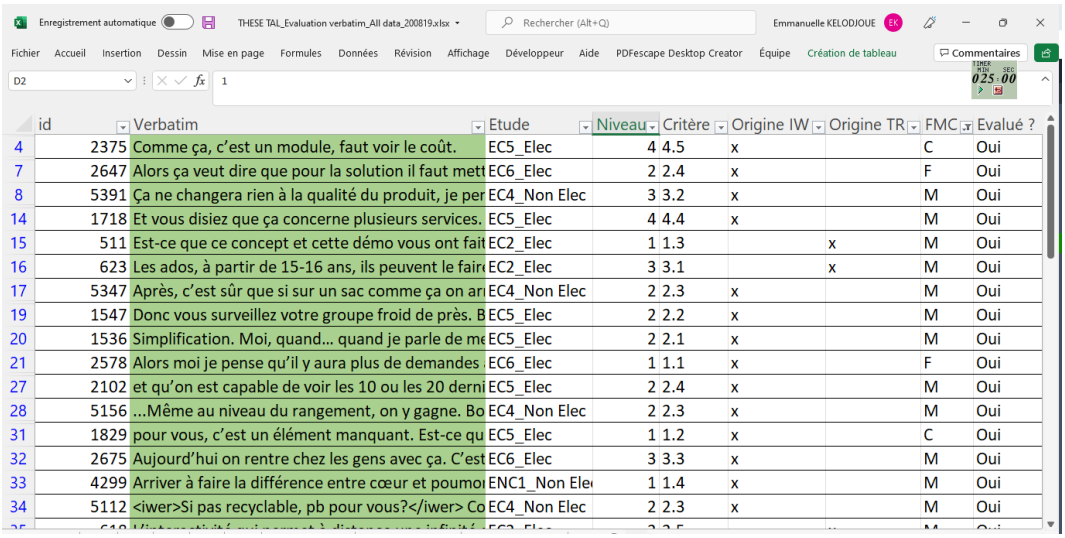
15. Au départ, nous avions 5 433 verbatims. Une phase de filtrage pour supprimer les doublons a été mise en place. En effet, un verbatim pouvait être présent comme illustratif et discours, dans ce cas, nous supprimons l'un d'eux.

16. Cette colonne n'était pas systématiquement présente dans le fichier d'analyse. Parfois un code couleur était juste appliqué sur la cellule du verbatim pour différencier les verbatims de type de freins (*rouge*), motivations (*vert*) ou conditions (*orange*).

vrables clients vers notre fichier général.

Le **fichier général** de recueil du corpus dont un extrait est présenté à la figure 5.2 se compose :

- d'un tableau Excel comprenant :
  - **l'identifiant** du verbatim ;
  - **le verbatim collecté** ;
  - **le nom de l'étude de laquelle le verbatim est extrait** ;
  - **la catégorie CAUTIC®** [17] ;
  - **la catégorie FMC** ;
  - **l'origine du verbatim** : est-ce qu'il provient d'une transcription d'un entretien ou d'une prise note de table ronde ;
  - **des informations annexes** relatives à la date à laquelle le verbatim a été évalué par l'expert et le nom du groupe d'évaluation.



	id	Verbatim	Etude	Niveau	Critère	Origine IW	Origine TR	FMC	Evalué ?
4	2375	Comme ça, c'est un module, faut voir le coût.	EC5_Elec	4	4.5	x		C	Oui
7	2647	Alors ça veut dire que pour la solution il faut mett	EC6_Elec	2	2.4	x		F	Oui
8	5391	Ça ne changera rien à la qualité du produit, je per	EC4_Non Elec	3	3.2	x		M	Oui
14	1718	Et vous disiez que ça concerne plusieurs services.	EC5_Elec	4	4.4	x		M	Oui
15	511	Est-ce que ce concept et cette démo vous ont fait	EC2_Elec	1	1.3		x	M	Oui
16	623	Les ados, à partir de 15-16 ans, ils peuvent le fair	EC2_Elec	3	3.1		x	M	Oui
17	5347	Après, c'est sûr que si sur un sac comme ça on ar	EC4_Non Elec	2	2.3	x		M	Oui
19	1547	Donc vous surveillez votre groupe froid de près. B	EC5_Elec	2	2.2	x		M	Oui
20	1536	Simplification. Moi, quand... quand je parle de me	EC5_Elec	2	2.1	x		M	Oui
21	2578	Alors moi je pense qu'il y aura plus de demandes	EC6_Elec	1	1.1	x		F	Oui
27	2102	et qu'on est capable de voir les 10 ou les 20 derni	EC5_Elec	2	2.4	x		M	Oui
28	5156	...Même au niveau du rangement, on y gagne. Bo	EC4_Non Elec	2	2.3	x		M	Oui
31	1829	pour vous, c'est un élément manquant. Est-ce qu	EC5_Elec	1	1.2	x		C	Oui
32	2675	Aujourd'hui on rentre chez les gens avec ça. C'est	EC6_Elec	3	3.3	x		M	Oui
33	4299	Arriver à faire la différence entre cœur et poumo	ENC1_Non Ele	1	1.4	x		M	Oui
34	5112	<iwer>Si pas recyclable, pb pour vous?</iwer> Co	EC4_Non Elec	2	2.3	x		M	Oui

FIGURE 5.2 – Image du fichier de collecte.

17. Voir section 2.5.3

5.2.2.1 Description du corpus

À la suite de cette collecte, **4367** verbatims ont été recueillis. Sur ces **4367** verbatims, **2423** verbatims ont été uniquement codés en CAUTIC® et **1944** verbatims en CAUTIC® et en FMC. Sur les **1944** verbatims, 408 sont des freins, 978 motivations et 558 des conditions.

5.2.2.2 Nature des données

Les entretiens semi-directifs ou *focus groups* réalisés durant les études sont systématiquement retranscrits. Les retranscriptions textuelles de ces entretiens comportent différentes caractéristiques. On y retrouve des éléments typiques de l'oral conversationnel comme les indicateurs de disfluences (par exemple, euh, ah, quoi). Bien que présents à l'oral, ces disfluences ne sont pas toutes traduites afin de ne pas nuire à la compréhension globale du texte [18]. Ainsi, une répétition de *euh... je... euh... euh...* pourra simplement être restituée par un seul *euh...* ou pas. En effet, l'objectif de la transcription est de restituer les informations utiles à la compréhension pour une analyse qualitative. De même, quelques corrections sont permises ou nécessaires afin d'éviter les contresens ou de surcharger inutilement la transcription : *Je sais pas* pourra être remplacé par *Je ne sais pas*, les *Oui*, *Ok*, *Bien*, *D'accord* de l'intervieweur ne sont pas transcrits lorsqu'ils sont utilisés pour encourager l'interviewé à s'exprimer davantage. De plus, les transcriptions sont également ponctuées, ce qui renforce l'interprétation faite de la personne en charge de la transcription. Au final, les textes obtenus sont ainsi bien écrits et ne présentent pas de faute d'orthographe. D'autres caractéristiques incluent :

— **la mention du *time code* ou code temps** sur certains passages pour indiquer qu'à cette partie l'audio était inaudible comme c'est le cas avec l'exemple suivant : *La situation est réaliste **[?00 :25 :00-00 :26 :00 ?]**. On la rencontre tous les jours.* Sur les livrables de présentation, les time code sont systématiquement enlevés.

— **la mention des expressions de rires, soupirs, et etc** dans certains passages pour signifier que l'interlocuteur riait, soupirait au moment où il s'exprimait.

Le tableau 5.3 présente une répartition en termes de mots et de phrases du corpus collecté.

18. Cette façon de faire est caractéristique de la méthodologie appliquée chez Ixiade.

	Jeu de données de transcription et de prises de notes 4367 verbatim
Nombre de phrases total	13 933
Nombre de tokens (mots) total	105 266
Nombre moyen de phases par verbatim	3,19
Nombre de verbatim ayant une phrase	1553
Nombre de verbatim ayant 2 phrases	938
Nombre de verbatim ayant 3 phrases	605
Nombre de verbatim ayant 4 phrases	401
Nombre de verbatim ayant plus de 5 phrases	870

TABLE 5.3 – Répartition du jeu de données agrégé en termes de mots et phrases.

5.2.3 Méthodologie d'évaluation du corpus

Le corpus de verbatim collecté et issu d'études passées présenté à la section précédente a été donné à évaluer en critères CAUTIC® (4 niveaux et 20 critères) et FMC (3 classes) par vague de 200 verbatims aux chargés d'études (6 annotateurs + 1 un autre annotateur en renfort appelé *annotateur commun* [19]). Tout d'abord, une première phase d'expérimentation où toute l'équipe a évalué les 100 premiers verbatims du corpus collecté a été réalisée en juillet 2019. Cette phase avait pour objectif de donner des instructions sur la procédure d'annotation. Ainsi, pour chaque verbatim, les chargés d'études devaient assigner le niveau, le critère CAUTIC® et la classe FMC. Ensuite, à partir de septembre 2019, deux groupes d'annotateurs ont été mis en place. Chaque groupe était constitué de trois annotateurs devant évaluer chacun 200 verbatims. En outre, un quatrième annotateur que nous appellerons *annotateur commun* a été sollicité, et ce, pour les deux groupes pour pallier l'indisponibilité d'un ou deux annotateurs en raison de leurs charges de travail. Plus spécifiquement, si parmi les trois annotateurs d'un groupe uniquement deux évaluent la vague de verbatim qui leur a été soumis, l'*annotateur commun* vient en renfort pour équilibrer les évaluations. En outre, les verbatims (200) étaient envoyés mensuellement afin de ne pas surcharger les chargés d'études dans leur fonction. Comme le montre la figure 5.3, les différentes situations d'évaluations suivantes ont été observées. Elles concernent uniquement les 6 annotateurs et non l'annotateur commun.

— **Première situation** : Tous les annotateurs évaluent un verbatim.

— **Deuxième situation** : Au moins un annotateur n'a pas évalué un verbatim quel que soit le groupe.

— **Troisième situation** : Au moins deux annotateurs n'ont pas évalué un verbatim quel que soit le groupe.

19. L'annotateur commun évaluait les verbatims uniquement en FMC et non en Cautic.

— **Dernière situation** : Aucun verbatim n'est évalué par le groupe d'annotateur.

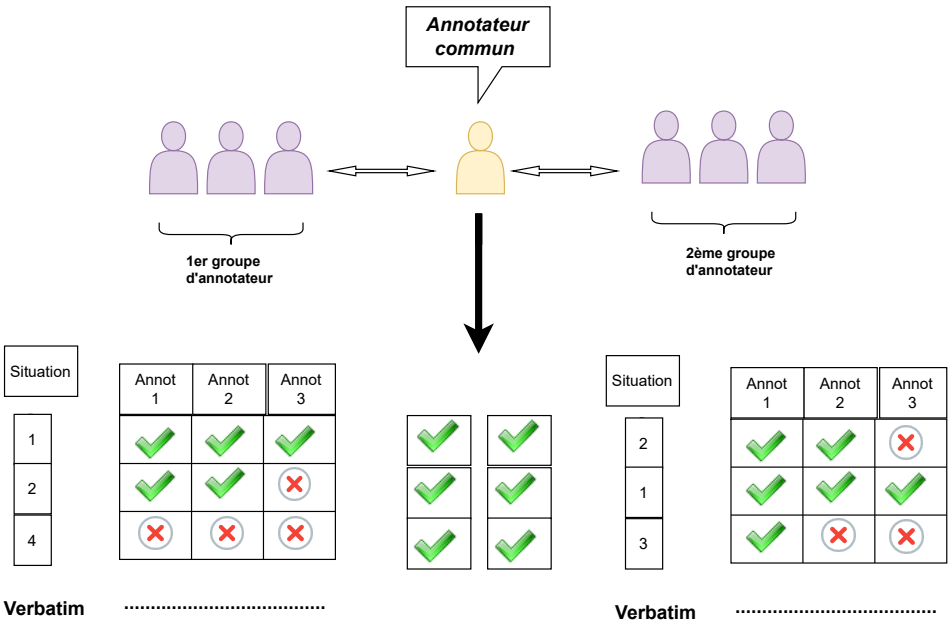


FIGURE 5.3 – Aperçu du processus d'annotation selon les différentes situations observées.

5.2.4 Accord sur l'évaluation

L'annotation de données est un sujet qui implique toujours une part de subjectivité (Benamara et al., 2007). Cette subjectivité peut différer selon les annotateurs ou annotatrices. Le calcul de l'accord inter-annotateur ou interjuge permet ainsi d'estimer la fiabilité des données. Il existe dans la littérature plusieurs méthodes pour l'évaluation de l'accord interannotateur. Les plus répandues dans le domaine du TALN sont le *kappa* de Cohen Cohen (1960) et l' *alpha* de Krippendroff (Krippendorff, 2009). Étant donné le contexte de notre travail et la non-disponibilité d'annotateurs pendant des semaines, voire sur de longues périodes, afin de constituer notre corpus final d'apprentissage de verbatim, nous avons retenu les verbatims dont la classification obtenait un accord interjuge selon les modalités suivantes :

1. Chaque verbatim de notre corpus collecté (4367) doit être évalué par au minimum trois personnes.

— Si une classe (frein, motivation, ou condition) donne lieu à un accord supérieur ou égal à 50% [20] [21] et qu'il n'y a pas 50/50 [22] sur deux classes, les actions suivantes sont menées :

  - si la classe retenue correspond à celle initialement choisie [23], on conserve cette classe ;
  - si la classe retenue est différente de celle initialement choisie, on réassigne la nouvelle classe au verbatim ;
  - si la classe retenue est inclassifiable [24], le verbatim est éliminé du corpus.

— Si l'accord interjuge est inférieur à 50% ou s'il y a 50/50 sur deux classes, le verbatim est éliminé du corpus.

Ce processus d'annotation a été effectué de juillet 2019 à juillet 2020. Le graphique 5.4 illustre la situation d'annotation pour la catégorisation FMC à l'été 2020 :

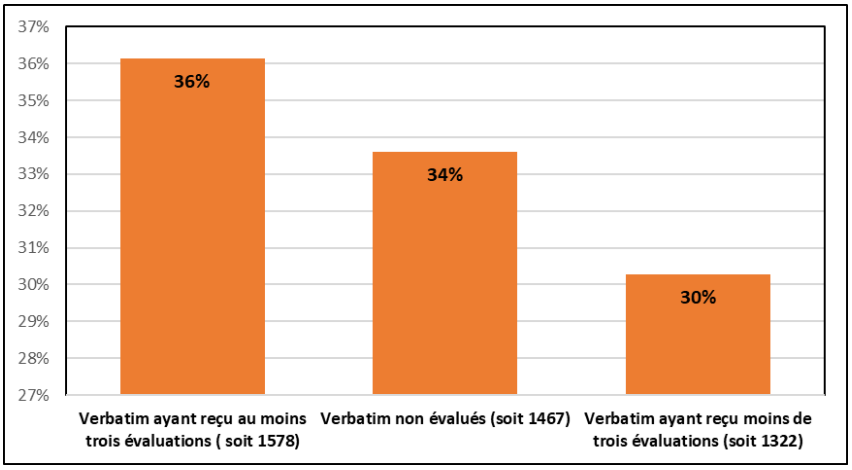


FIGURE 5.4 – Répartition de l'évaluation au niveau du corpus.

20. Pour rappel, une première vague de verbatim de 100 a été soumis en premier lieu à toute l'équipe des annotateurs sauf l'annotateur commun.

21. Si minimum trois personnes attribuent la même classe à un verbatim, ce dernier est retenu.

22. Un verbatim reçoit deux évaluations (par exemple frein et motivation) et pour chaque classe, on observe le même nombre d'annotateurs, soit 2/2 ou 3/3.

23. Il s'agit de la classe FMC d'origine ou de référence du verbatim.

24. Si une majorité d'annotateurs n'arrive pas à classifier le verbatim dans l'une de nos trois classes.

1. 36% sur l'ensemble des verbatims soit **1578** verbatims a reçu au moins trois évaluations. Sur les **1578**, seuls **839** ont obtenu un accord interjuge supérieur ou égal 50%. Parmi les **839** verbatims conservés, **274** ont conservé leur classe d'origine et **565** ont obtenu une **nouvelle classe** différente de la classe d'origine (tableau 5.4).
2. 30% sur l'ensemble des verbatims (**4367**) soit **1322** a reçu moins de trois évaluations (entre 1 et 2).
3. 34% sur l'ensemble des verbatims soit **1467** n'a reçu aucune évaluation.

Malheureusement, ce processus d'évaluation du corpus n'a pas pu être finalisé et nous avons ainsi fait le choix de travailler uniquement avec les 839 verbatims dont l'accord était supérieur à 50% sur nos classes respectives. Dans la suite, nous l'appellerons *corpusIx*. Le corpusIx obtenu après évaluation par les différents annotateurs est non seulement très faible, mais également déséquilibré au niveau des classes. Une large majorité appartiennent à la classe « motivation » suivi de la « classe condition » et ensuite « frein ». Les tableaux 5.4 et 5.5 présentent la répartition des verbatims retenus en fonction de leur classe, mais également les caractéristiques sur le nombre de phrases et de mots.

	Conservés	**Réassignés**	**Total**
Freins	53	136	189
Motivations	161	246	407
Conditions	60	183	243
Total	274	565	**839**

TABLE 5.4 – Distribution des classes en FMC sur le corpus en fonction des valeurs *conservés* ou *réassignés*. Conserver signifie que le verbatim a conservé sa classe de référence. Réassigner signifie que le verbatim a été réassigné à une nouvelle classe différente de sa classe de référence. La classe de référence est la classe d'origine, celle assignée au cours de l'étude.

Le tableau 5.4 montre la difficulté de la tâche d'annotation en FMC. En effet, sur les 407 motivations finales, seuls 161 ont conservé leur classe de référence et 246 avaient préalablement été étiquetées soit en frein ou en condition. Ce constat est également observé sur les classes frein et condition où la proportion de verbatim réassignée est supérieure à celle ayant conservé sa classe d'origine. Ainsi, sur les 839 verbatims obtenus plus de la moitié ont été assignés à une nouvelle classe différente de celle qu'ils avaient à l'origine. Par ailleurs, sur les 839 verbatims retenus, 274 (33%) ont conservé leur classe d'origine tandis que 565 ont obtenu une

nouvelle classe. Ce qui montre la complexité de catégorisation en FMC pour les chargés d'études.

	Jeu de données de transcription corpusIx - 839 verbatim
Nombre de phrases total	2648
Nombre de tokens (mots) total	47 026
Nombre moyen de phrases par verbatim	3,15
Nombre de verbatim ayant une phrase	268
Nombre de verbatim ayant 2 phrases	193
Nombre de verbatim ayant 3 phrases	123
Nombre de verbatim ayant 4 phrases	79
Nombre de verbatim ayant plus de 5 phrases (le verbatim le plus long en contient 19)	176

TABLE 5.5 – Répartition du jeu de données d'apprentissage en terme de phrases et de mots.

Nous nous sommes également rendus compte que les chargés d'études au cours de leur analyse des textes transcrits pouvaient faire la démarche de tronquer les textes en récupérant les parties ou portions d'une phrase qu'ils estimaient appartenir à telle ou telle catégorie. Ainsi, en comparant certains fichiers d'analyse avec leurs transcriptions initiales, nous nous sommes rendus compte que certains verbatims étaient des portions d'autres verbatims extraits ou phrases présentes dans les textes initiaux. Ceci peut s'expliquer par le fait que les transcriptions étaient d'abord catégorisées en CAUTIC® avant d'être classées en FMC. Nous prenons l'exemple suivant pour illustrer notre propos : *Ce concept semble intéressant, facile à comprendre et je crois qu'il aurait un réel intérêt pour les sportifs amateurs.* La première portion *Ce concept semble intéressant, facile à comprendre* a été assignée au niveau 1 et sous-critère 1.1 de CAUTIC® et l'autre portion (*je crois qu'il aurait un réel intérêt pour les sportifs amateurs*) au niveau 3 et sous-critère 3.1. Chaque portion a été assignée à la classe motivation. Cette phase de collecte et d'évaluation a été longue et fastidieuse et n'a pas permis de récolter assez données pour notre tâche. Parallèlement donc à cette procédure, nous avons décidé de collecter un deuxième corpus issu d'un autre type de données que nous présentons au chapitre 5.3.

5.3 Corpus Amazon

Le recours à un autre corpus de données poursuivait deux objectifs :

— pallier le manque de données à notre disposition en constituant un nouveau corpus complémentaire au corpus de transcription (voir 5.2) en cours d'évaluation à l'époque.

— développer une méthode d'extraction des freins, motivations et conditions en vue d'extraire dans d'autres corpus des données similaires aux données traitées chez Ixiade.

5.3.1 Recueil des données

Nous souhaitons compléter le corpus de transcription [25] que nous utiliserons pour la tâche de classification en freins, motivations et conditions. Notre choix s'est porté sur le corpus Amazon [26], car il est constitué essentiellement de commentaires dans lesquels les internautes expriment leurs opinions sur des objets précis même si ces objets ne sont pas des innovations ou des concepts d'innovation et qu'ils sont déjà accessibles [27]. Le corpus Amazon est composé de commentaires/critiques de produits de la part d'internautes en langue française [28](11998 commentaires). Chaque critique appartient à une des trois catégories suivantes : livres, films ou musique, contient un titre, le texte de la critique et une note de 1 à 5 étoiles. La répartition est présentée au tableau 5.6.

	Jeu de données Amazon
Nombre total de phrases	61 146 phrases
Nombre total de tokens (mots)	1 339 306 mots
Nombre moyen de phrases par commentaire	5,09

TABLE 5.6 – Répartition du jeu de données en termes de mots et phrases.

25. À ce moment-là, seuls 500 verbatims des 4367 avaient déjà reçu au moins 3 évaluations dont 260 avaient été retenus pour faire partir de notre corpus final.

26. https://zenodo.org/record/3251672

27. Les études chez Ixiade portent en grande majorité sur des innovations réelles qui sont soit à l'état d'idée ou de concept, soit des maquettes 3d mais pas encore accessibles sur le marché. Les utilisateurs dans le cadre d'une étude donnent leur opinion dans une projection de l'utilisation et des désagréments qu'ils pourraient rencontrer. Or, s'agissant des commentaires en ligne, les utilisateurs ont été en contact direct avec l'objet dans leur espace de vie.

28. Ce corpus existe aussi pour la langue anglaise, japonaise et allemande.

La constitution de ce deuxième corpus a nécessité d'extraire dans un premier temps un ensemble de données, plus spécifiquement des phrases. Dans cet ensemble de phrases, nous avons décidé d'extraire des phrases de types frein. Pour cela nous avons défini deux critères d'extraction :

— **Critère 1 : Présence du connecteur « mais » et d'expressions négatives**. Nous partons de l'hypothèse qu'un énoncé (phrase ou paragraphe) peut être considéré comme un frein si le marqueur "mais" y est retrouvé. Par exemple, *Je n'ai rien contre Eve Angeli, mais je trouve qu'il manque une certaine profondeur et de l'émotion dans sa façon de chanter* ou *"J'ai toujours écouté les albums """"Star Academy"""", mais celui-là est le plus mauvais..."""*).

— Critère 2 : **Présence des mots négatifs**.

Nous avons utilisé le lexique de sentiment Feel (Abdaoui et al., 2017) pour cette phase d'extraction via des expressions régulières et la librairie Spacy [29].

Le tableau 5.7 présente la répartition du nombre de phrases extraites :

	Critère 1	**Critère 2**
Total des phrases sur le corpus Amazon	**6737 (11%)**	**28079 (45%)**

TABLE 5.7 – Résultats obtenus après extraction.

L'intersection des phrases recueillies en fonction de ces deux critères (6399 phrases) a été soumise à l'annotation des chargés d'études.

5.3.2 Résultats de l'annotation

	Nombre de phrases	Pourcentage sur les 6399 phrases
Motivations	**1059**	16%
Freins	**2020**	32%
Conditions	**381**	6%
Inclassifiable	1270	20%
Non annoté	1669	26%

TABLE 5.8 – Répartition des annotations.

Les 6399 phrases ont été soumises pour évaluation à nos 7 annotateurs. Nous avons donné à chaque annotateur 1000 phrases à annoter sauf pour le dernier qui en a reçu 399. Les résultats sont décrits dans le tableau 5.8.

29. Spacy est une bibliothèque gratuite et open-source pour le traitement du langage naturel en Python. Elle permet de faire de l'étiquetage morphosyntaxique, la reconnaissance d'entités nommés, etc.

Pour ce deuxième corpus, les données soumises à l'évaluation n'étaient pas les mêmes pour chaque annotateur. En effet, l'objectif était de disposer rapidement d'un corpus d'apprentissage pour commencer le cadre expérimental. Ainsi, 3460 phrases sur les 6399 ont été collectées pour faire un mini-jeu de données. Ces phrases correspondent à celles annotées en FMC. Ce corpus a été constitué sur la même période que le corpus de transcription et de prises de note (entre septembre et décembre 2019). Le tableau 5.9 présente la répartition en termes de token et phrases.

	Jeu de données Amazon (3460 phrases)
Nombre total de phrases	3460 phrases
Nombre total de tokens (mots)	105 266 mots

TABLE 5.9 – Répartition du jeu de données Amazon en termes de mots et phrases.

Comme nous l'avons mentionné en début de cette section, l'objectif d'avoir recours à un corpus similaire à nos données initiales poursuivaient deux finalités : compléter les données initiales et développer une méthode d'extraction. Malheureusement, ces objectifs n'ont pas pu être atteints en raison du temps nécessaire à l'évaluation d'un tel corpus. C'est pour cela que nous avons opté pour une évaluation individuelle d'une portion du corpus par chaque annotateur.

5.4 Méthode d'extraction des freins, conditions et motivations

Toujours dans l'objectif de pouvoir rassembler suffisamment de données pour une classification en FMC, nous avons également exploré la création d'un ensemble de règles morphosyntaxiques pour caractériser premièrement les freins et les motivations identifiés dans le corpus Amazon.

Nous avons rédigé des règles à partir de l'observation des deux corpus de référence (corpus frein et motivation) issus de l'annotation réalisée par l'équipe Ixiade. Nous définissons ainsi des règles pour la catégorie frein et également pour la catégorie motivation. L'écriture des règles est réalisée en utilisant la librairie Spacy pour l'étiquetage et quatre lexiques de sentiments. En outre, nous précisons également que les règles sont écrites en prenant en compte uniquement les catégories morphosyntaxiques des termes, de ce fait, nous ne procédons pas à une analyse en dépendance syntaxique. Les différents lexiques que nous avons utilisés sont décrits

en 3.4.3 : le lexique Affect (Augustyn et al., 2006), la ressource JeuxDeMots (Lafourcade et al., 2015) le lexique Polarimots (Gala et Brun, 2012) et la ressource Feel (Abdaoui et al., 2017).

5.4.1 Méthode d'extraction : création des règles

Pour élaborer les règles pour le sous-corpus frein (2020 verbatims) que nous utiliserons dans la suite, nous nous sommes basés sur les travaux de Corblin et Tovena (2003). Leurs travaux portent sur l'étude des systèmes de négation dans les langues romanes. Généralement, on qualifie de négative une phrase dont la représentation sémantique comporte une négation. Considérons les exemples suivants :

— Je n'ai pas accepté un compromis

— J'ai refusé un compromis

La première phrase comporte une négation alors que la deuxième non. Les phrases négatives peuvent être aussi identifiées par une série de propriétés sémantiques, syntaxiques ou discursives, néanmoins, nous faisons le choix de nous concentrer uniquement sur les éléments de négation et les utiliser pour construire des règles d'extraction qui nous permettront de pouvoir extraire dans différents textes des phrases de type frein ou motivation.

5.4.1.1 Corpus frein

Nous écrivons des règles à partir de l'observation des deux corpus de référence (corpus frein et motivation) issus de l'annotation réalisée par l'équipe Ixiade. Nous définissons ainsi des règles pour la catégorie frein et également pour la catégorie motivation. L'écriture des règles est réalisée en utilisant la librairie Spacy et quatre lexiques de sentiments. En outre, nous précisons également que les règles sont écrites en prenant en compte uniquement les catégories morphosyntaxiques des termes, de ce fait, nous ne procédons pas à une analyse en dépendance syntaxique.

Dans notre cadre de travail, l'hypothèse est d'exploiter la présence du connecteur "mais" dans la phrase pour déterminer si cette dernière est un frein. Si le connecteur "mais" est précédé d'un terme (nous prenons en compte les adjectifs, adverbes, noms et verbes) et que ce dernier est positif, alors nous supposons que la suite de la phrase peut-être négative car "mais" est un connecteur qui exprime une opposition et parfois une concession. Le tableau 5.10 détaille la répartition du terme "mais" dans le corpus Amazon.

	Corpus Amazon annoté en frein
Nombre de "mais"	1896
Nombre de phrases contenant le "mais"	1814
Nombre de phrases sans le "mais"	206

TABLE 5.10 – Aperçu du corpus Amazon annoté en frein.

Nous avons préalablement établi un ensemble de règles[30] à partir du sous-corpus d'Amazon constitué uniquement de phrases annotées en frein. À partir du corpus de référence de frein, nous cherchons et définissons un schéma général de structures négatives se trouvant après le connecteur "mais". Pour cela, nous utilisons le logiciel TXM . TXM est une plateforme modulaire et open-source de textométrie. La sélection des règles est basée sur les travaux de Corblin et Tovena (2003) et sur la fréquence d'apparition de ces dernières dans le corpus. En utilisant ces règles pour valider leur robustesse sur le corpus de référence de motivation, nous avons constaté que ces règles permettent également de retourner des phrases annotées en motivations dans notre corpus de référence de motivation (648/1059, soit 61%). Les résultats sont détaillés au tableau 5.11.

	Corpus Frein (2020)	**Corpus Motivation (1059)**
Total des phrases extraites sans doublons	**1249 (61%)**	**648 (61%)**

TABLE 5.11 – Résultats des extractions sur les deux corpus avec les règles de frein.

Le pourcentage de l'union des phrases extraites par toutes les règles ne dépasse par les + 61% pour le corpus frein. Ceci reste relativement inférieur au pourcentage attendu (+ 80%). Pour conclure, ces règles ne permettent pas de clairement différencier les phrases appartenant au corpus de référence de frein de celles du corpus de motivation. Dans la suite, nous présentons les résultats obtenus par les règles mises en œuvre à partir du corpus de référence de motivation.

5.4.1.2 Corpus motivation

L'objectif de cette partie est de rédiger des règles de motivations afin d'extraire des phrases de type motivation. Les règles sont rédigées en analysant linguistiquement une vingtaine de phrases à partir desquelles nous réalisons des premières extractions. Le tableau 5.12 détaille la répartition du terme "mais" dans le sous-ensemble du corpus Amazon, annoté uniquement en motivation.

30. Un tableau de ces règles est donné en annexe.

	Corpus Amazon annoté en motivation
Nombre de "mais"	1035
Nombre de phrases contenant le "mais"	994
Nombre de phrases sans le "mais"	65

TABLE 5.12 – Aperçu du corpus Amazon annoté en motivation.

En ce qui concerne le corpus de référence de motivation, nous avons mis en place une vingtaine de règles. Nous prenons également en compte la présence de structures négatives dans les phrases de motivation en rédigeant nos règles Le tableau 5.13 détaille les résultats obtenus.

	Corpus Frein (2020)	**Corpus Motivation (1059)**
Total phrases retournées avec règles de motivation	940 (46%)	568 (53%)

TABLE 5.13 – Résultats des extractions sur les deux corpus avec les règles de motivation.

Limitations : Nous observons que les règles d'extraction écrites pour le corpus motivation permettent d'extraire 53% (568 phrases sur 1059) dans le corpus de référence de motivation, mais également environ 940 phrases dans le corpus frein (soit 46% sur 2020). Nous concluons qu'une partie (environ la moitié) des phrases des deux corpus semblent donc être assez proche au niveau de la structure. Pour véritablement valider les extractions renvoyées par la méthode d'extraction à base de règles, nous proposons d'utiliser des méthodes de filtrage pour valider la classe finale des extractions.

L'ensemble des premières règles élaborées permettent difficilement de dépasser les plus de 60% du total des phrases annotées. Par ailleurs, comme souligné dans les analyses ci-dessus, certaines règles construites pour le repérage des phrases de type "frein" permettent également de remonter des phrases annotées en motivations. Le cas contraire est également observé pour les règles de motivations qui permettent également de retourner une grande quantité de phrases annotées originellement en freins comme des motivations. Pour pallier ce problème, nous avons également proposé de mettre en œuvre une méthode de filtrage pour valider la classe définitive de la phrase extraite et réduire la précision des phrases extraites comme des freins ou motivations, mais qui n'appartiendraient pas à la classe d'extraction bien qu'extraite avec l'une des règles de cette même classe. Plus spécifiquement, si une phrase est extraite avec une règle de frein, la méthode de filtrage nous permettra de valider ou non l'appartenance de la phrase à cette classe ou à la classe de motivation.

5.4.2 Méthode de filtrage par somme

Cette méthode consiste à valider la classe d'une phrase en faisant la somme des mots positifs et des mots négatifs avant et après le connecteur "mais". Ensuite, nous élaborons une liste de conditions pour déterminer la classe finale de la phrase extraite. Cette liste est détaillée au tableau 5.14. Plus spécifiquement, si la somme des mots négatifs est supérieur à celle des mots positifs avant le "mais" et qu'après le "mais", on observe la même tendance, la phrase est classée comme appartenant à la classe frein.

Récapitulatif des conditions
Si polarité positive (à gauche) **mais** et polarité négative (à droite) => Frein
Si polarité positive (à gauche) **mais** et polarité positive (à droite) => Motivation
Si polarité négative (à gauche) **mais** et polarité négative (à droite) => Frein
Si polarité négative (à gauche) **mais** et polarité positive (à droite) => Frein
Si polarité neutre (à gauche) **mais** et polarité négative (à droite) => Frein
Si polarité négative (à gauche) **mais** et polarité neutre (à droite) => Frein
Si polarité neutre (à gauche) **mais** et polarité positive (à droite) => Motivation
Si polarité positive (à gauche) **mais** et polarité neutre (à droite) => Motivation
Si polarité neutre (à gauche) **mais** et polarité neutre (à droite) => Neutre

TABLE 5.14 – Liste des conditions pour la méthode de filtrage par somme.

Observations : Nous remarquons que la méthode proposée permet de valider + 70% des phrases extraites comme des motivations tandis que le nombre de phrases validées pour le corpus frein se limitent à peine à 36%. Au vu de cela, nous proposons d'explorer une autre méthode de filtrage (voir tableau 5.17).

	Validées en frein	***Validées en motivation***	**Total extrait**	**Total corpus**
Corpus frein - Extraites avec règles de frein	*458 (36%)*	*791*	*1249 (61%)*	2020
Corpus motivation - Extraites avec règles de motivation	*160*	*408 (72%)*	*568 (53%)*	1059

TABLE 5.15 – Résultats des phrases extraites avec les règles de motivations comme freins et validées avec la méthode de filtrage par somme.

5.4.3 Méthode de filtrage par pivot

La méthode de filtrage par somme a montré des résultats peu satisfaisants (36% de phrases validées en frein sur les 61% extraites avec des règles de freins) pour le

corpus frein. Les résultats pour les freins ne sont clairement pas satisfaisants bien que pour les motivations, nous obtenons environ 72% de validation. Nous décidons d'exploiter une autre méthode de filtrage et nous décrivons les résultats obtenus dans la suite.

La méthode de filtrage par pivot utilise la position de chaque mot dans la phrase. L'hypothèse est de considérer que les mots présents après le connecteur "mais" ont plus de chance d'exprimer le sentiment de l'auteur. Plus spécifiquement :

1. nous déterminons la position de tous les mots de la phrase ;
2. nous déterminons si le mot est avant le connecteur "mais" ou après :
   - si le mot est avant (qu'il soit positif ou négatif), on diminue la valeur de l'index de la position selon une valeur fixe établie au départ (entre 1 à 4) ;
   - si le mot est après et qu'il est négatif, on ajoute le double de la valeur initialisée à la valeur de son index ;
   - si le mot est après et qu'il est positif, on diminue la valeur de son index par la valeur initialisée en début.
3. nous faisons la somme des valeurs obtenues pour tous les mots positifs et négatifs avant et après le "mais" et nous divisons cette somme par le nombre de mots de la phrase. La valeur obtenue pour les mots positifs et négatifs avant et après le connecteur "mais" est comparée en fonction des conditions énumérées au tableau 5.16. Les résultats sont donnés au tableau 5.17.

Récapitulatif des conditions
Si polarité positive (à gauche) **mais** et polarité négative (à droite) => Frein
Si polarité positive (à gauche) **mais** et polarité positive (à droite) => Motivation
Si polarité négative (à gauche) **mais** et polarité négative (à droite) => Frein
Si polarité négative (à gauche) **mais** et polarité positive (à droite) => Frein
Si polarité neutre (à gauche) **mais** et polarité négative (à droite) => Frein
Si polarité négative (à gauche) **mais** et polarité neutre (à droite) => Frein
Si polarité neutre (à gauche) **mais** et polarité positive (à droite) => Motivation
Si polarité positive (à gauche) **mais** et polarité neutre (à droite) => Motivation
Si polarité neutre (à gauche) **mais** et polarité neutre (à droite) => Neutre

TABLE 5.16 – Liste des conditions pour la méthode de filtrage par pivot.

Nous observons que la méthode de filtrage par pivot affiche un score 70% de phrases extraites en freins et validées en freins pour le corpus frein contrairement à la méthode de filtrage par somme (36%). S'agissant des phrases extraites avec

les règles de motivations et validées comme motivations, nous avons observé une baisse par rapport à la méthode de filtrage par somme (45% contre 72%). De manière générale, nous avons observé que les règles élaborées peinent difficilement à caractériser les phrases de types frein et les phrases de type motivation. Ainsi, les règles élaborées pour une classe spécifique permet également de remonter les phrases appartenant à une autre classe. De même les méthodes de filtrage ne permettent pas de filtrer au maximum les phrases.

	Validées en frein	***Validées en motivation***	**Total extrait**	**Total corpus**
Corpus frein - Extraites avec règles de frein	*873 (70%)*	*376*	*1249 (61%)*	2020
Corpus motivation - Extraites avec règles de motivation	*315*	*253 (45%)*	*568 (53%)*	1059

TABLE 5.17 – Résultats des phrases extraites avec les règles de motivations comme freins et validées avec la méthode de filtrage par pivot.

5.4.4 Conclusion

Les règles établies n'ont pas permis de véritablement discriminer les classes de notre jeu de données. Une hypothèse pourrait être également liée au fait que le corpus ait été annoté par différentes personnes sans mise en place d'un accord inter-annotateur. Devant le manque de résultats, nous avons décidé d'explorer le domaine de l'amplification de données du corpus de verbatim.

5.5 Corpus Yoomaneo

5.5.1 Présentation de la plateforme Yoomaneo

Yoomaneo [31] (figure 5.5) est une application communautaire gratuite et ouverte à tous. Elle a été créée en 2020 par la société Ixiade. La création de Yoomaneo poursuit deux objectifs :

1. construire une base de données d'individus prêts à participer à des études d'Ixiade ;
2. créer et animer différentes communautés d'individus désireux de partager leurs opinions ou avis sur des concepts innovants.

31. https://www.yoomaneo.com/fr

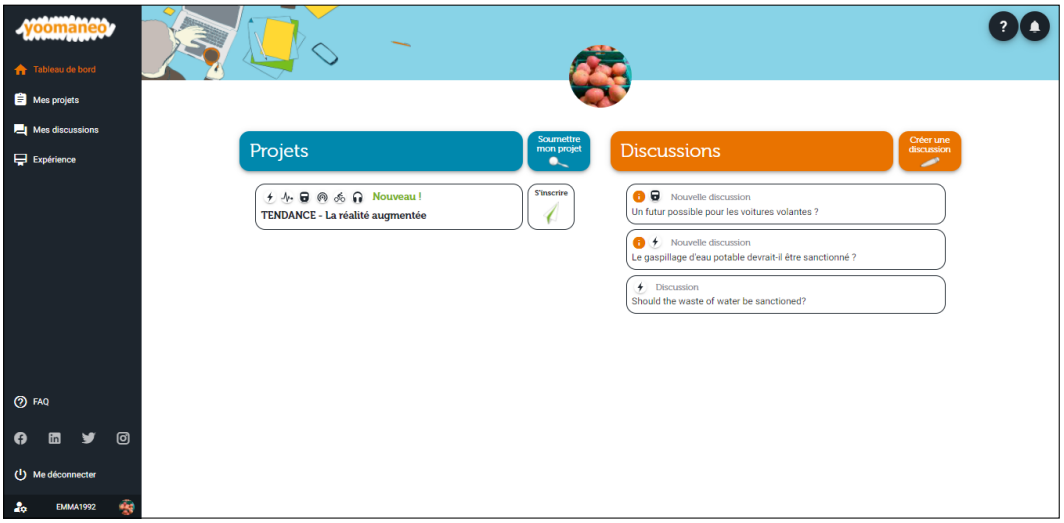


FIGURE 5.5 – Aperçu de la page d'accueil de la plateforme Yoomaneo.

Yoomaneo est une application accessible sur iOS, Android et sur WebApp (via la majorité des navigateurs internets). Depuis 2021, l'application est utilisée comme un outil de collecte de commentaires en ligne sur des concepts innovants pour de véritables études d'**acceptabilité**. Elle se compose de deux espaces :

1. **Un espace de discussions et d'échanges** autour de **sujets variés et de réflexions** autour de l'innovation. L'espace discussion contient à ce jour 6 communautés publiques (**MobilitéTransport**, **Santé**, **Univers digital et connecté**, **Énergie**, **Modes de vie/consommation**, **Loisirs et Bien-être**) amené à croître. Ces 6 communautés ont été réfléchies et conçues après discussions et échanges au sein d'Ixiade. À travers ces 6 thématiques, Yoomaneo a le potentiel d'adresser un large nombre d'individus. De plus, les individus peuvent choisir les communautés pour lesquelles ils souhaitent être membres et ils ont également le libre choix de changer de communauté ; leur choix initial n'étant pas figé. Lorsqu'un membre de la communauté ou un animateur publie une discussion, chaque utilisateur reçoit la notification.
2. **Un espace projet** où les membres de la communauté répondant aux critères d'éligibilité [32] peuvent participer à des **projets d'innovation réels**. Les questions sont posées au jour le jour à une heure définie afin de permettre au plus grand nombre de participer. Durant l'étude, les participants répondent

32. Ces critères peuvent être liés à la catégorie socioprofessionnelle, au genre, etc. Ils sont définis par le chargé de projet en fonction de l'objectif de projet.

de manière régulière aux questions ouvertes. Ensuite, les répondants peuvent également voir les réponses des autres utilisateurs et échanger avec les autres utilisateurs. Les études durent en moyenne 5 à 10 jours. À la fin, une extraction de ses réponses donne un matériel riche afin de produire une analyse pertinente du concept testé.

Yoomaneo s'adresse autant aux professionnels qu'au grand public. Le recrutement pour participer aux projets se fait directement par Ixiade qui identifie les profils des individus correspondant aux études en cours. Une fois contacté, les individus sont invités à télécharger l'application. Après avoir participé à l'étude, ces nouveaux membres restent sur l'application et deviennent des membres actifs dans l'espace discussion.

5.5.2 Nature des données

Les données provenant de la plateforme Yoomaneo sont de nature différente des données issues de transcription. Ce sont des données saisies en ligne par des participants contrairement aux données orales transcrites. Ces dernières peuvent contenir des fautes de frappe, d'accord, de ponctuation et l'usage d'apostrophes et d'espaces peut diverger de l'usage standard. D'autre part, elles ne sont pas révisées comme les transcriptions. Nous pouvons également observer l'absence de hashtags dans ces posts. De manière générale, même si nous pouvons constater des similitudes entre les posts sur Yoomaneo avec des contenus postés en ligne, nous sommes dans un type d'écrit qui s'éloigne de celui que l'on peut retrouver sur des réseaux sociaux grand public type Facebook, Twitter ou des simples SMS. Sur ces réseaux sociaux, les individus écrivent presque instantanément en essayant d'utiliser le moins de ponctuation possible. Dans notre cadre, un temps de réflexion est donné aux participants pour répondre et passé ce délai, ils n'ont plus la possibilité de revenir, voire de modifier leurs réponses.

5.5.3 Constitution du corpus

Après avoir présenté la nature des données Yoomaneo, nous en venons à présent sur la méthodologie poursuivie pour constituer un corpus d'apprentissage. Pour construire notre nouveau corpus d'apprentissage, nous avons recueilli 755 réponses ou posts provenant de 4 études réalisées dans l'espace projet sur la plateforme Yoomaneo. Ces projets portaient sur l'évaluation de différents concepts innovants dans 3 domaines différents : santé, bien-être et électrique (2 projets). Ces projets touchaient autant les professionnels que le grand public. Ces données ont été exportées

directement depuis la plateforme Yoomaneo dans un tableur .xlsx où chaque ligne contient les informations suivantes : le numéro de la question, l'auteur du post, le code de la réponse auquel le post répond, les points de pertinence attribués au post par les autres répondants et les émoticônes [33]. Toutefois, les participants peuvent utiliser les smileys, mais cela reste très minime dans les textes observés (les commentaires présents dans la partie discussion).

Les posts recueillis ont été ensuite donnés à évaluer à trois chargés d'études. La procédure d'évaluation est similaire à celle mentionnée à la section 5.2.3. Elle a été réalisée de juillet 2021 à décembre 2021. Nous retenons uniquement les posts qui reçoivent au minimum deux évaluations similaires. De ce fait, sur les **755 évalués**, **433** ont été assignés à la classe *motivation*, **112** à la classe *frein*, **97** à la classe *condition*, **65** ont été jugés inclassifiables et **48** n'ont reçu aucun accord. Sur l'ensemble des posts annotés (figure 5.6) :

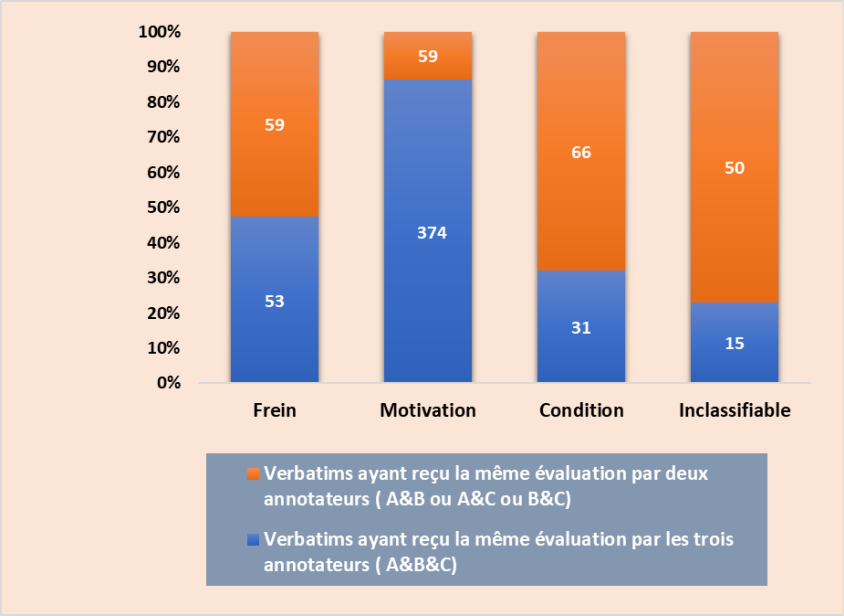


FIGURE 5.6 – Répartition des évaluations par classe selon les différents couples d'annotateurs.

- **473** posts ont reçu la même évaluation par les 3 annotateurs. Sur les 473, **374**

33. Il est possible d'intégrer des émoticônes dans Yoomaneo. Par contre, ces derniers sont proposés dans un champ séparé de la réponse. Ce mode a été mis en œuvre pour une autre méthode d'analyse utilisée par l'entreprise.

sont de la classe motivation, **53** sont des freins, et **31** des conditions et le reste est jugé inclassifiable, soit **15**.

- **234** ont reçu la même évaluation par deux annotateurs quel que soit le couple d'annotateurs (Annotateur A et B ou A et C ou encore C et B). Sur les 234, **59** sont de la classe motivation, **59** sont des freins, et **66** des conditions et le reste est jugé inclassifiable, soit **50**.
- **48** n'ont pas reçu la même classe de la part des trois annotateurs.

En somme, 642 posts ont été retenus pour faire partie du corpus d'apprentissage Yoomaneo. Ils sont répartis comme le montre le tableau 5.18. Au tableau 5.19, nous présentons les informations liées au nombre de phrases et de mots dans le corpus Yoomaneo.

	Corpus Yoomaneo
Nombre de freins	112
Nombre de motivations	433
Nombre de conditions	97
Total	**642**

TABLE 5.18 – Répartition des classes sur le corpus Yoomaneo.

	Jeu de données de Yoomaneo corpusY - 642 verbatim
Nombre total de phrases	934 phrases
Nombre total de tokens (mots)	15 530 mots
Nombre moyen de phases par verbatim	1,45
Nombre de post ayant une phrase	461 commentaires
Nombre de post ayant 2 phrases	117 commentaires
Nombre de post ayant 3 phrases	39 commentaires
Nombre de post ayant 4 phrases	11 commentaires
Nombre de post ayant plus de 5 phrases (le post le plus long en contient 6)	14 commentaires

TABLE 5.19 – Répartition du jeu de données en termes de mots et phrases.

5.6 Conclusion

Dans ce chapitre, nous nous sommes intéressés aux différentes procédures mises en place pour constituer des données d'apprentissage pour notre tâche de classification. Nous avons présenté tout d'abord la méthodologie élaborée pour le corpus de transcriptions et de prise de notes. Nous avons décrit ensuite la méthode appliquée pour constituer le jeu de données Amazon. Puis, nous avons proposé de catégoriser nos classes en mettant en œuvre une méthode d'extraction de freins et de filtrage. Cette méthode s'est révélée peu robuste et efficiente pour être utilisée pour amplifier notre corpus Amazon. Nous avons enfin présenté la procédure pour obtenir des données d'apprentissage issues de la plateforme Yoomaneo. Comme mentionné au chapitre 2, la clé de la construction d'un bon modèle d'apprentissage automatique réside dans les données sur lesquelles le modèle a été appris. Comme nous pouvons l'observer, pour l'ensemble des données collectées, la quantité est non seulement faible, mais les corpus sont assez déséquilibrés au niveau de la répartition des classes. En effet, l'apprentissage en général et plus particulièrement l'apprentissage profond requiert une grande quantité de données (**beaucoup plus qu'un modèle traditionnel**) et correctement étiquetées. Ayant à notre disposition une quantité de données insuffisante, nous nous sommes penchés sur les techniques permettant d'augmenter nos jeux de données pour construire des modèles robustes.

Chapitre 6

Techniques d'amplification de corpus

Ce chapitre s'intéresse à l'amplification [1] des données et en présente un état de l'art général. L'amplification des données pour le TAL permet de disposer d'un corpus d'entraînement plus important quantitativement. Ce corpus permet ainsi aux modèles de généraliser plus facilement sur des nouvelles données. La section **??** présente la problématique de l'amplification des données textuelles pour le TAL. La section 6.2 présente un état de l'art des méthodes d'amplification pour différentes tâches et la section **??** détaille les méthodes implémentées dans cette thèse.

6.1 Problématique de l'amplification des données textuelles pour le TAL

Un des problèmes majeurs dans le domaine du langage naturel est le manque de données étiquetées. En effet, l'insuffisance de données pour les tâches d'apprentissage classique comme profond est un défi permanent. Les données disponibles ne sont pas suffisantes pour entraîner un classifieur afin de permettre à ce dernier de pouvoir repérer les régularités dans les textes fournis. De plus, l'annotation de telles données consomme énormément de temps et demande beaucoup de ressources. C'est dans cette perspective qu'il devient de plus en plus pressant pour certaines tâches et en fonction des algorithmes à utiliser (i.e, réseau de neurones) d'amplifier les données. L'amplification des données regroupe toutes les techniques permettant d'amplifier la quantité de données disponibles en ajoutant des copies légèrement

1. Coulombe (2020) précise qu'il est préférable d'utiliser le terme *amplification* qu'*augmentation*, car il s'agit de la création de « nouvelles données en préservant le sens qui demeure invariant ». Le deuxième terme est généralement utilisé par abus de langage.

modifiées des données initiales (Li et al., 2021a) ou de générer artificiellement des données à partir des données initiales au moyen de transformations (Taylor et Nitschke, 2018).

Dans le domaine de la vision par ordinateur, il existe de nombreuses méthodes pour créer artificiellement de telles données. En ce qui concerne les images, les transformations telles que les rotations (voir figure 6.1 ou les modifications du canal RVB [2] sont appropriées.

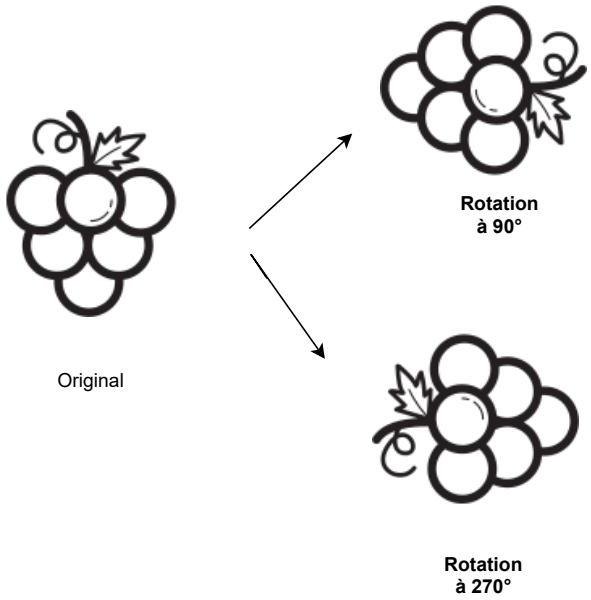


FIGURE 6.1 – Exemples d'augmentation d'une image en utilisant la méthode de rotation.

Comme on peut l'observer sur la figure 6.1, nous obtenons, après rotation, diverses images qui seront uniques pour le modèle de reconnaissance. L'amplification des données est une tâche très populaire en classification d'images, car il est assez aisé et facile de générer de nouvelles images en les modifiant légèrement (He et al., 2016). Pour la reconnaissance vocale, on utilisera des procédures qui peuvent modifier l'intensité du son ou la vitesse des données audios (Park et al., 2019). Néanmoins, dans le domaine du langage naturel, c'est une tâche assez complexe, car les données textuelles sont difficiles à traiter, car elles peuvent être hautement bruitées (Coulombe, 2020) à l'exemple des commentaires ou tweets. Ce bruit peut notamment être des fautes d'orthographes, des lettres manquantes, etc. En outre,

2. En anglais « RGB : red, green, blue » est un système de codage des couleurs.

l'amplification des données pour le traitement du langage naturel doit pouvoir établir des règles universelles pour la transformation de données textuelles et ces dernières doivent être exécutées automatiquement tout en maintenant la qualité de l'étiquetage des données (Bayer et al., 2021). Cette complexité est également liée à la structure des langues, aux différentes catégories de mots et les multiples niveaux d'organisation de la langue. S'il est plus facile de transformer des données d'images et sons, il est bien plus compliqué de le faire avec du texte en demeurant compréhensible. La notion d'amplification des données est très vaste et englobe diverses recherches dans différents sous-domaines de l'apprentissage automatique. Même si de nombreux travaux scientifiques relient simplement l'amplification des données à l'apprentissage profond, elle est fréquemment appliquée dans tout contexte de l'apprentissage automatique. En outre, même si une amplification des données d'apprentissage ne permet pas toujours de trouver une solution au problème d'apprentissage, les données restent toujours décisives pour la qualité d'un classifieur supervisé (Kobayashi, 2018; Wei et Zou, 2019). Dans leurs travaux, Banko et Brill (2001) démontrent que disposer d'une très grande quantité de données améliore la qualité d'un processus d'apprentissage automatique supervisé pour une tâche de désambiguïsation pour différents algorithmes d'apprentissage classiques et super simples comme le perceptron ou le Naive Bayes. Pour eux, le choix du classifieur ne conduit pas forcément à un changement significatif. Toutefois, ils recommandent quand même de réfléchir ou d'examiner davantage la sélection et le développement d'algorithmes robustes plutôt que de miser sur le développement de gros corpus dont le coût en temps est non négligeable.

Lorsque l'on parle d'amplification des données, une notion importante à prendre en considération est la *préservation des étiquettes*. En effet, si le remplacement de mots par des mots similaires permet de préserver le sens de la phrase ou du document, l'ajout aléatoire de certains mots peut modifier la classe de la phrase comme le montre la figure 6.2. Par exemple, en analyse de sentiment, l'ajout de certains mots à des positions critiques pourraient altérer la classe du nouveau texte généré par rapport à l'original.

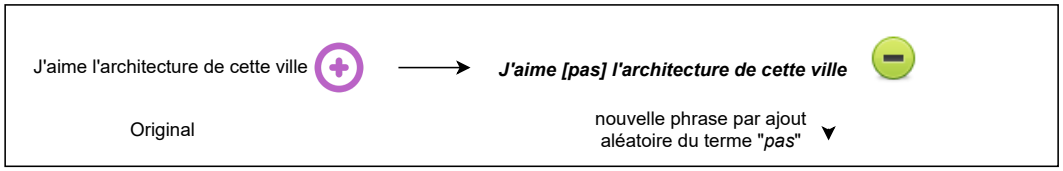


FIGURE 6.2 – Exemple d'une altération d'une phrase par ajout d'un mot à une position aléatoire.

De ce fait, il est important de bien comprendre ces données et de choisir la méthode appropriée pour amplifier le volume de ses données d'apprentissage. Toutes les méthodes n'apportent pas forcément le même gain au processus d'apprentissage.

6.2 Méthodes d'amplification des données : état de l'art

6.2.1 Approches par injection de bruit synthétique

Les approches par injection de bruit synthétique sont des méthodes dont l'objectif est d'altérer légèrement les mots ou les phrases en vue de préserver l'orientation sémantique de ces derniers et de permettre à l'algorithme d'apprendre des généralisations sur des données proches de la réalité et d'être plus robuste aux perturbations ou erreurs présentes dans les données d'entraînement (Bayer et al., 2021). Les méthodes par injection de bruit synthétique peuvent être appliquées au niveau du caractère ou du mot. Elles ont essentiellement été appliquées pour amplifier des corpus pour des tâches de classification thématique de texte, d'analyse de sentiment, de traduction automatique ou encore de génération.

• La technique d'échange ou encore *swapping* consiste à échanger la position de deux mots dans une phrase ou de deux caractères dans un mot. Bien que certaines langues soient sensibles à l'ordre des mots, cette technique peut être utilisée comme technique d'amplification de données si elle est faiblement appliquée sur le jeu de données. Ceci suppose de définir préalablement le nombre de mots/caractères à échanger, de la position dans la phrase ou au niveau du mot. Cette estimation doit également s'appuyer sur le nombre de mots présents dans la phrase, le nombre de caractères des mots ou encore la longueur des phrases.

• L'insertion aléatoire ou encore *insertion* consiste à insérer de manière aléatoire des caractères dans un mot ou des mots dans une phrase ou une phrase dans un texte.

• La suppression aléatoire ou encore *deleting* consiste à supprimer aléatoirement un mot dans la phrase ou un caractère dans un mot.

• La substitution ou encore *substitution* consiste à remplacer un mot ou un caractère par un autre mot ou un autre caractère alphabétique. Ce mot peut être un mot de la même nature (partie du discours), c'est-à-dire un synonyme.

Le tableau 6.1 présente des exemples d'altération de phrases au niveau du caractère en utilisant les différentes techniques mentionnées ci-dessus.

Méthodes	Niveau	Exemples	Phrase générée
Echange	Caractère	Une fille nettoie la salle de conférence.	Une f**li**le nettoie la s**la**le de conférence.
	Mot	Une fille nettoie la salle de conférence.	Une **nettoie fille** la salle de conférence.
Insertion	Caractère	Une fille nettoie la salle de conférence.	Une fil**u**le nettoie la salle de confér**m**ence.
	Mot	Une fille nettoie la salle de conférence.	Une **jeune** fille nettoie la **table** salle de conférence.
Suppression	Caractère	Une fille nettoie la salle de conférence.	Un fille nettoie la sall de conférenc.
	Mot	Une fille nettoie la salle de conférence.	Une fille nettoie la salle de .
Substitution	Caractère	Une fille nettoie la salle de conférence.	Une fille nettoie **z**a salle de conférence.
	Mot	Une fille nettoie la salle de conférence.	Une **dame** nettoie **une** salle de conférence.
Mélange de toutes les méthodes	Caractère	Une fille nettoie la salle de conférence.	Un f**li**le nettoie **z**a salle de**e** conférence.
	Mot	Une fille nettoie la salle de conférence.	Une **jeune nettoie fille une** salle de .

TABLE 6.1 – Exemples de phrases obtenues en implémentant les méthodes d'injection de bruit.

6.2.1.1 Tâches et méthodes : niveau du caractère

Belinkov et Bisk (2017) utilisent des méthodes d'injection de bruit synthétique dans leurs travaux pour l'amélioration de la robustesse des modèles de traduction neuronaux sur des textes bruités afin que leur modèle de traduction neuronale soit le moins sensible aux *exemples contradictoires* [3]. Ils proposent ainsi quatre techniques d'injection de bruit synthétique : *swap, middle random, fully random* et *keyboard typo*. L'objectif pour eux est d'évaluer la performance et la robustesse de différents modèles pour une tâche de traduction neuronale de textes en français, allemand et tchèque vers l'anglais. Ils utilisent trois corpus distincts : WiCoPaCo [4], CzeSL [5] et un jeu de données regroupant deux corpus en langue allemande : MERLIN [6] et RWSE [7].

3. Adversial examples ou exemples contradictoires en français sont des entrées de modèles d'apprentissage automatique conçues intentionnellement pour que le modèle fasse une erreur (Saporta et al., 2021).

4. **WiCoPaCo** (Max et Wisniewski, 2010) est un corpus en langue française de réécritures naturelles extraites de l'historique des révisions de Wikipédia.

5. Le **CzeSL** (Šebesta et al., 2017) est un jeu de données en langue tchèque sur la correction des erreurs grammaticales.

6. Le corpus **MERLIN** (Wisniewski et al., 2013) est constitué de 2 286 textes destinés aux apprenants d'italien, d'allemand et de tchèque, issus d'examens écrits d'institutions de test reconnues.

7. Le **RWSE Wikipedia Revision Dataset** (Zesch, 2012) est constitué d'un ensemble de données d'erreurs orthographiques de mots réels extraites de l'historique des révisions de Wikipédia. Chaque instance est constituée de la phrase originale contenant une erreur et de la phrase où l'erreur a été corrigée. Une instance contient également l'identifiant de l'article Wikipédia ainsi que de la révision, de sorte que l'instance peut être retracée jusqu'à l'article Wikipédia original.

— La technique *Swap*, en français *échange* ou *permutation* consiste à permuter deux lettres dans un mot. La permutation est limitée à une fois par mot à l'exception de la première et la dernière lettre qui ne sont pas concernées. De ce fait, cette technique est appliquée uniquement aux mots ayant une longueur supérieure à quatre. (i.e, *order* → *od**re**r*)

— La technique *Middle random* consiste à permuter toutes les lettres d'un mot à l'exception de la première et la dernière lettre. (i.e, *order* → *o**erd**r*)

— La technique *Fully random* consiste à permuter toutes les lettres d'un mot. Cette technique ne tient pas compte de l'ordre des lettres comme les précédentes.(i.e, *order* → ***rrode***)

— La technique *Keyboard typo* consiste à remplacer de manière aléatoire une lettre d'un mot par une lettre de l'alphabet. (i.e, *order* → *or**z**er*)

Belinkov et Bisk (2017) constatent que les modèles CNN [8] formés sur un seul type de bruit ne sont pas bons sur un autre type. Par contre, les modèles formés sur différents types de bruits qu'on appellera *modèles mixtes* sont plus robustes aux types de bruit sur lesquels ils ont été entraînés. Pour la langue française, le modèle mixte entraîné sur le corpus constitué avec les techniques *Fully random* et *Keyboard typo* obtient un score BLEU [9] de 39.13 sur le corpus test (*Fully random*) par rapport à 39.73 obtenu avec le modèle entraîné uniquement sur le corpus obtenu avec la technique *Fully random*.

Feng et al. (2020) ont également exploré diverses méthodes d'amplification de données textuelles dont des méthodes par injection de bruit textuel (*échange de deux caractères côte à côte, suppression et insertion de caractère*) pour produire différentes versions de leur corpus d'entraînement [10] et analyser leurs effets pour la tâche de génération. Le corpus d'apprentissage qu'ils utilisent est constitué d'un sous ensemble de commentaires/avis en langue anglaise sur des commerces locaux tels que les restaurants, hôtels, etc. provenant de la plateforme Yelp [11]. Ensuite, ils

8. Un réseau de neurones convolutif est un type de réseau de neurones utilisé essentiellement pour la classification d'images. Il est constitué d'un empilement de couches (LeCun et al., 2015) : une couche de convolution qui constitue la base du réseau, une couche de pooling qui applique une opération de pooling à la sortie de la couche de convolution (l'opération consiste à réduire la taille des images tout en conservant les caractéristiques jugées importantes de ces dernières), une couche de correction qui joue le rôle de fonction d'activation et une couche *fully connected* qui constitue la dernière couche (c'est celle-ci qui classifie l'image donnée en entrée au réseau CNN).

9. BLEU est un algorithme qui évalue la qualité d'une traduction en fournissant un score.

10. Ils récupèrent un sous ensemble du corpus Yelp soit 67 000 commentaires sur près de 6 990 280 commentaires. 50 000 commentaires sont utilisés pour l'apprentissage, 15 000 pour le corpus de validation et 2 000 pour le test.

11. https://www.yelp.com/dataset.

adaptent finement GPT-2 [12] sur les différents corpus obtenus. Tout comme Belinkov et Bisk (2017), ils conservent le premier et le dernier caractère du mot à bruiter pour imiter le plus fidèlement le bruit naturel et les fautes de frappe. Ils observent que l'injection de bruit apporte des résultats bien meilleurs que s'ils se contentaient uniquement du corpus initial (le sous ensemble récupéré).

Cette technique a également été utilisé par Coulombe (2020) dans ces travaux relatifs à l'amplification de données textuelles pour une tâche de prédiction de la polarité d'opinions sur un corpus de critiques de films en langue anglaise. Le nombre de classes était de deux : positive et négative. Il propose 6 méthodes simples d'amplification des données dont une méthode d'injection de bruit textuel faible. Il retire, remplace ou ajoute de manière aléatoire un caractère alphabétique provenant soit d'une table de caractères, soit d'une table préétablie (table des erreurs les plus fréquentes en Reconnaissance optique de caractères et une table des erreurs de frappes au clavier). Coulombe (2020) propose également d'insérer des fautes d'orthographes. L'idée est de générer des textes contenant des fautes d'orthographe courantes afin d'entraîner des modèles qui deviendront ainsi plus robustes à ce type particulier de bruit textuel. Il observe que l'injection de fautes d'orthographe apporte de bonnes performances pour la tâche de détection de la polarité des critiques. Le meilleur résultat est ainsi obtenu par le modèle Xgboost qui augmente la précision de bonnes étiquettes de 1.5% par rapport à la base de référence sans amplification.

6.2.1.2 Tâches et méthodes : niveau du mot

L'injection du bruit au niveau du mot a été introduit par Wei et Zou (2019) dans leurs travaux relatifs à l'amélioration des performances des modèles neuronaux pour différentes tâches de classification de textes en langue anglaise. Ces tâches incluent :

- l'analyse de sentiment sur les corpus SST [13], CR [14] et le PR [15] ;

12. GPT-2 est un modèle pré-entrainé pour la génération de texte (voir chapitre 4 section 4.4).

13. **Le corpus Stanford Sentiment Treebank** (Socher et al., 2013) est un corpus de commentaires annoté avec 5 étiquettes (SST-5 : de très positif à très négatif) et également en deux étiquettes (SST-2 : positive et négative)

14. **Un corpus de commentaires** (Hu et Liu, 2004).

15. **Un corpus de phrases comparatives** (Ganapathibhotla et Liu, 2008) dans lesquels les auteurs comparent deux entités différentes. Par exemple, la qualité de l'appareil photo X est meilleure que celle de Y.

- la détection de la subjectivité sur le corpus SUBJ [16] ;
- la classification des types de questions pour le corpus TREC [17].

Ils proposent un ensemble de techniques regroupés sous l'appellation *EDA (Easy data augmentation*). Pour une phrase donnée, ils choisissent aléatoirement et effectuent une insertion, un échange aléatoire ou encore une suppression aléatoire d'un mot qui n'est pas un mot vide. Wei et Zou (2019) font varier le nombre de mots modifiés, n, pour l'insertion, l'échange et le remplacement (section **??**) en fonction de la longueur de la phrase l avec la formule $n{=}\alpha l$, où α est un paramètre qui indique le pourcentage de mots à modifier dans une phrase. Ils réalisent leurs expériences en utilisant des réseaux neuronaux récurrents (RNN) et des réseaux neuronaux convolutifs (CNN). Les résultats montrent que les architectures entraînées sur les 5 corpus augmentés avec toutes les méthodes *EDA* obtiennent une moyenne d'exactitude d'ensemble de 88.6% (en utilisant 50% des données d'entraînement disponibles) par rapport à 88.3% obtenu sur les données non augmentées pour l'ensemble des tâches.

La méthode de Wei et Zou (2019) a été utilisée dans d'autres travaux comme ceux de Feng et al. (2020) sur une tâche de génération de texte (corpus Yelp[11]), Rastogi et al. (2020) pour une tâche de classification de commentaires [18] ou encore dans Longpre et al. (2020) pour différentes tâches comme l'analyse de sentiment, la détection de subjectivité, la classification de type de question et la tâche de similarité sémantique et d'inférence de textes du référentiel d'évaluation Flue. Longpre et al. (2020) analysent l'apport de deux méthodes d'amplification de données dont l'injection de bruit et de la rétrotraduction (voir 6.2.3). Pour cela, ils utilisent trois architectures de type transformer BERT, XLNET-BASE [19] et RoBERTa-BASE sur 6 corpus différents. Ils observent pour les deux techniques d'amplification une amélioration marginale (+1% pour les tâches de classification) pour 5 des 6 corpus utilisés pour le modèle BERT et aucune amélioration pour les deux autres architectures.

16. **Le corpus de subjectivité** (Pang et Lee, 2004) est un corpus de phrases annoté en deux étiquettes : subjectif et objectif.

17. **Le corpus TREC** (Li et Roth, 2002) est un jeu de données pour la tâche de classification composé de six types de questions. Les questions peuvent porter sur une personne, un lieu ou encore une information, etc.

18. Le corpus utilisé est le *Wikipedia Toxic Comments* provenant de Kaggle. Les étiquettes sont limitées à deux : toxique et non toxique.

19. XLNET (Yang et al., 2019) XLNet est un modèle de pré-entrainé autorégressif. Un modèle autorégressif est simplement un modèle de type feed-forward, qui prédit le futur mot à partir d'un ensemble de mots donnés dans un contexte.

Rastogi et al. (2020) ont recours à un SVM, un LR, et un LSTM bidirectionnelle pour évaluer la méthode EDA par rapport à la rétrotraduction. Ils observent que les deux techniques d'amplification des données montrent une amélioration par rapport à la base sans amplification pour la tâche de classification de commentaires avec une amélioration moyenne du score F1 de 3% pour EDA et de 2,3% pour la rétrotraduction pour toutes les combinaisons (méthode et modèle). Les résultats obtenus montrent que l'amplification des données peut améliorer les performances des classifieurs. Le tableau 6.2 présente les avantages et les inconvénients de la méthode d'injection de bruit.

6.2.1.3 Substitution

La méthode de substitution a été explorée dans les travaux de Wei et Zou (2019) qui a proposé de remplacer les mots choisis aléatoirement par leur synonyme pour différentes tâches (voir section 6.2.1.2). Cette méthode est communément appelée *substitution lexicale par synonymes* et est détaillée à la section 6.2.2. Néanmoins, certains travaux ont proposé de remplacer les mots soit par leurs versions mal orthographiées à partir d'une base de données préalablement définie Belinkov et Bisk (2017) soit par "_" (Xie et al., 2017) pour rajouter du bruit dans le texte. Cette technique a été appliquée pour une tâche de traduction. Xie et al. (2019) proposent de faire usage de méthodes d'amplification de données pour de l'apprentissage semi-supervisé de texte. Ils proposent une méthode nommée *Unsupervised data augmentation* (UDA) regroupant des méthodes d'amplification robustes de données pour des tâches de vision par ordinateur et des tâches de classification de textes (analyse de sentiment et classification thématique). S'agissant du texte, ils proposent deux méthodes : la rétrotraduction et le remplacement de mots. Ils choisissent de remplacer les mots ayant un faible score tfidf par d'autres mots aléatoires provenant du vocabulaire du texte. Ils conservent ceux ayant une valeur tfidf élevée. UDA a permis ainsi de réduire l'erreur sur une tâche d'analyse de sentiment de 6.50 à 4.20 pour le corpus de commentaires IMdb [20] avec BERT.

20. https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

Synthèse des méthodes d'injection de bruit
Avantages : • L'injection de bruit au niveau du caractère et du mot, si elle est faible, n'altère pas l'étiquette finale du texte d'origine (Coulombe, 2020). • L'injection de bruit est adaptée aux tâches d'analyse de sentiment d'avis/-commentaires. • La technique de suppression comme d'insertion peut permettre aux modèles d'apprendre différentes représentations des textes. **Inconvénients :** • L'estimation du bruit à générer est une tâche assez complexe. En effet, (Coulombe, 2020) souligne que l'injection de bruit fort est une méthode qui altère le sens du texte d'origine et produit des mots qui s'éloignent de la réalité. • La suppression aléatoire peut supprimer des informations cruciales pour la phrase et potentiellement changer l'étiquette de la phrase nouvellement générée.

TABLE 6.2 – Présentation des avantages et des inconvénients de l'injection de bruit.

6.2.2 Approches par substitution lexicale

La substitution lexicale est une méthode très populaire pour l'amplification des données textuelles. Les approches par substitution lexicale sont des méthodes qui consistent à remplacer un mot ou des mots dans un texte par des mots similaires. La grande majorité des travaux remplacent les mots dans le texte original par leurs synonymes afin de préserver le sens sémantique du texte initial et de l'étiquette. Pour cela, ils ont recours soit à une grande base de données lexicale, soit à des plongements de mots ou encore à un modèle de langue.

6.2.2.1 Substitution lexicale par l'emploi d'un dictionnaire

Beaucoup de travaux (Bayer et al., 2021) que nous détaillons dans la suite se sont penchés sur l'emploi de la substitution lexicale pour l'amplification de données textuelles sur une tâche bien spécifique. Comme évoqué en introduction de cette section, substituer consiste à remplacer un élément, un mot, par un équivalent sémantique. Ainsi, les synonymes présents ou compilés dans un dictionnaire ou

un thesaurus ont été majoritairement utilisés comme mots de substitution pour les mots originaux. WordNet (Strapparava et al., 2004) reste la base la plus citée dans les travaux en substitution lexicale pour les corpus en langue anglaise.

L'amplification des données par substitution lexicale a été appliquée pour diverses tâches et différentes méthodes de sélection des termes ont été proposées.

- La **reconnaissance des expressions temporelles** : Kolomiyets et al. (2011) choisissent de ne remplacer que les mots de tête des expressions temporelles dans l'optique de préserver la sémantique des phrases. Néanmoins, pour la tâche de reconnaissance des expressions temporelles pour laquelle ils adoptent cette méthode, ils ne constatent aucune amélioration. La méthode est appliquée sur un corpus de documents d'actualités (Reuters) et des données de Wikipédia ;
- La **classification thématique** : Zhang et al. (2015) ont été les premiers à utiliser un thesaurus dérivé de WordNet pour amplifier leurs données. L'amplification pour cette tâche est réalisée sur 8 jeux de données à grande échelle dont deux corpus d'articles de presse en entraînant un réseau de neurone convolutionnel. Les autres corpus concernent la tâche d'analyse de sentiment. Le thésaurus est obtenu à partir du composant *Mytheas* utilisé dans LibreOffice project. Pour chaque phrase de leur corpus, ils extraient tous les mots remplaçables et choisissent aléatoirement r d'entre eux à remplacer. Néanmoins, ils ne mentionnent pas l'impact de cette technique sur la performance des modèles développés.
- L'**analyse de sentiment** (Zhang et al., 2015; Wei et Zou, 2019; Coulombe, 2020; Xie et al., 2017; Marivate et Sefara, 2020) pour des textes commentaires/avis/critiques. Une méthode très populaire en amplification des données est l'EDA (Easy data amplification). EDA est un ensemble 4 techniques d'amplification des données simples proposé par Wei et Zou (2019) afin d'améliorer la performance sur la tâche cette classification de textes. Une de ces techniques repose sur le remplacement de mot par leurs synonymes en utilisant WordNet. Ils choisissent de manière aléatoire n mots pour chaque phrase qui ne sont pas des mots outils (stopwords) et remplacent chacun de ces mots par un de ses synonymes. Par ailleurs, le nombre de mots à remplacer varie en fonction de la longueur de la phrase. Les expérimentations montrent que de simples techniques d'amplification de données textuelles boostent les performances sur une tâche de classification pour les deux modèles. Ils obtiennent une moyenne d'amélioration de l'exactitude de +0.8% par rapport à la base de référence sans amplification pour l'ensemble des modèles (CNN et RNN) entraînés sur l'ensemble du corpus d'apprentissage.

Coulombe (2020) utilise un algorithme de substitution lexicale qui récupère dans WordNet l'ensemble des synonymes d'un mot. Étant donné que chaque entrée de WordNet est associé à un ensemble de synset et que chaque synset correspond à

un sens particulier de l'entrée, il définit une fonction qui utilise des informations contextuelles (les définitions, les exemples accompagnant chaque synset) en calculant un score de similarité entre ces informations et le contexte de mot. Finalement, l'algorithme choisit le synset qui est le plus similaire [21] au contexte du mot. Une dernière étape consiste à vérifier les synonymes retournés avec une liste d'antonymes. En outre, il propose également d'utiliser des hyperonymes pour remplacer les mots initiaux des textes. Il se limite au remplacement des adverbes, adjectifs et les noms tandis que Marivate et Sefara (2020) remplacent uniquement les verbes et les adjectifs. Cette méthode d'amplification génère un gain de performance de +0.5% pour le modèle Xgboost et +4.92% pour le perceptron multicouche par rapport à la base de référence sans amplification de données.

• La **génération de texte** : Une des quatre méthodes d'amplification des données proposée par Feng et al. (2020) consiste à remplacer aléatoirement les mots clés par leurs synonymes, hyponymes et hyperonymes issus de la base lexicale WordNet. Ces derniers sont également choisis aléatoirement. Pour sélectionner les mots clés, ils utilisent un algorithme d'extraction de mots clés RAKE (Rose et al., 2010). Ils remplacent jusqu'à 3 mots clés par commentaire avec des mots ayant la même classe grammaticale que le mot à remplacer. Ils observent que la méthode de substitution par hyperonymes est la plus efficace pour la tâche de génération.

• La **classification de relations** : Giridhara et al. (2019) se limitent au replacement des noms, des adjectifs et des adverbes pour des textes en langue anglaise. Ils utilisent le corpus KBP37 (Zhang et Wang, 2015) contenant des annotations en type de relations (per :origin, per :spouse, etc) et d'entités (nom de personne, organisations, villes, etc) et le corpus SemEval2010 [22]. Ils implémentent deux architectures : un CNN et un LSTM bidirectionnel. Ils observent que les scores F1 pour la base de référence et les scores de F1 obtenus à partir de modèles entraînés sur des données amplifiées varient d'environ +2%. En outre, une augmentation des scores F1 est également observée pour tous les modèles d'apprentissage profond sur les deux ensembles de données utilisés lors de l'apprentissage sur des données amplifiées générées par la substitution par synonymes sur 75% des échantillons d'entraînement.

Les travaux énumérés dans cette section ont montré que la substitution lexicale à partir d'une base lexicale telle que WordNet a été appliquée pour diverses tâches allant de l'analyse de sentiment à la classification de relations. L'amélioration des modèles pour ces différentes tâches est notable. Le principe général consiste à rem-

21. Utilisation de la similarité cosinus.

22. Un corpus annoté manuellement et généralement accepté pour le benchmark des tâches de classification de relations.

placer de manière aléatoire ou selon une distribution les mots clés ou les mots appartenant à des catégories grammaticales précises par leurs synonymes. Certains travaux ont proposé d'utiliser les hyperonymes et même les hyponymes comme dans Feng et al. (2020). Tout comme l'injection de bruit, la substitution est une technique qui peut être utilisée pour amplifier les données. Le tableau 6.3 présente les inconvénients et les avantages de la substitution par l'emploi d'un dictionnaire.

Synthèse des méthodes de substitution par l'emploi d'un dictionnaire
Avantages :
• Facile à utiliser et à implémenter.
Inconvénients :
• Remplacer trop de mots peut affecter le sens des phrases.
• La substitution ne permet pas de produire des phrases syntaxiquement différentes des textes originaux.

TABLE 6.3 – Présentation des avantages et des inconvénients de la substitution lexicale par l'emploi d'un dictionnaire.

6.2.2.2 Substitution lexicale par l'emploi des vecteurs de mots

La substitution lexicale basée sur les vecteurs de mots consiste à utiliser des vecteurs de mots pré-entraînés tels que Word2vec (Mikolov et al., 2013a), Glove (Pennington et al., 2014) ou encore Fastext (Joulin et al., 2016) afin de récupérer dans leur espace vectoriel, les mots similaires ou les plus proches à substituer aux mots d'origine. Cette approche nécessite soit des modèles de mots pré-entraînés pour la langue en question, soit suffisamment de données de la tâche visée pour pouvoir construire des modèles pré-entraînés.

• L' **analyse de sentiment** : Wang et Yang (2015) montrent que l'utilisation de plongements de mots (Word2vec) comme méthode d'amplification de données améliore significativement les performances d'un classifieur statistique sur une tâche de classification de tweets gênants en langue anglaise [23]. Ils utilisent un corpus de tweets qu'ils amplifient en générant de nouveaux exemples pour chaque tweet. Ils

23. La tâche consistait à catégoriser des tweets gênants selon 60 classes comme l'irrespect, le caractère sexuel du tweet, le caractère arrogant, l'hypocrisie, etc.

recherchent le mot *w* le plus proche pour un terme choisi dans un tweet en utilisant la similarité cosinus entre le terme et le mot ciblé. Pour chaque mot dans un tweet, ils interrogent la ressource de plongement de mots et le remplace avec le mot le plus proche dans l'espace de plongement. Par exemple, la phrase *Being late is terrible* devient *Be behind are bad* après la recherche des mots voisins les plus proches pour chaque token de la phrase. Les résultats montrent que l'utilisation de plongements pour amplifier les données d'entraînement améliore la F1 score de +3,8% par rapport à la base de référence sans amplification de données (AD).

Coulombe (2020) utilise l'algorithme AdaGram (Bartunov et al., 2016) pour amplifier leurs données. AdaGram est algorithme qui produit des vecteurs-mots répartis par sens et peut ainsi gérer la polysémie des mots. De plus, la liste de vecteurs-mots retournés par AdaGram est accompagnée d'un coefficient de similarité avec le mot choisi. La technique qu'ils proposent se subdivise en trois étapes. La première étape consiste en une analyse lexicale phrase par phrase de l'ensemble des textes. Ensuite, une liste de mots candidats pour la substitution est définie avec le mot, son étiquette morphosyntaxique, sa position dans la phrase et la phrase d'origine du mot. La deuxième étape consiste à produire pour chacun des mots candidats un vecteur-mot dans le but d'obtenir les différents synonymes « qui correspondent à un mot son contexte (Coulombe, 2020)». Une recherche des plus-proches-voisins est exécutée pour retenir uniquement les synsets qui sont proches du mot. La troisième phase consiste à produire des nouvelles phrases à partir du fichier des mots de substitution généré. Ils observent une amélioration de la mesure F1 (0.91 contre 0.89) par rapport à la base de référence sans amplification pour le perceptron multicouche sur des données de critiques de films. La substitution lexicale par l'emploi des vecteurs de mots pour l'analyse de sentiment a également été utilisée dans les travaux de (Rizos et al., 2019; Marivate et Sefara, 2020; Huong et Hoang, 2020). Le principe de l'algorithme demeure le même : remplacer de manière aléatoire le mot à substituer dans une phrase par un mot similaire en utilisant le score de similarité produit en utilisant modèles des plongements.

- La **classification de relations** : Giridhara et al. (2019) utilisent la même méthode de Wang et Yang (2015) décrite en début de cette sous-section pour amplifier leur corpus au moyen de Word2vec. Ils notent une amélioration marginale de la mesure F1 (50.86 contre 50.74) pour le corpus KBP37 sans Amplification des données pour la tâche visée avec un CNN.

- L'**évaluation automatique des questions à réponses courtes** : La technique proposée par Xie et al. (2019) (section 6.2.1.3) est également employée par Poulain et Connes (2021). Ces derniers utilisent la méthode des K plus proches voisins et les vecteurs pré-entraînés FastText pour le Français (Grave et al.) pour générer un

ensemble de mots candidats pour le mot d'origine. Les 20 mots les plus proches sémantiquement du mot d'origine sont sélectionnés et utilisés pour le remplacer. Cette méthode leur permet de passer de 67 questions et 3736 réponses initialement à 272 questions et 9153 réponses pour un corpus de questions et réponses d'étudiants. Leurs travaux n'ont pas but d'examiner l'apport de l'amplification de données sur leur corpus, mais d'évaluer l'apport des traits lexicaux (nombre de mots, nombre de mots mal orthographiés, etc).

Les travaux énumérés dans cette section ont montré que la substitution lexicale à partir de vecteurs de mots est possible. Cette méthode a été appliquée pour diverses tâches allant de l'analyse de sentiments à l'évaluation automatique des questions à réponses courtes. Le tableau 6.4 présente les inconvénients et les avantages de la substitution lexicale par l'emploi des vecteurs de mots.

Synthèse des méthodes de substitution par l'emploi des vecteurs de mots
Avantages :
• Facile à utiliser et à implémenter.
• L'estimation et les étiquettes grammaticales des mots à remplacer n'est plus limitée.
• Cette méthode peut être utilisée pour les langues qui ne disposent pas de grandes bases lexicales, mais pour lesquelles de grandes quantités de données textuelles sont disponibles pour la construction de modèles de plongements.
• L'utilisation des plongements de mots permet aux mots ayant une signification similaire d'avoir une représentation similaire.
Inconvénients :
• Remplacer trop de mots peut affecter le sens des phrases.

TABLE 6.4 – Présentation des avantages et des inconvénients de la substitution lexicale par l'emploi des vecteurs de mots.

6.2.2.3 Substitution lexicale basée sur des modèles de langues

Contrairement au remplacement basé sur les plongements de mots, les modèles de langue sont capables de représenter le langage en prédisant les prochains mots ou mots manquants en fonction du contexte.

• La **traduction automatique** : Fadaee et al. (2017) expliquent qu'ils ont utilisé un modèle de langue entraîné avec des réseaux de neurones récurrents à longue mémoire à court terme (LSTM) pour trouver les mots communs à substituer dans la langue source pour une tâche de traduction. Les mots de remplacement proviennent d'une liste de mots établie en sélectionnant les mots apparaissant peu fréquemment dans le corpus parallèle. Une fois le mot trouvé et remplacé dans la phrase source, une nouvelle phrase est générée. Puis, ils utilisent l'algorithme *fast-align* (Dyer et al., 2013) pour remplacer la traduction du mot remplacé dans la phrase source par la traduction de sa substitution dans la phrase cible. Ils considèrent deux configurations : la première concerne le remplacement d'un seul mot dans la phrase et la deuxième concerne le remplacement de plusieurs mots. Cette méthode améliore la qualité de la traduction pour un score BLEU de 2.9 par rapport à la base de référence et jusqu'à 3.2 par rapport à la rétrotraduction.

• L' **analyse de sentiment** : Kobayashi (2018) propose une nouvelle méthode d'amplification pour un corpus de critiques de films *« amplification contextuelle avec contrainte conditionnelle de l'étiquette* [24] *»* afin de proposer des mots plus variés à utiliser pour remplacer les mots d'origine d'une phrase. Au lieu d'utiliser des synonymes provenant d'une base lexicale comme WordNet pour la tâche d'amplification des données, ils utilisent les mots qui sont prédits par un modèle de langue bidirectionnel LSTM-RNN. Par exemple, le terme *actor* dans la phrase *The actors are fantastic* peut être remplacé par des mots non-synonymes tels que *characters, movies, stories, etc,* tout en maintenant le sentiment positif de la phrase d'origine. Par contre, si l'amplification est réalisée juste sur le terme *fantastic*, il y a une forte probabilité que les mots assignés soient *bad, terrible or good, entertaining* bien que certains de ces mots soient contraires à l'étiquette de la phrase. Ainsi, la génération de telles phrases pose un problème pour l'étape d'entraînement. Pour préserver l'étiquette de départ, ils modifient l'architecture du modèle de langue afin qu'il intègre les étiquettes des données initiales dans la tâche de prédiction des mots.

Ils concatènent chaque plongement du label y avec une couche cachée du réseau de neurones à propagation avant dans le modèle de langue afin que le résultat soit calculé à partir d'un mélange d'information provenant de l'étiquette et du contexte. Ils testent la combinaison de trois méthodes d'amplification (une méthode de substitution lexicale par synonymes, une d'amplification contextuelle avec et sans contrainte conditionnelle de l'étiquette) avec deux architectures typiques basées sur un LSTM-RNN et un CNN avec une fonction dropout [25] sur 6 corpus différents pour une tâche

24. En anglais : « Label conditional contextual amplification ».

25. Le dropout est une technique qui permet d'éviter le sur-apprentissage sur les données d'entraînement.

de classification. Ils reprennent les mêmes corpus utilisés dans les travaux de Wei et Zou (2019); Longpre et al. (2020); Feng et al. (2019) et Feng et al. (2020). Les tâches concernaient la détection de la polarité (2 corpus de critiques étiquetés en deux classes pour le premier et 5 classes pour le dernier), la détection de la subjectivité et la classification de questions selon 6 types avec le corpus TREC (Li et Roth, 2002). Ces questions peuvent porter sur un lieu, une personne, ou encore une information, etc.

Kobayashi (2018) déclare que la méthode d'amplification proposée améliore légèrement les performances de classification des modèles neuronaux (**78.20%** [26] pour le CNN et **77.83%** le RNN) par rapport à l'amplification basée sur les synonymes (77.50% pour le CNN et 77.40% le RNN), l'amplification contextuelle sans architecture conditionnelle de l'étiquette (78.02% pour le CNN et 77.62% le RNN) et la base de référence sans amplification (77.53% pour le CNN et 77.43% le RNN) .

Dans la même lancée, Wu et al. (2019) proposent une méthode d'amplification contextuelle avec contrainte de l'étiquette similaire à celle de Kobayashi (2018) mais reposant sur l'utilisation du modèle de langue pré-entraîné BERT. Tout comme avec Kobayashi (2018), le modèle de langue sur lequel s'appuie BERT prédit le mot masqué en se basant uniquement sur son contexte, donc le mot prédit peut être incompatible avec les étiquettes annotées des phrases originales. Afin de résoudre ce problème, ils introduisent un nouvel objectif : un modèle de langue masqué conditionnel [27]. Le modèle de langue masqué conditionnel masque aléatoirement certains des tokens d'une entrée. L'objectif est de prédire un mot correspondant avec l'étiquette du mot de la phrase en se basant à la fois sur le contexte et l'étiquette de la phrase. Ils observent que la méthode d'amplification proposée permet d'obtenir une amélioration évidente des performances pour différentes tâches (analyse de sentiment : +0,2 et +0,8 pour la classification en deux classes et 5 classes respectivement, détection de la subjectivité : +0,8) par rapport à la méthode d'amplification reposant sur l'utilisation du modèle de langue BERT.

Les travaux énumérés dans cette section ont montré que la substitution lexicale à partir d'un modèle de langues est possible et améliore les performances des modèles sur la tâche d'analyse de sentiment. Le tableau 6.5 présente les inconvénients et les avantages de la substitution lexicale par l'emploi des modèles de langues.

26. Les chiffres énoncés représentent la moyenne des exactitudes sur les 6 corpus de référence.
27. En anglais : « Conditional masked language model (C-MLM) ».

Synthèse des méthodes de substitution par l'emploi des modèles de langues
Avantages : • Ce type de méthode prend en compte le contexte et la sémantique des textes. • Certaines méthodes proposées prennent en compte l'étiquette dans la génération de nouvelles données (Wu et al., 2019).

TABLE 6.5 – Présentation des avantages et des inconvénients de la substitution lexicale par l'emploi des modèles de langues.

6.2.3 Approches par rétrotraduction

La rétrotraduction[28] ou encore traduction inversée est une tâche qui permet d'obtenir des paraphrases. Elle consiste à traduire une phrase d'une langue source en une langue cible. La phrase obtenue après traduction de la langue source dans la langue cible est ensuite retraduite dans la langue source. Cette démarche permet ainsi d'obtenir différentes possibilités ou variantes d'une même phrase. Les phrases qui sont similaires sont supprimées. Deux manières de réaliser une rétrotraduction existent dans la littérature :

— La première consiste à utiliser un service en ligne de traduction comme Google ou Amazon soit directement (Marivate et Sefara, 2020), soit via une API comme celle proposée par GoogleTranslate (Coulombe, 2020) pour générer des phrases et amplifier son corpus d'entraînement.

— La deuxième consiste à utiliser un modèle de traduction neuronale (Sennrich et al., 2015a; Edunov et al., 2018).

La rétrotraduction peut également être réalisée en utilisant plusieurs langues pour générer différentes variantes pour une même phrase. Nous pouvons ainsi traduire une phrase en langue A vers une langue unique ou vers plusieurs langues et inversement vers la langue A depuis la langue unique ou les autres langues. C'est une méthode généralement utilisée en traduction automatique pour évaluer les modèles de traduction (Li et Specia, 2019; Edunov et al., 2019), mais elle a également été utilisé en analyse de sentiment (Aroyehun et Gelbukh, 2018; Shleifer, 2019; Kruspe et al., 2018; Mercadier, 2020). S'agissant de l'analyse de sentiment, Marivate et Sefara (2020) observent une amélioration de +0.33 par rapport à la base

28. En anglais, « back-translation ».

de référence sur un corpus d'articles de journaux [29] (*AG News*). Coulombe (2020) observe également une amélioration de +5.8% par rapport à la base de référence pour un corpus de critiques de films [30].

De manière globale, la rétrotraduction permet d'améliorer les performances des modèles de classifications en terme d'exactitude sur la tâche d'analyse de sentiment. Les données variaient entre les critiques de films, les avis provenant d'Amazon, les textes médicaux, ou encore les tweets. Le tableau 6.6 présente les inconvénients et les avantages de la substitution lexicale par rétrotraduction.

Synthèse des méthodes de rétrotraduction
Avantages :
• Méthode robuste.
• Préservation de la sémantique de la phrase initiale. (Wu et al., 2019).
• Création de réelles paraphrases.
Inconvénients :
• Les traductions incorrectes pourraient altérer la performance du modèle de classification.

TABLE 6.6 – Présentation des avantages et des inconvénients de la rétrotraduction.

6.2.4 Conclusion

Dans cette section, nous avons présenté une vue générale des différentes méthodes d'amplification des données à partir de données initiales pour générer de nouvelles données proposées pour le traitement automatique du langage naturel. Ces techniques ont été majoritairement appliquées dans diverses tâches de TAL telles que l'analyse de sentiment, la traduction ou encore la recherche d'information et sur n'importe quel type de corpus (commentaires, avis, questions-réponses et, etc.). La classification ne se limitait pas qu'à la classification binaire, mais pouvaient concerner plusieurs classes (Wang et Yang, 2015). Comme énoncé en début de ce chapitre, l'amplification des données prend son origine dans le domaine de

29. Les articles ont été traduits de l'anglais vers le français ou l'allemand et retraduits vers l'anglais avec l'API de GoogleTranslate.

30. Les critiques ont été traduits de l'anglais vers le français et retraduits vers l'anglais avec l'API de GoogleTranslate.

la vision par ordinateur ou les transformations (*cropping, flipping, zooming, rotation et, etc.*) apportées au jeu de données et sont très utilisées avant comme pendant l'entraînement de modèles. Par contre, pour le TAL, ces transformations doivent être effectuées de manière judicieuse en fonction des données, des étiquettes et de la tâche à réaliser.

De manière synthétique, l'amplification des données a été généralement pratiquée pour les cas de figures suivants :

1. Dans le premier cas, lorsqu'on constate que l'on dispose d'une quantité de données insuffisante et que l'on souhaite utiliser un modèle neuronal quel que soit le type de tâche. Ce premier cas est majoritaire dans la littérature présentée.
2. Pour le deuxième cas, lorsque l'on dispose de données sensibles qui peuvent ne pas être mises à disposition (Claveau et al., 2021), on cherche ainsi des données de substitution.
3. Pour le troisième cas, il s'agissait de comparer l'apport des différentes méthodes sur une tâche précise telle que l'analyse de sentiment.

S'il est vrai que l'amplification des données (toute méthode confondue) permet de grossir son corpus de départ, elle ne garantit pas systématiquement un gain de performance sur la tâche explorée. Ceci vaut particulièrement pour la génération comme le démontre les résultats d'expériences pour les travaux de Mercadier (2020). D'autre part, comme l'a fait remarquer Feng et al. (2019), le nombre de données (commentaires/phrases) a augmenté ou encore le mixage (Coulombe, 2020) de différentes méthodes d'amplification est également une piste à exploiter pour obtenir de bien meilleurs résultats.

Dans le cadre de nos travaux, nous nous situons dans le premier cas de figure où nous disposons d'un faible jeu donné. Nous décidons de sélectionner quelques-unes des méthodes détaillées dans ce chapitre pour amplifier notre jeu de donnée et dans un second temps d'examiner l'apport de ces méthodes sur notre tâche de classification en FMC. Dans la prochaine section, nous présentons les différentes méthodes implémentées pour amplifier nos jeux de données.

6.3 Amplification des données initiales

La précédente section nous a permis de présenter un état de l'art des différentes méthodes généralement utilisées pour amplifier des données textuelles en traitement automatique du langage naturel. Ce chapitre s'intéressa à différentes méthodes simples et peu profondes que nous avons sélectionnées pour grossir notre corpus de

départ. La section 6.3.2 s'intéressera à la méthode de rétrotraduction. La section 6.3.3 présentera la méthode de substitution lexicale, la section 6.3.4 présentera la méthode d'injection.

6.3.1 Introduction

Cette section présente les méthodes d'amplification des données utilisées pour grossir premièrement notre corpus de verbatim et les données Yoomaneo. Comme évoqué au chapitre précédent, différentes méthodes d'amplification des données textuelles inspirées directement de la vision par ordinateur ont été appliquées en traitement de la langue naturelle et de l'apprentissage automatique à tout type de texte (des commentaires, des documents juridiques, des données issues de Wikipédia, des textes d'apprenants de langue, etc.) en fonction de la tâche visée. Dans le cadre de notre travail, nous avons fait le choix de nous concentrer sur quelques méthodes (trois principalement) pour générer des nouvelles données. En outre, le but n'était pas de générer des données de manière aveugle, mais de privilégier les méthodes facilement implémentables et reproductibles. Feng et al. (2020), Wei et Zou (2019), et Marivate et Sefara (2020) ont montré que les méthodes d'injection et de substitution permettait d'améliorer la performance de leurs modèles sur des tâches de classification de sentiment et pour des corpus de commentaires. Nous avons ainsi fait le choix d'utiliser ces méthodes pour amplifier nos données. Nous avons également rajouté la rétrotraduction.

Pour augmenter nos deux jeux de données (corpus de transcriptions d'entretiens et de tables rondes et données Yoomaneo), nous avons opté pour les méthodes suivantes :

1. la rétrotraduction ;
2. la substitution lexicale ;
3. l'injection de bruit.

Toutes ces méthodes sont présentées au chapitre 6. Pour augmenter nos données, nous les divisons en deux parties : le train (80%) et le test (20%) comme observé au tableau 6.7. Nous nous focalisons uniquement sur le train que nous subdivisons ensuite deux autres lots : le train (80%) et le dev (20%). Nous augmentons séparément chaque lot avec chacune des méthodes énumérées plus haut. Pour chaque corpus généré et pour chaque phrase, nous appliquons un filtre. Ainsi, si une phrase est générée plus d'une fois ou qu'elle correspond à la phrase d'origine, elle est supprimée.

	CorpusIx	CorpusY
Train	**503**	**384**
Dev	168	129
Test	168	129
Total	839 verbatims	642 commentaires

TABLE 6.7 – Répartition des corpus à amplifier en test, Transformer et train. *CorpusIx* se réfère au corpus de transcriptions et *CorpusY* au corpus Yoomaneo.

6.3.2 La rétrotraduction

La rétrotraduction ou encore *Round Trip translation ou Back-translation* est un processus au cours duquel on traduit un mot, une phrase ou un texte dans une autre langue (traduction vers l'avant), puis on retraduit la traduction obtenue dans la langue d'origine (traduction vers l'arrière) (Aiken et Park, 2010). Nous avons implémenté la rétrotraduction en utilisant le service de traduction Deepl [31] car il est basé sur des réseaux de neurones et est beaucoup plus performant que celui de Google (Macketanz et al., 2021). Nous avons utilisé toutes les langues (de l'anglais au polonais en passant par le japonais) de DeepL pour l'amplification de notre corpus initial, soit au total 25 langues [32]. Cette méthode est appliquée sur les deux corpus. La rétrotraduction est la première méthode que nous avons appliquée sur le corpus de transcription et de prises de note. Le tableau 6.8 présente les caractéristiques du jeu de données après amplification. Nous rajoutons une colonne pour comparer avec le corpus initial (transcriptions et prises de notes).

En utilisant toutes les langues présentes dans DeepL, nous augmentons ainsi jusqu'à 30 fois notre jeu initial. Cette opération est répétée pour le corpus Yoomaneo dont nous présentons le détail au chapitre suivant. De manière générale, nous avons constaté que les langues asiatiques et de l'Europe de l'Est donnait des rétrotraductions assez variées par rapport à l'original ou aux langues romaines (italien, espagnol, etc.). Nous donnons une liste d'exemple au tableau 6.9.

6.3.3 La substitution lexicale

Pour rappel, la substitution lexicale consiste à remplacer dans une phrase ou un texte, un élément par un autre élément qui est généralement un synonyme, tou-

31. https://www.deepl.com/fr/translator.

32. Bulgare, néerlandais, anglais britannique, anglais américain, finnois, français, allemand, grec, roumain, italien, japonais, chinois, lituanien, slovène, slovaque, hongrois, danois, polonais, portugais, russe, serbe, croate, espagnol, suédois et letton.

	CorpusIx initial Train	**CorpusIx augmenté par rétrotraduction**
Nombre total de verbatim/verbatim augmenté	503	11 236
Nombre total de phrases	1 623	38895
Nombre total de tokens (mots)	28 532	636 158
Nombre moyen de phrase par verbatim	3,22	3,46
	CorpusY initial Train	**CorpusY augmenté par rétrotraduction**
Nombre total de verbatim/verbatim augmenté	384	7 691
Nombre total de phrases	547	14 138
Nombre total de tokens (mots)	9 436	211 044
Nombre moyen de phase par verbatim	1,4	1,86

TABLE 6.8 – Répartition du jeu de données initial avant et après amplification pour le corpusIx et le corpusY en utilisant la rétrotraduiction.

Langues	Texte
Origine	Si on veut rajouter des prises, enfin des trous, qui peut les installer ? Est-ce que ça nécessite des capacités ? Parce que des fois les électriciens ne sont pas au niveau de la technologie. . . Est ce qu'il faut des électriciens ?
via le japonais	Si je veux ajouter un **bouchon** ou un **trou de forage**, qui peut l'installer ? Avez-vous besoin de compétences ? Parfois, les électriciens ne peuvent pas suivre le **rythme** de la technologie. ..... Ai-je besoin d'un électricien ?
Via le suédois	Si nous voulons ajouter des **bouchons**, des **trous de puits**, qui peut les installer ? Des compétences **sont-elles requises** ? Comme les électriciens ne sont parfois pas à la pointe de la technologie... Avons-nous besoin d'électriciens ?
via l'allemand	Si nous voulons ajouter des prises de courant, des **trous de puits**, qui peut les installer ? Cela nécessite-t-il des compétences ? Parce que parfois les électriciens ne sont pas à la hauteur.... Avons-nous besoin d'électriciens ?

TABLE 6.9 – Exemple d'une phrase obtenue par rétrotraduction avec le japonais, l'allemand et le suédois pour le corpusIx.

tefois certains travaux prennent en compte les hyperonymes ou encore les hyponymes (Feng et al., 2020). La majorité des travaux en substitution lexicale par synonymie concerne la langue anglaise. À cet effet, c'est la base WordNet qui est généralement utilisée. Pour le cas du français, nous avons opté pour l'utilisation de la base DBnary (Sérasset, 2012; Sérasset et Tchechmedjiev, 2014).

6.3.3.1 Présentation de DBnary

DBnary est une ressource lexicale multilingue au format RDF (Klyne, 2004) extrait du wiktionnaire [33]. Cette ressource a été collectée et assemblée par Sérasset

33. http://kaiko.getalp.org/fct/rdfdesc/%EF%BF%BD%20http://fr.wiktionary.or

en 2012 (Sérasset, 2012). Dans DBnary, les données lexicales ont été représentées à l'aide du vocabulaire LEMON [34] (McCrae et al., 2011) mais depuis juillet 2017, le vocabulaire ontolex [35] qui étend LEMON est utilisé. La première version de DBnary contenait peu de langues dont le français, l'anglais, l'allemand, etc. Au fil du temps, d'autres langues ont été ajoutées pour arriver aujourd'hui à 22 langues [36]. En plus de ces langues, DBnary inclut également des données provenant du projet DILAF [37] (Dictionnaire de langue africaine dont le bambara et autres) et des données morphologiques pour les langues françaises et allemandes. DBnary est disponible en téléchargement ou directement accessible en ligne via un point d'accès SPARQL. Le diagramme en figure 6.3 représente le modèle de base Ontolex sur lequel DBnary est construit.

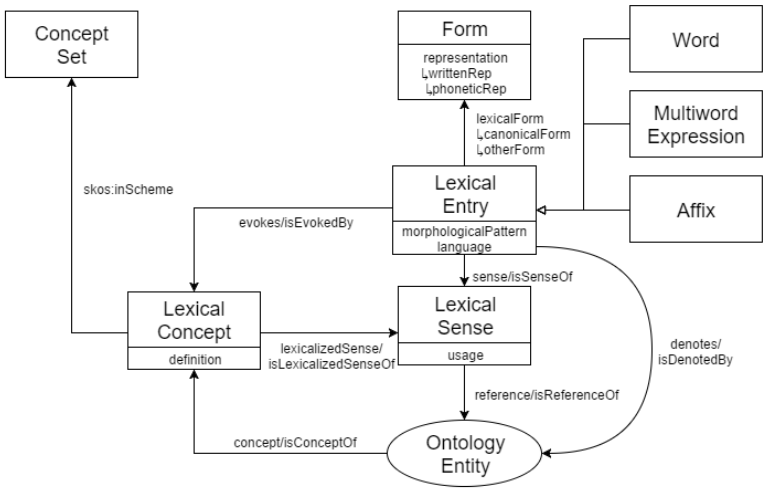


FIGURE 6.3 – Visualisation du modèle Ontolex (McCrae et al., 2011).

Les carrés représentent les classes du modèle. Les flèches à tête remplie représentent les propriétés des objets, tandis que les flèches à tête vide représentent les relations entre les sous-classes. Dans les flèches étiquetées *"X/Y"* (par exemple, *sense/isSenseOf*), *X (sense)* est le nom de la propriété de l'objet et *Y (isSenseOf)* le

34. https://lemon-model.net/.
35. https://www.w3.org/2016/05/ontolex/#purpose-of-the-model
36. Bulgare, allemand, grec (moderne), anglais, espagnol, finnois, français, indonésien, italien, japonais, kurde, latin, lituanien, malgache, néerlandais, norvégien, polonais, portugais, russe, serbe-croate, suédois et turc.
37. http://pagesperso.ls2n.fr/~enguehard-c/DiLAF/index.php.

nom de la propriété inverse [38]. Nous présentons sans détailler les classes principales du modèle Ontolex. Pour plus de détails, le lecteur peut se référer aux explications fournies sur le site du W3C [39].

1. Une des classes les plus importantes dans DBnary est la classe *Lexical Entry*. Lexical Entry ou entrée lexicale représente l'unité d'analyse du lexique qui comprend différentes formes grammaticales et un ensemble de définitions et de sens associés à toutes ces formes. Ainsi, une entrée lexicale est un mot, une expression multimots ou un affixe avec une seule partie du discours, un seul modèle morphologique, une seule étymologie et un seul ensemble de sens.
2. *Lexical Sense* ou sens lexical représente la signification lexicale d'une entrée lexicale lorsqu'elle est interprétée comme faisant référence à l'élément ontologique correspondant. Une entrée lexicale peut avoir ainsi plusieurs sens.
3. *Lexical concept* ou concept lexical représente une abstraction mentale, un concept ou une unité de pensée qui peut être lexicalisée par une collection donnée de sens.
4. *Form* ou forme représente une réalisation grammaticale d'une entrée lexicale. Ces différentes réalisations grammaticales sont représentées comme différentes formes de l'entrée lexicale. Par exemple, nous pouvons avoir comme entrée lexicale, *enfants* et ses différentes formes peuvent être *enfants* pour la forme plurielle et *enfant* pour le singulier.

	Général	Langue français
Nombre d'entrée	6 272 969	478 624
Nombre de sens	4 644 000	608 200
Nombre de Traduction	8 408 523	1 077 227

TABLE 6.10 – Détail des données existantes dans DBnary de manière générale et pour la langue considérée.

Dbnary comprend plus de 6,2 millions d'entrées et plus de 8,4 millions de traductions depuis les 22 langues vers d'autres langues dont plus de 1500 langues cibles différentes. Les entrées lexicales incluent la forme canonique des mots, les différents sens avec des exemples et des définitions, les catégories grammaticales, etc. DBnary contient également des relations lexico-sémantiques (Servan et al.,

38. Par exemple, « être le parent de » est la propriété inverse d'« être l'enfant de ».
39. Le W3C ou World Wide Web Consortium est un organisme international de standardisation à but non lucratif chargé de définir les standards techniques liés au web.

2016) comme la relation de syno/antonymie, hypo/hyperonymie, mero/holonymie et troponymie.

6.3.3.2 Substitution lexicale via la base DBnary

Pour la méthode de substitution, nous avons utilisé la base DBnary. Dans un premier temps, nous avons extrait via des requêtes SPARQL la liste des mots (adjectif, nom, adverbe, verbe) avec leurs synonymes. Le site public de DBnary propose par ailleurs des exemples de requêtes SPARQL. Nos requêtes récupéraient les entrées lexicales qui possédaient des relations sémantiques du type synonymie et de catégories grammaticales *nom, verbe, adjectif et adverbe*. Finalement, nous nous sommes heurtés à un problème, avec nos requêtes, il était impossible de récupérer toutes les entrées que nous souhaitions du fait d'une limite imposée sur le nombre d'éléments à extraire. Finalement, nous avons privilégié de télécharger directement la base de données depuis le site [40]. La base est au format Turtle [41]. Nous avons utilisé la librairie Rdflib [42] pour parser la base de données et extraire la liste de tous les mots ayant une catégorie grammaticale parmi celles mentionnées plus haut et leurs synonymes. Nous avons constitué une base de 12 120 mots avec 33 910 synonymes. Ensuite, nous avons implémenté et adapté pour le français la méthode de substitution lexicale décrite dans Wei et Zou (2019) pour l'appliquer à nos données en utilisant notre base de termes extraits. Comme eux, pour chaque phrase, nous remplaçons de manière aléatoire les mots qui sont soit des adjectifs, des noms, des verbes ou des adverbes par leurs synonymes. Par ailleurs, nous mettons en minuscule tout le texte et retirons la ponctuation avant amplification. Wei et Zou (2019) intègrent dans leur algorithme un paramètre α qui correspond au pourcentage de mots à remplacer. Ils précisent que remplacer trop de mots amène une baisse de performance pour les modèles développés, c'est pour cela qu'ils proposent d'utiliser les paramètres présentés au tableau 6.11.

Dans ce cas précis, nous choisissons de produire des phrases par substitution lexicale en fixant le taux de remplacement à 0.05 pour 16 amplifications au maximum par phrase. Cette méthode est appliquée aux corpus de verbatim et de posts. Le tableau 6.12 donne une répartition des datasets générés par cette technique. Nous

40. http://kaiko.getalp.org/about-dbnary/download/
41. Turtle (Terse RDF Triple Language) est une syntaxe d'un langage qui permet une sérialisation non-XML des modèles RDF.
42. https://github.com/RDFLib/rdflib.

Nombre d'éléments dans le corpus d'entraînement	Pourcentage de mots à remplacer	Nombres d'éléments à augmenter par entrée
500	0.05	16
2 000	0.05	8
5000	0.1	4
Plus	0.1	4

TABLE 6.11 – Paramètres recommandés par (Wei et Zou, 2019) pour l'amplification d'un corpus. En fonction du nombre d'éléments (phrase, paragraphe, commentaire entier), différents paramètres d'amplification peuvent être choisis comme le pourcentage de mots à remplacer où le nombre de phrases à augmenter.

produisons ainsi trois nouveaux jeux de données en remplaçant les termes par leurs synonymes, hyponymes et ensuite hyperonymes.

	CorpusIx augmenté avec la substitution lex. par synonymes	**CorpusIx augmenté avec la substitution lex. par hyponymes**	**CorpusIx augmenté avec la substitution lex. par hyperonymes**
Nombre total de verbatim/verbatim augmenté	6 822	2 915	1 286
Nombre total de phrases	25 191	12 574	4 462
Nombre total de tokens (mots)	548 737	269 949	94 698
Nombre moyen de phrase par verbatim	3,69	4,31	3,46
	CorpusY augmenté avec la substitution lex. par synonymes	**CorpusY augmenté avec la substitution lex. par hyponymes**	**CorpusY augmenté avec la substitution lex. par hyperonymes**
Nombre total de verbatim/verbatim augmenté	3854	1 596	785
Nombre total de phrases	6128	2 712	1 155
Nombre total de tokens (mots)	132 626	57 551	21 733
Nombre moyen de phrase par verbatim	1, 59	1,69	1,47

TABLE 6.12 – Répartition du jeu de données après amplification pour le corpusIx et le corpusY en utilisant la substitution lexicale par synonymie, par hyponymie et hyperonymie.

6.3.3.3 Combinaison des méthodes de substitution lexicale par syno-, hypo- et hyperonymes

Nous avons également proposé d'amplifier nos deux jeux de données initiaux en générant pour chaque verbatim/post de notre corpus différentes phrases amplifiées soit en utilisant la liste de syno-, d'hypo- et d'hyperonymes. Le résultat est montré au tableau 6.13.

	CorpusIx augmenté	CorpusY augmenté
Nombre total de verbatim/verbatim augmenté	3 887	2413
Nombre total de phrases	13 643	6 638
Nombre total de tokens (mots)	297 056	74 321
Nombre moyen de phrase par verbatim	3, 50	1,50

TABLE 6.13 – Répartition du jeu de données après amplification pour le corpusIx et le corpusY avec la substitution lexicale par syno-,hypo- et hyperonyme.

6.3.3.4 Substitution lexicale avec un plongement de mots

Nous avons également décidé de remplacer les mots de nos corpus initiaux avec leurs équivalents similaires provenant d'un plongement mot. Pour ce faire, nous avons utilisé le modèle FastText (voir 4.3.4) pour le français. Nous détaillons les caractéristiques des textes obtenus dans le tableau 4.3.

	CorpusIx augmenté avec la substitution lex. par plongement	CorpusY augmenté avec la substitution lex. par plongement
Nombre total de verbatim/verbatim augmenté	8 403	6 168
Nombre total de phrases	27 250	8 946
Nombre total de tokens (mots)	581 260	174 729
Nombre moyen de phrase par verbatim	3,24	1,45

TABLE 6.14 – Répartition du jeu de données après amplification avec modèle de langue pour le corpusIx et le corpusY avec la substitution par plongements de mots.

6.3.3.5 Substitution lexicale avec un modèle de langue

Nous avons utilisé la librairie nlpaug [43](Ma, 2019) qui est une librairie qui permet d'augmenter ces propres corpus d'apprentissage. Elle permet entre autres de trouver le mot le plus approprié à l'augmentation. Elle s'appuie sur l'utilisation des modèles de langue de type BERT et RoBERTa. Pour le modèle de langue français choisi, nous avons utilisé le modèle de CamemBERT BASE puisqu'il s'appuie sur RoBERTa. Nous détaillons les caractéristiques des textes obtenus dans le tableau 6.15.

Le tableau 6.16 présente des exemples de phrases générées à partir des différentes méthodes de substitution présentées dans cette section.

43. https://github.com/makcedward/nlpaug

	CorpusIx augmenté avec un modèle de langue	**CorpusY augmenté avec un modèle de langue**
Nombre total de verbatim/verbatim augmenté	8 549	6 501
Nombre total de phrases	26 519	9 151
Nombre total de tokens (mots)	486 000	159 685
Nombre moyen de phrase par verbatim	3,10	1,40

TABLE 6.15 – Répartition du jeu de données après amplification pour le corpusIx et le corpusY avec la substitution lexicale avec un modèle de langue.

	Texte
Origine	La cave à vin peut être traitée séparément, dans le sens où aujourd'hui, un consommateur peut très bien avoir... on va dire... allez, je vais aller à l'extrême : a envie d'avoir un réfrigérateur type, euh... réfrigérateur américain, parce que je vais avoir mon distributeur d'eau, mon distributeur de glaçons, ça fait classe, etc. Et à côté de ça, je veux une petite cave à vins pour avoir, comme vous l'avez noté sur une de vos diapos... pour avoir sous la main le vin à la bonne température dans ma cuisine.
Substitution par synonymes	La cave à **pinard** peut être traitée séparément , dans le sens où aujourd'hui , un consommateur peut très bien **posséder** ... on va dire ... allez , je vais aller à l' extrême : a envie d' posséder un réfrigérateur type , euh ... réfrigérateur américain , parce que je vais posséder mon distributeur d'eau , mon distributeur de glaçons , ça fait classe , etc. Et à côté de ça , je veux une petite cave à vins pour posséder , comme vous l' avez noté sur une de vos **diapositive** ... pour **posséder** sous la main le **pinard** à la bonne température dans ma cuisine.
Substitution par hyponymes	La cave à **picrate** peut être traitée séparément , dans le sens où aujourd'hui , un consommateur peut très bien avoir ... on va dire ... allez , je vais aller à l' **suraigu** : a envie d' avoir un réfrigérateur type , euh ... réfrigérateur américain , parce que je vais avoir mon distributeur d' **eau de mer** , mon distributeur de glaçons , ça fait classe , etc. Et à côté de ça , je veux une petite cave à **blanc** pour avoir , comme vous l' avez noté sur une de vos diapos ... pour avoir sous la main le **picrate** à la bonne température dans ma cuisine .
Substitution par hyperonymes	La cave à **picole** peut être traitée séparément , dans le sens où aujourd'hui , un consommateur peut très bien avoir ... on va dire ... allez , je vais aller à l' extrême : a péché capital d'avoir un réfrigérateur type , euh ... réfrigérateur américain , parce que je vais avoir mon distributeur d' **liquide** , mon distributeur de glaçons , ça fait classe , etc. Et à côté de ça , je veux une petite cave à **alcool** pour avoir , comme vous l' avez noté sur une de vos diapos ... pour avoir sous la main le **picole** à la bonne température dans ma cuisine .
Substitution par plongement de mots	La **brasserie** à vin peut être traitée séparément , dans le **questionnement** où aujourd'hui , un consommateur peut très bien avoir ... on va rétorquer ... allez , je vais aller à l'extrême : a envie d' avoir un réfrigérateur type , euh ... réfrigérateur américain , parce que je vais avoir Mon distributeur d' eau , Mon distributeur de glaçons , ça fait classe , etc. Eh à côté de ça , je veux une petite **brasserie** à vins pour avoir ,comme vous l' avez noté sur une de vos diapos ... pour avoir sous la main le vin à la bonne température dans ma cuisine .
Substitution à partir d'un modèle de langue	La cave à vin peut être traitée séparément, dans le sens où aujourd'hui, **LE** consommateur peut très bien avoir... on va..... allez, je vais aller à l **ultraextrême** : a envie d'avoir son réfrigérateur type, euh... réfrigérateur américain, parce que je vais avoir ma distributeur d'eau, mon **bain** de glaçons, ça **devenait** classe, etc. Et à côté de ça, je veux une petite cave à vins pour avoir, comme vous le'avez noté sur une de vos précédentes diapos... pour avoir sous la main le **rouge** à la bonne température dans ma cuisine.

TABLE 6.16 – Exemple d'une phrase généré en utilisant différentes méthodes de substitution.

6.3.4 Injection du bruit

En ce qui concerne l'injection de bruit, nous l'appliquons au niveau du caractère et du mot. Par contre, nous ne l'appliquons pas au niveau de la phrase, tout simplement parce que nous disposons d'un corpus avec une grande majorité de commentaires/verbatim qui sont constitués d'une seule phrase. L'intérêt de procéder à des changements de positions de phrases ou suppression auraient tendance à changer la sémantique et finalement l'étiquette de la phrase.

6.3.4.1 Injection du bruit au niveau du caractère

En ce qui concerne l'injection du bruit au niveau caractère, nous suivons la méthode de (Feng et al., 2020). Pour chaque phrase donnée, nous supprimons, rajoutons et échangeons des caractères de manière aléatoire en fonction d'un taux de pourcentage de remplacement. Nous produisons ainsi pour une phrase, des variations de bruit de 0.05, 0.1 et 0.15 comme observé sur le tableau 6.17 pour la phrase « *Plutôt sceptique par rapport au lien entre utilisation du produit et remise sur mutuelle* ». Cette technique est appliquée sur nos deux jeux de données.

Taux	Exemples générés
0.05	Plutôt sceptique par rapport au lien entre **utilsation** du **pvroduit** et remise sur mutuelle
0.10	**Pluktôt** sceptique par **raport** au lien entre **utilisaiton** du **prodiut** et **renmise** sur mutuelle
0.15	Plutôt **sceyptwiuqe** par rapport au lien entre **uttilisation dku poroudit** et remise sur mutuelle

TABLE 6.17 – Exemples de variation de bruit pour une phrase unique.

Le tableau 6.18 présente le jeu de donnée augmenté pour le corpusIx et corpusY respectivement.

	CorpusIx augmenté	**CorpusY augmenté**
Nombre total de verbatim/verbatim augmenté	1 981	1 465
Nombre total de phrases	6 442	2 098
Nombre total de tokens (mots)	113 296	36 772
Nombre moyen de phrase par verbatim	3,25	1,43

TABLE 6.18 – Répartition du jeu de données après amplification pour le corpusIx et le corpusY en utilisant l'injection du bruit au niveau du caractère.

6.3.4.2 Injection du bruit au niveau du mot

Cette technique permet d'échanger la position de deux mots, d'insérer un synonyme aléatoire d'un mot lexical à une place aléatoire et supprimer un mot dans

une phrase. Un exemple est donné au tableau 6.19. Pour compenser la différence de longueur entre les phrases dans leurs corpus, Wei et Zou (2019) font varier le pourcentage de mots n pour la technique de suppression, d'insertion et d'échange en fonction de la longueur de la phrase l avec la formule n=αl, où α est un paramètre qui indique le pourcentage de mots à modifier dans la phrase. Pour augmenter nos jeux de données, nous fixons ce taux à 0.05 pour 16 amplifications. Pour chaque phrase, nous supprimons aléatoirement un mot, remplaçons aléatoirement un mot ou échangeons aléatoirement la position de deux mots dans le verbatim. Cette technique est appliquée sur les deux jeux de données. Le tableau 6.20 présente le jeu de donnée initial et augmenté pour le corpusIx et corpusY respectivement.

	Texte
Origine	une fonction permettant de mesurer son rythme cardiaque pendant la pratique sportive serait bien je pense
Echange	sportive fonction permettant de mesurer son rythme cardiaque pendant la pratique une serait bien je pense
Insertion	une fonction permettant de mesurer son cardiaque pendant la pratique serait bien je pense
Suppression	une fonction permettant de mesurer son rythme cardiaque pendant la pratique sportive serait bien pense

TABLE 6.19 – Exemple d'une phrase généré par injection de bruit au niveau du mot (InjM) .

	CorpusIx augmenté	**CorpusY augmenté**
Nombre total de verbatim/verbatim augmenté	8 024	5 481
Nombre total de phrases	26 897	8 204
Nombre total de tokens (mots)	572 787	165 240
Nombre moyen de phrase par verbatim	3,35	1,49

TABLE 6.20 – Répartition du jeu de données après amplification pour le corpusIx et le corpusY en utilisant l'injection du bruit au niveau du mot.

6.3.4.3 Combinaison des méthodes d'injection de bruit

Nous avons également proposé de combiner les méthodes d'injection de bruit pour générer de nouveaux corpus. Chaque méthode est appliquée séparément sur le corpus initial et ensuite, nous additions les données amplifiées obtenues pour former un nouveau jeu de donnée. Le texte généré est décrit au tableau 6.21.

6.3.5 Addition des jeux de données pour l'ensemble des méthodes

Nous avons également décidé d'additionner les jeux de données générés à l'aide de chacune des méthodes à l'exception des méthodes de combinaison. Dans le ta-

	CorpusIx augmenté	CorpusY augmenté
Nombre total de verbatim/verbatim augmenté	9 483	6 605
Nombre total de phrases	31 664	9 783
Nombre total de tokens (mots)	657 701	193 598
Nombre moyen de phrase par verbatim	3,33	1,48

TABLE 6.21 – Répartition du jeu de données après amplification pour le corpusIx et le corpusY en combinant les méthodes d'injection de bruit.

bleau 6.22, nous détaillons la composition de chacun des jeux de données obtenus sur le train.

	CorpusIx	CorpusY
Nombre total de verbatim/verbatim augmenté	56 023	37 322
Nombre total de phrases	194 261	58 961
Nombre total de tokens (mots)	3 907 673	1 103 351
Nombre moyen de phrase par verbatim	3, 46	1,57

TABLE 6.22 – Répartition du jeu de données après addition des corpus amplifiés pour chaque corpus initial en excluant les méthodes de combinaison.

6.4 Conclusion

Dans ce chapitre, nous avons exploré de simples méthodes d'amplification de données textuelles pour notre jeu de données. Nous avons également fait le choix de combiner certaines dont les méthodes d'injection et les méthodes de substitution via la ressource Dbnary pour générer d'autres corpus. Dans le chapitre suivant, nous décidons de les utiliser dans une tâche de classification et d'explorer leur influence sur le développement de divers modèles. En termes de bilan, nous constatons que les simples méthodes d'amplification permettent générer des nouveaux corpus jusqu'à 25 fois supérieur au corpus initial. En ce qui concerne les modèles utilisés, nous avons fait le choix d'utiliser des architectures de type transformer (variants français de BERT et la version multilingue de ce dernier) que nous avons décrit au chapitre 4. En effet, ces derniers ont considérablement facilité les tâches en TAL et ont amélioré les performances dans de nombreuses tâches. En outre, ces dernières utilisent le mécanisme d'attention qui permet aux modèles d'avoir une compréhension globale de la phrase. Nous présentons dans le chapitre suivant les résultats obtenus sur les corpus générés pour la tâche de classification en FMC.

Chapitre 7

Classification FMC et amplification des données

L'objectif original de cette thèse était de proposer un modèle de classification de textes (verbatim et posts) utilisant des réseaux de neurone de type Transformer. Ce modèle devra être ensuite imbriqué dans une plateforme d'analyse pour un usage industriel par la société Ixiade. Étant donné une transcription d'entretien, le modèle devra être capable de classifier automatiquement les phrases présentes selon les trois classes principales. Pour Ixiade, l'automatisation de cette tâche réduirait le temps d'analyse des contenus d'étude, mais également le nombre de ressources affectées à cette tâche. In fine, Ixiade proposerait à ses clients des prestations moins coûteuses et riches en termes de données à traiter. Dans les précédents chapitres, nous avons proposé d'amplifier notre jeu de données à l'aide de diverses techniques d'amplification de données. Cette tâche a été réalisée à la fois sur le corpus de transcriptions et celui de posts issus de Yoomaneo. L'objectif poursuivi était double :

1. disposer de suffisamment de données d'apprentissage ;
2. pouvoir comparer l'effet de chacune des méthodes retenues sur notre tâche de classification.

Au total, onze corpus ont été ainsi générés pour chaque jeu de données (22 corpus d'apprentissage au total). Chaque corpus a été ensuite utilisé pour développer un modèle de détection de FMC. Pour le développement des modèles, nous avons opté pour des modèles pré-entrainés pour le français tels que FlauBERT, CamemBERT et mBERT (modèle multilingue). Nous les avons comparés et avons examiné l'apport des différentes méthodes d'amplification pour chacun d'eux. La disponibilité de ces modèles étant très récente au début de nos travaux de recherche, notre étude est la première à porter sur l'utilisation des Transformer sur des transcrip-

tions d'entretiens et des posts pour de la classification multi-classes dans le cadre de l'évaluation de l'acceptabilité d'une innovation. Dans un premier temps, seuls les corpus provenant des transcriptions ont été utilisés. Dans un second temps, les données provenant de Yoomaneo ont été utilisées. Les modèles générés à partir de l'apprentissage sur les différents corpus amplifiés ont été ensuite évalués sur des données de tests (transcriptions ou posts).

7.1 Protocole expérimental

Dans cette sous-partie, nous décrivons le protocole d'expérimentation mis en œuvre, les méthodes et les outils utilisés.

7.1.1 Modèles et architectures

Pour entraîner les corpus générés pour chaque type de données (transcriptions comme posts), nous avons utilisé une méthode d'apprentissage fine pour entraîner des réseaux de neurones de type Transformer. Chaque corpus est utilisé pour développer un modèle de classification en fonction d'un type de modèle Transformer. Pour chaque modèle sélectionné, nous utilisons toutes les architectures disponibles. Les différentes architectures utilisées sont présentées aux tableaux 7.1 à 7.3. Nous précisons également les données sur lesquelles elles ont été pré-entraînées et le nombre de paramètres. Nous rajoutons une dernière architecture « *IxBERT base cased* » basée sur FlauBERT que nous avons directement pré-entraîné sur les transcriptions originelles.

Modèle	Paramètres	Architecture	Corpus d'apprentissage
FlauBERT-base cased	138M	Base	24 sous-corpus de types divers (71 GB)
FlauBERT-base uncased	137M	Base	24 sous-corpus de types divers (71 GB)
IxBERT-base cased	138M	Base	Corpus de transcriptions (2 Mo)
FlauBERT-Large	373M	Large	24 sous-corpus de types divers (71 GB)

TABLE 7.1 – Modèles pré-entrainés pour FlauBERT (Le et al., 2020).

Pour CamemBERT, les premiers modèles mis à disposition de la communauté scientifique ont été entraînés sur 138 GB et 135 GB. Plus tard, d'autres modèles plus petits ont été proposés. Ces derniers ont été entraînés sur seulement 4 GB d'échantillons sélectionnés de manière aléatoire à partir des corpus OSCAR, CC-

Net et Wikipédia.

Modèle	Paramètres	Architecture	Corpus d'apprentissage
CamemBERT-base	110M	Base	corpus OSCAR (138 GB)
CamemBERT-large	335M	Large	corpus CCNet (135 GB)
CamemBERT-base-ccnet	110M	Base	corpus CCNet (135 GB)
CamemBERT-base-wikipedia-4gb	110M	Base	Wikipedia (4 GB)
CamemBERT-base-oscar-4gb	110M	Base	sous-ensemble d'OSCAR (4 GB)
CamemBERT-base-ccnet-4gb	110M	Base	sous-ensemble de CCNet (4 GB)

TABLE 7.2 – Modèles pré-entrainés pour CamemBERT (Martin et al., 2020a,b).

Enfin, nous testons également les architectures du modèle multilingue de BERT.

Modèle	Paramètres	Architecture	Corpus d'apprentissage
Multilingual BERT (mBERT-base cased)	110M	Base	Wikipedia (104 langues)
Multilingual BERT (mBERT-base uncased)	110M	Base	Wikipedia (102 langues)

TABLE 7.3 – Modèles pré-entrainés pour mBERT (Devlin et al., 2019).

7.1.2 Outils

Pour l'écriture et l'entraînement des modèles, nous avons utilisé le langage de programmation Python. Les architectures de type Transformer ont été codées en utilisant la librairie PyTorch. Tout ce qui concerne le prétraitement a nécessité l'utilisation de la librairie Scikit-learn et de Spacy-stanza. D'autres librairies telles que Pandas, SciPy et NumPy ont été également utilisées. Au niveau des machines de calcul, nous avons dans un premier temps utilisé les calculateurs du LIG ensuite, nous nous sommes tournés vers Jean Zay [1] pour les modèles plus larges. Un GPU TESLA V100 a été utilisé pour les calculs.

7.1.3 Phase d'apprentissage et d'évaluation

Les données ont été toutes nettoyées à l'aide d'expressions régulières. Le format d'encodage des données choisi était en UTF-8 (Universal Character Set Transfor-

1. Jean Zay est un supercalculateur de type HPE SGI 8600 installé à l'IDRIS (centre national de calcul du CNRS). Il est équipé à la fois d'unités centrales de traitement (CPU) et d'unités de traitement graphique (GPU).

mation Format - 8).

Modèles. Tous les modèles sélectionnés pour la comparaison sont disponibles gratuitement en open source. En ce qui concerne l'affinage des modèles, nous avons suivi le processus préconisé par Devlin et al. (2018) et suivi par Le et al. (2020). L'entrée de tous les modèles est une seule entrée. Les dimensions des entrées des couches linéaires sont respectivement égales à la taille du Transformer. Les têtes de classification ajoutées au-dessus de chaque modèle pré-entrainé sont les mêmes que celles présentées dans Liu et al. (2019); Le et al. (2020) et Devlin et al. (2018).

Paramètres. En ce qui concerne les hyperparamètres, ils sont tous fixés au moment de l'apprentissage avec une taille de lot à 8 pour tous les modèles. Le nombre d'époques a été fixé à 5 et le taux d'apprentissage à 5e-5 pour la première époque puis diminue linéairement. Nous nous retrouvons ainsi avec 12 architectures pour 24 corpus d'apprentissage (en incluant les deux corpus initiaux). Chaque architecture est utilisée pour générer un modèle de classification en fonction d'une méthode d'amplification et d'un type de corpus (transcriptions ou posts).

Phase d'amplification. Toutes les méthodes d'amplification des données décrites au chapitre 6 sont utilisées pour amplifier nos corpus d'apprentissage initiaux.

Phase d'apprentissage supervisé. 80% des données ont été dédiées à l'apprentissage supervisé des modèles. Ces données sont réparties en un échantillon d'entraînement et de validation. Seul l'échantillon d'entraînement a été amplifié. L'apprentissage pour chaque combinaison (modèle + méthode d'amplification) est répété dix fois en faisant varier la graine aléatoire.

Phase de test. Le meilleur modèle sur le corpus de validation est ensuite évalué sur le corpus de test. Les résultats sont évalués en fonction de la méthode d'amplification utilisée et de l'architecture utilisée. La base de référence choisie est le modèle sans amplification affiné sur le corpus initial.

Métriques. Les métriques utilisées pour mesurer la performance de chaque modèle et évaluer l'apport de l'amplification sont la F1-macro [2] et l'exactitude qui est égale à la micro-précision.

7.2 Expériences et résultats : Données de verbatim

Dans cette section, nous présentons les résultats obtenus avec les meilleurs modèles développés pour chaque couple de combinaison (architecture du modèle +

2. C'est une métrique utilisée pour évaluer la performance d'un modèle de classification à plus de deux classes. Il est également recommandé pour les données présentant un déséquilibre au niveau des classes.

méthode d'amplification [3]) à partir des corpus de verbatim issus des transcriptions, et ce, pour l'ensemble des expériences. Les résultats présentés dans la suite de cette section ont été obtenus en évaluant les modèles sur le corpusIx (168 verbatims) et le corpusY (129 posts). Ils sont donnés en utilisant les métriques de l'exactitude et de la F1-macro.

7.2.1 Modèle FlauBERT

Nous détaillons les résultats obtenus sur les deux corpus de test avec les modèles de classification générés avec le corpus initial d'apprentissage de verbatim et ses versions amplifiées dans les tableaux 7.4 et 7.5. Les améliorations/dégradations statistiquement significatives de l'exactitude par rapport à la base de référence sont notées (+/-) dans les tableaux. Le meilleur résultat obtenu pour chaque corpus de test est illustré dans une cellule colorée en gris.

A Résultat général

En utilisant toutes les méthodes d'amplification des données avec les différentes architectures de FlauBERT (tableau 7.4 et tableau 7.5), nous avons obtenu la meilleure valeur d'exactitude de 0,714 pour 0,683 de F1-macro avec l'architecture FlauBERT BC [4] et le corpus amplifié avec l'injection de bruit au niveau du mot (InjM). Ceci correspond à une augmentation de **+0,21** pour l'exactitude et de **+0,42** pour la F1-macro par rapport à la base de référence (0,482 en exactitude et 0,217 en F1-macro). Ce résultat démontre l'importance d'avoir un volume de données assez important pour la phase d'entraînement. En testant le même modèle sur le corpusY, nous avons obtenu une valeur d'exactitude de 0,791 pour 0,703 de F1-macro par rapport à la base de référence qui se situait à 0,659 pour l'exactitude et 0,265 pour la F1-macro. Bien que nous ayons observé une amélioration du score tant au niveau de l'exactitude comme de la F1-macro, la meilleure valeur d'exactitude sur le corpusY est obtenue avec FlauBERT L et la substitution lexicale par plongement

3. Les différentes méthodes d'amplification sont la rétrotraduction (1), la substitution lexicale par synonymes (2), la substitution lexicale par hyperonymes (3), la substitution lexicale par hyponymes (4), la substitution lexicale par plongement de mots (5), la substitution à partir d'un modèle de langue (6), l'injection de bruit au niveau du mot (7), l'injection de bruit au niveau du caractère (8), la combinaison des méthodes d'injection de bruit (9), la substitution lexicale via une base lexicale par syno-, hypo- et hyperonymes(10) et l'addition de tous les jeux de données (11).

4. Dans la suite, nous utiliserons les abréviations suivantes : BC pour BASE CASED, BU pour BASE UNCASED et L pour LARGE.

(0,798 d'exactitude et 0,690 de F1-macro). Ce qui correspond à une amélioration de **+0,09** pour l'exactitude et de **+0,11** pour la F1-macro.

TAD	CorpusIx - test							
	FlauBERT BC		FlauBERT BU		IxBERT		FlauBERT L	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,482	0,217	0,500	0,267	0,536	0,482	0,589	0,538
1 - Rétrotraduction	0,667 +0,18	0,604 +0,39	0,690 +0,19	0,650 +0,38	0,554 +0,02	0,484 0,00	**0,690** **+0,10**	**0,657** **+0,12**
2 - Subs - Syn	0,589 +0,11	0,574 +0,36	0,649 +0,15	0,607 + 0,34	0,560 +0,02	0,492 +0,01	0,690 +0,10	0,641 +0,10
3 - Subs - hype	0,637 +0,15	0,549 +0,33	0,464 -0,04	0,330 +0,06	0,536 0,00	0,466 -0,02	0,649 +0,06	0,593 +0,06
4 - Subs - hypo	0,577 +0,10	0,516 +0,30	0,667 +0,17	0,620 +0,35	0,565 +0,03	0,538 +0,06	0,655 +0,07	0,592 +0,05
5 - Subs - plongement	**0,679** **+0,20**	**0,618** **+0,40**	0,548 +0,05	0,482 +0,21	0,536 0,00	0,481 0,00	0,679 +0,09	0,637 +0,10
6 - Subs - modèle de langue	0,625 +0,14	0,546 +0,33	0,685 +0,18	0,637 +0,37	0,560 +0,02	0,487 +0,01	0,589 0,00	0,491 -0,05
7 - InjM	0,595 +0,11	0,558 +0,34	**0,714** **+0,21**	**0,683** **+0,42**	0,536 0,00	0,433 -0,05	0,583 -0,01	0,411 -0,13
8 - InjL	0,673 +0,19	0,591 +0,37	0,685 +0,18	0,641 +0,37	**0,583** **+0,05**	**0,520** **+0,04**	0,625 +0,04	0,514 -0,02
9 - AllInj (7+8)	0,667 +0,18	0,624 +0,41	0,631 +0,13	0,511 +0,24	0,554 +0,02	0,504 +0,02	0,685 +0,10	0,651 +0,11
10 - AllSubs (2+3+4)	0,637 +0,15	0,564 +0,35	0,524 +0,02	0,381 +0,11	0,565 +0,03	0,480 0,00	0,685 +0,10	0,623 +0,09
11 - 1+2+3+4+5+6+7+8	0,613 +0,13	0,561 +0,34	0,667 +0,17	0,627 +0,36	0,554 +0,02	0,496 +0,01	0,673 +0,08	0,610 +0,07

TABLE 7.4 – Résultats d'évaluation et gains obtenus sur le corpusIx pour les différentes architectures de FlauBERT avec les différentes méthodes d'amplification.

B Effet de l'amplification des données : corpusIx

Nous avons obtenu une meilleure valeur d'exactitude sur notre tâche avec la substitution par plongement pour FlauBERT BC, soit 0,679 pour 0,618 de F1-macro. La technique de rétrotraduction obtient le meilleur score pour l'architecture FlauBERT L (0,690 en exactitude pour 0,657 de F1-macro). Nous avons obtenu la meilleure valeur d'exactitude à 0,583 pour 0,520 de F1-macro pour l'architecture IxBERT avec la méthode d'injection de bruit au niveau du caractère.

Si nous examinons de près chaque architecture, nous remarquons que toutes les méthodes d'amplification améliorent les performances du modèle FlauBERT BC dans une fourchette de +0,10 (substitution lexicale par hyponymes) à +0,20 (substitution par plongement). S'agissant du modèle FlauBERT L, les améliorations observées sont moins significatives (entre +0,04 à +0,10) qu'avec FlauBERT BC ou encore FlauBERT BU. Par ailleurs, nous remarquons que l'injection de bruit au

niveau du mot dégrade la performance de FlauBERT L jusqu'à -0,13 pour la F1 macro par rapport à la base de référence (0,538). La rétrotraduction, la substitution lexicale par synonymes, la combinaison des méthodes d'injection et celles de substitution lexicale par syno-, hypo- et hyperonymes sont celles qui performent le mieux avec FlauBERT L (jusqu'à +0,10 de gain obtenu).

Nous avons également constaté une amélioration de la performance pour FlauBERT BU en terme d'exactitude dans une fourchette allant de +0,02 (combinaison des méthodes de substitution) à +0,21 (injection de bruit au niveau du mot). Pour la F1-macro, la fourchette d'amélioration de l'exactitude se situe entre +0,06 à +0,42. Toutefois, la méthode de substitution lexicale par hyperonymes performe moins bien (-0,04 pour l'exactitude par rapport à la référence) que les autres méthodes lorsqu'elle est utilisée avec FlauBERT BU.

Globalement, les résultats obtenus par l'ensemble des modèles demeurent très significatifs si nous nous focalisons sur la mesure F1 avec une valeur maximale de 0,683 (score du meilleur modèle) par rapport à 0,267 pour la base de référence. Nous remarquons également que les méthodes d'amplification ont moins d'effet lorsqu'elles sont appliquées avec le modèle IxBERT. L'amélioration reste très faible entre +0,02 à +0,05. Une explication pourrait être liée au trop peu de données sur lequel le modèle de langue a été pré-entraîné. Une autre remarque à souligner est que l'ajout de plusieurs jeux de données amplifiées (méthodes 9, 10, et 11) n'améliore pas significativement les résultats de classification pour l'exactitude.

C Effet de l'amplification des données : corpusY

Les meilleurs résultats d'exactitude sont obtenus avec l'injection de bruit au niveau du caractère pour FlauBERT BC (0,767 pour 0,636 de F1-macro), l'injection de bruit au niveau du mot pour FlauBERT BU (0,791 et 0,703 de F1-macro), la substitution par plongement de mots pour FlauBERT L (0,798 pour 0,690 de F1-macro). La méthode 6 (substitution lexicale à partir d'un modèle de langue) est celle qui performe le mieux avec l'architecture *IxBERT* (0,736 d'exactitude et 0,583 de F1-macro). Nous avons également constaté une amélioration de la mesure F1 dans l'ensemble (jusqu'à +0,44 pour le meilleur modèle).

TAD	CorpusY - test							
	FlauBERT BC		FlauBERT BU		IxBERT		FlauBERT L	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,674	0,269	0,659	0,265	0,574	0,455	0,713	0,579
1 - Rétrotraduction	0,736 +0,06	0,563 +0,29	0,752 +0,09	0,641 +0,38	0,566 -0,01	0,413 -0,04	0,721 +0,01	0,608 +0,03
2 - Subs - Syn	0,581 +0,05	0,526 +0,27	0,682 +0,02	0,601 +0,34	0,620 +0,05	0,513 +0,06	0,767 +0,05	0,665 +0,09
3 - Subs - hype	0,729 0,00	0,542 +0,24	0,628 -0,03	0,263 0,00	0,651 +0,08	0,507 +0,05	0,698 -0,02	0,575 0,00
4 - Subs - hypo	0,674 +0,04	0,513 +0,31	0,752 +0,09	0,613 +0,35	0,620 +0,05	0,508 +0,05	0,752 +0,04	0,641 +0,06
5 - Subs - plongement	0,713 +0,05	0,583 +0,34	0,574 -0,09	0,435 +0,17	0,651 +0,08	0,450 0,00	**0,798** **+0,09**	**0,690** **+0,11**
6 - Subs - modèle de langue	0,721 -0,09	0,607 +0,26	0,721 +0,06	0,604 +0,34	**0,736** **+0,16**	**0,583** **+0,13**	0,752 +0,04	0,616 +0,04
7 - InjM	0,550 -0,12	0,486 +0,22	**0,791** **+0,13**	**0,703** **+0,44**	0,721 +0,15	0,517 +0,06	0,705 -0,01	0,393 -0,19
8 - InjL	**0,767** **+0,09**	**0,636** **+0,37**	0,775 +0,12	0,661 +0,40	0,550 -0,02	0,451 0,00	0,720 +0,15	0,592 +0,14
9 - AllInj (7+8)	0,729 +0,05	0,607 +0,26	0,744 +0,09	0,553 +0,29	0,612 +0,04	0,483 +0,03	0,729 +0,02	0,624 +0,05
10 - AllSubs (2+3+4)	0,721 +0,05	0,530 +0,34	0,698 +0,04	0,412 +0,15	0,690 +0,12	0,550 +0,09	0,752 +0,04	0,628 +0,05
11 - 1+2+3+4+5+6+7+8	0,721 +0,05	0,562 +0,29	0,760 +0,10	0,637 +0,37	0,667 +0,09	0,490 +0,04	0,736 +0,02	0,561 -0,02

TABLE 7.5 – Résultats d'évaluation et gains obtenus sur le corpusY pour les différentes architectures de FlauBERT avec les différentes méthodes d'amplification.

D Résultats pour les meilleures méthodes d'amplification pour les deux corpus de test

Les résultats des précédents tableaux montrent clairement que les différentes méthodes d'amplification des données avec les différentes architectures de type Transformer ont permis d'améliorer les performances (exactitude et F1) des modèles par rapport à la base de référence que ce soit pour le corpusIx comme pour le corpusY. Parmi toutes les méthodes d'amplification utilisées et évaluées sur le corpusIx, c'est l'injection au niveau du mot avec FlauBERT BU qui a permis d'obtenir le meilleur résultat sur le corpusIx. S'agissant du corpusY, c'est la substitution par plongement de mots avec FlauBERT L qui performe le mieux suivi de l'injection au niveau du mot avec FlauBERT BU. La méthode d'injection de bruit au niveau du mot est celle qui dégrade le plus la performance pour l'architecture FlauBERT L (-0,13 et -0,19 en F1 pour le corpusIx et corpusY respectivement). Les meilleurs résultats obtenus avec FlauBERT pour chacun des corpus sont détaillés dans les tableaux 7.6 et 7.7. Dans la prochaine sous-section de cette partie, nous détaillons les résultats obtenus avec le modèle pré-entraîné (CamemBERT).

TAD	CorpusIx - test		CorpusY - test	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,500	0,267	0,659	0,265
7 - InjM + FlauBERT BU	**0,714 (+0,21)**	**0,683 (+0,42)**	0,791 (+0,13)	0,703 (+0,44)

TABLE 7.6 – Meilleur résultat obtenu pour le corpusIx avec FlauBERT.

TAD	CorpusY - test		CorpusIx - test	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,713	0,579	0,589	0,538
5 - Subs -plongement + FlauBERT L	**0,798 (+0,09)**	**0,690 (+0,11)**	0,679 (+0,09)	0,637 (+0,10)

TABLE 7.7 – Meilleur résultat obtenu pour le corpusY avec FlauBERT.

7.2.2 Modèle CamemBERT

CamemBERT a été pré-entraîné sur 6 corpus de différentes tailles (voir la section 7.1.1 et 4.4.2.2). Nous choisissons de tester toutes ces architectures (6) afin d'évaluer l'impact des méthodes d'AD pour chacun des corpus de test sur notre tâche de classification. Les tableaux 7.8 à 7.11 illustrent les résultats d'évaluation des modèles pré-entraînés soit sur l'ensemble du corpus d'apprentissage, soit sur 4 GB de données en terme d'exactitude et de F1-macro pour les corpusIx et corpusY.

A Résultat général

Nous avons obtenu la meilleure valeur d'exactitude de 0,780 avec le modèle CamemBERT B (soit 0,742 en F1-macro) pré-entraîné sur le corpus CCNet de 4 GB avec la méthode de substitution lexicale par synonymes pour le corpusIx. Nous avons obtenu un gain de performance de **+0,19** pour l'exactitude et de **+0,32** en F1-macro par rapport à la base de référence. Le meilleur résultat pour l'exactitude est ainsi obtenu en utilisant le modèle CamemBERT B entraîné sur une version réduite du corpus CCNet même s'il ne surpasse pas massivement celui des modèles entraînés sur l'ensemble des corpus OSCAR (0,732 pour CamemBERT B) et CCNet (0,738 pour CamemBERT B et 0,774 pour CamemBERT L). De même, pour le corpusY, nous avons obtenu une meilleure valeur d'exactitude égale à 0,837 (soit 0,748 en F1-macro) en utilisant également CamemBERT B pré-entraîné sur le corpus CCNet de 4 GB avec la méthode *AllSubs* [5]. Nous avons obtenu un gain de performance

5. *AllSubs* correspond à la méthode de substitution lexicale via une base lexicale par syno-, hypo- et hyperonymes.

de **+0,11** pour l'exactitude et de **+0,30** en F1-macro par rapport à la base de référence. Une fois de plus, le meilleur résultat est obtenu avec CamemBERT B entraîné sur 4 GB de données. Cela démontre que la taille des données d'entraînement n'a pas eu autant d'impact sur la performance globale des différentes architectures évaluées.

B Effet de l'amplification des données : corpusIx

De manière générale, toutes les méthodes d'amplification sans exception améliorent l'exactitude et la F1-macro sur l'ensemble des résultats (tableaux 7.8 et 7.9) pour le corpusIx par rapport à la base de référence. Cette amélioration de l'exactitude correspond à une fourchette de +0,05 à +0,28 pour l'ensemble des architectures de CamemBERT.

TAD	CorpusIx - test					
	CamemBERT B 138 GB (OSCAR)		CamemBERT B 135 GB (CCNet)		CamemBERT L - 135 GB (CCNet)	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,482	0,217	0,607	0,458	0,494	0,318
1 - Rétrotraduction	0,696 +0,21	0,640 +0,42	0,732 +0,13	0,687 +0,47	0,673 +0,18	0,611 +0,39
2 - Subs - Syn	0,714 +0,23	0,663 +0,45	0,702 +0,10	0,653 +0,44	0,649 +0,15	0,581 +0,36
3 - Subs - hype	0,65 +0,17	0,595 +0,38	0,684 +0,08	0,624 +0,41	0,678 +0,18	0,632 +0,42
4 - Subs - hypo	0,649 +0,17	0,604 +0,39	0,667 +0,06	0,601 +0,38	0,685 +0,19	0,606 +0,39
5 - Subs - plongement	**0,732** **+0,25**	**0,702** **+0,49**	0,673 +0,07	0,609 +0,39	0,702 +0,21	0,665 +0,45
6 - Subs - modèle de langue	0,661 +0,18	0,611 +0,39	0,679 +0,07	0,611 +0,39	0,726 +0,23	0,688 +0,47
7 - InjM	0,714 +0,23	0,648 +0,43	0,655 +0,05	0,594 +0,38	0,756 +0,26	0,720 +0,50
8 - InjL	0,685 +0,20	0,642 +0,43	0,667 +0,06	0,621 +0,16	0,714 +0,22	0,687 +0,37
9 - AllInj (7+8)	0,696 +0,21	0,657 +0,44	**0,738** **+0,13**	**0,707** **+0,25**	**0,774** **+0,28**	**0,739** **+0,42**
10 - AllSubs (2+3+4)	0,673 +0,19	0,634 +0,42	0,702 +0,10	0,652 +0,19	0,661 +0,17	0,618 +0,30
11 - 1+2+3+4+5+6+7+8	0,613 +0,13	0,579 +0,36	0,679 +0,07	0,607 +0,15	0,714 +0,22	0,673 +0,36

TABLE 7.8 – Résultats d'évaluation et gains obtenus sur le corpusIx pour les différentes architectures de CamemBERT avec les différentes méthodes d'amplification. Les modèles utilisés sont ceux entraînés sur l'ensemble du corpus d'apprentissage (138/135 GB)..

Modèles 138/135 GB. L'utilisation du modèle CamemBERT améliore considérablement l'exactitude (jusqu'à +0,28 pour la méthode *Allinj* [6] avec CamemBERT L)

6. *Allinj* correspond à la combinaison des méthodes d'injection de bruit au niveau du caractère

pour toutes les architectures (voir tableau 7.8). La substitution par plongement de mots est celle qui permet d'obtenir le meilleur résultat pour CamemBERT B pré-entraîné sur OSCAR avec 0,732 d'exactitude. La méthode *Allinj* permet d'obtenir les meilleurs résultats pour CamemBERT B pré-entraîné sur CCNet et CamemBERT L avec 0,738 et 0,774 d'exactitude respectivement.

Modèles 4 GB. Les méthodes de rétrotraduction, de substitution lexicale par synonymes et d'injection au niveau du mot sont celles qui performent les mieux (voir tableau 7.9). Les modèles entraînés sur 4 GB d'OSCAR et de CCNet ont obtenu des performances supérieures (0,780 et 0,720 d'exactitude) à celles du modèle entraîné sur les données de Wikipédia (0,673 d'exactitude). Une des raisons à cet écart de performance est liée au fait que notre tâche de classification implique des textes dont le genre et le style sont plus éloignés des données issues Wikipédia.

TAD	CorpusIx- test					
	CamemBERT B 4 GB (CCNet)		CamemBERT B 4 GB (OSCAR)		CamemBERT B 4 GB (Wiki)	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,589	0,418	0,500	0,258	0,482	0,217
1 - Rétrotraduction	0,726 +0,14	0,687 +0,27	0,679 +0,18	0,644 +0,39	**0,673** **+0,19**	**0,610** **+0,39**
2 - Subs - Syn	**0,780** **+0,19**	**0,742** **+0,32**	0,702 +0,20	0,653 +0,40	0,631 +0,15	0,572 +0,36
3 - Subs - hype	0,702 +0,11	0,673 +0,26	0,654 +0,15	0,612 +0,35	+0,601 +0,12	+0,485 +0,27
4 - Subs - hypo	0,673 +0,08	0,618 +0,20	0,655 +0,15	0,590 +0,33	0,595 +0,11	0,436 +0,22
5 - Subs - plongement	0,708 +0,12	0,676 +0,26	0,714 +0,21	0,660 +0,40	0,560 +0,08	0,448 +0,23
6 - Subs - modèle de langue	0,667 +0,08	0,626 +0,21	0,696 +0,20	0,662 +0,40	0,583 +0,10	0,403 +0,19
7 - InjM	0,702 +0,11	0,671 +0,25	**0,720** **+0,22**	**0,700** **+0,44**	+0,631 +0,15	+0,572 +0,36
8 - InjL	0,702 +0,11	0,674 +0,26	0,637 +0,14	0,595 +0,34	0,524 +0,04	0,351 +0,13
9 - AllInj (7+8)	0,696 +0,11	0,646 +0,23	0,685 +0,18	0,654 +0,40	0,560 +0,08	0,456 + 0,24
10 - AllSubs (2+3+4)	0,685 +0,10	0,631 +0,21	0,714 +0,21	0,661 +0,40	0,619 +0,14	0,526 +0,31
11 - 1+2+3+4+5+6+7+8	0,648 +0,06	0,618 +0,20	0,696 +0,20	0,668 +0,41	0,601 +0,12	0,465 +0,25

TABLE 7.9 – Résultats d'évaluation et gains obtenus sur le corpusIx pour les différentes architectures de CamemBERT avec différentes méthodes d'amplification. Les modèles utilisés sont ceux entraînés sur une version réduite du corpus d'apprentissage (4 GB).

et du mot.

C Effet de l'amplification des données : corpusY

De manière globale, nous avons obtenu une amélioration des performances des modèles de CamemBERT pour le corpusY. Les résultats sont illustrés aux tableaux 7.10 pour les modèles pré-entraînés sur 138/135 GB et 7.11 pour les modèles pré-entrainés sur 4 GB.

Modèles 138/135 GB. La substitution lexicale à partir d'un modèle de langue (E [7] : 0,822, F1 : 0,721), l'injection de bruit au niveau du caractère (E : 0,806, F1 : 0,700 ; E : 0,829, F1 : 0,743) sont celles qui performent le mieux avec les modèles de CamemBERT pré-entraînés sur l'ensemble du corpus d'apprentissage.

TAD	CorpusY - test					
	CamemBERT B 138 GB (OSCAR)		CamemBERT B 135 GB (CCNet)		CamemBERT L - 135 GB (CCNet)	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,674	0,269	0,736	0,494	0,698	0,351
1 - Rétrotraduction	0,744 +0,07	0,634 +0,36	0,760 +0,02	0,646 +0,15	0,791 +0,09	0,659 +0,31
2 - Subs - Syn	0,760 +0,09	0,606 +0,34	0,721 -0,02	0,625 +0,13	0,775 +0,08	0,634 +0,28
3 - Subs - hype	0,589 -0,08	0,479 +0,21	0,775 +0,04	0,647 +0,15	0,767 +0,07	0,679 +0,33
4 - Subs - hypo	0,752 +0,08	0,594 +0,32	0,775 +0,04	0,612 +0,12	0,752 +0,05	0,652 +0,30
5 - Subs - plongement	0,705 +0,03	0,613 +0,34	0,698 -0,04	0,513 +0,02	0,760 +0,06	0,656 +0,31
6 - Subs - modèle de langue	**0,822** **+0,15**	**0,721** **+0,45**	0,713 -0,02	0,542 +0,05	0,744 +0,05	0,640 +0,29
7 - InjM	0,752 +0,08	0,585 +0,32	0,760 +0,02	0,600 +0,11	0,822 +0,12	0,730 +0,38
8 - InjL	0,690 +0,02	0,600 +0,33	**0,806** **+0,07**	**0,700** **+0,21**	**0,829** **+0,13**	**0,743** **+0,39**
9 - AllInj (7+8)	0,775 +0,10	0,670 +0,40	0,775 +0,04	0,664 +0,17	0,705 +0,01	0,586 +0,24
10 - AllSubs (2+3+4)	0,729 +0,05	0,564 +0,29	0,822 +0,09	0,700 +0,21	0,729 +0,03	0,551 +0,20
11 - 1+2+3+4+5+6+7+8	0,698 +0,02	0,566 +0,30	0,698 -0,04	0,548 +0,05	0,736 +0,04	0,595 +0,24

TABLE 7.10 – Résultats d'évaluation et gains obtenus sur le corpusY pour les différentes architectures de CamemBERT avec les différentes méthodes d'amplification. Les modèles utilisés sont ceux entraînés sur l'ensemble du corpus d'apprentissage (135GB).

Modèles 4 GB. La rétrotraduction (E : 0,806, F1 : 0,719), la substitution lexicale

7. E : correspond à exactitude.

par plongements de mots (E : 0,760, F1 : 0,600) et la méthode *AllSubs* (E : 0,837, F1 : 0,748) sont celles qui performent le mieux avec les modèles pré-entraînés sur 4 GB.

TAD	CorpusY - test					
	CamemBERT B 4 GB (CCNet)		CamemBERT B 4 GB (OSCAR)		CamemBERT B 4 GB (Wiki)	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,729	0,451	0,674	0,269	0,674	0,269
1 - Rétrotraduction	0,760 +0,03	0,657 +0,21	**0,806** **+0,13**	**0,719** **+0,45**	0,690 +0,02	0,484 +0,27
2 - Subs - Syn	0,798 +0,07	0,694 +0,24	0,729 +0,05	0,623 +0,35	0,713 +0,04	0,532 +0,27
3 - Subs - hype	0,783 +0,05	0,673 +0,22	0,782 +0,11	0,685 +0,42	0,72 +0,05	0,460 +0,27
4 - Subs - hypo	0,752 +0,02	0,597 +0,15	0,767 +0,09	0,606 +0,34	0,705 +0,03	0,424 +0,27
5 - Subs - plongement	0,783 +0,05	0,685 +0,23	0,729 +0,05	0,595 +0,33	**0,760** **+0,09**	**0,600** **+0,27**
6 - Subs - modèle de langue	0,791 +0,06	0,698 +0,25	0,736 +0,06	0,617 +0,35	0,713 +0,04	0,410 +0,27
7 - InjM	0,721 -0,01	0,595 +0,14	0,798 +0,12	0,707 +0,4 4	0,744 +0,07	0,629 +0,27
8 - InjL	0,798 +0,07	0,695 +0,24	0,752 +0,08	0,645 + 0,38	0,698 +0,02	0,413 +0,27
9 - AllInj (7+8)	0,783 +0,05	0,679 +0,23	0,744 +0,07	0,670 +0,40	0,705 +0,03	0,526 +0,27
10 - AllSubs (2+3+4)	**0,837** **+0,11**	**0,748** **+0,30**	0,798 +0,12	0,697 +0,43	0,698 +0,02	0,541 +0,27
11 - 1+2+3+4+5+6+7+8	0,682 -0,05	0,586 +0,14	0,736 +0,06	0,626 +0,36	0,713 +0,04	0,466 +0,27

TABLE 7.11 – Résultats d'évaluation et gains obtenus sur le corpusY de test de posts pour les différentes architectures de CamemBERT avec les différentes méthodes d'amplification. Les modèles utilisés sont ceux entraînés sur une version réduite du corpus d'apprentissage (4 GB).

D Résultats pour les meilleures méthodes d'amplification pour les deux corpus de test

Nous avons examiné l'impact des méthodes d'amplification des données pour chaque corpus dans l'objectif de sélectionner la méthode qui permet d'atteindre la meilleure valeur d'exactitude avec le modèle CamemBERT pour le corpusIx et le corpusY. Nous détaillons les meilleurs résultats obtenus pour chacun des corpus dans les tableaux 7.12 et 7.13.

TAD	CorpusIx - test		CorpusY - test	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,589	0,418	0,760	0,606
2 - Subs-Syn CamemBERT B entraîné sur 4GB (CCNet)	**0,780 (+0,19)**	**0,742 (+0,32)**	0,798 (+0,07)	0,694 (+0,24)

TABLE 7.12 – Meilleur résultat obtenu pour le corpusIx avec CamemBERT.

TAD	CorpusY - test		CorpusIx - test	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,713	0,579	0,729	0,451
7 - Allsubs CamemBERT B entraîné sur 4GB (CCNet)	**0,837 (+0,11)**	**0,748 (+0,30)**	0,685 (+0,10)	0,631 (+0,21)

TABLE 7.13 – Meilleur résultat obtenu pour le corpusY avec CamemBERT.

Les résultats obtenus avec les différentes méthodes d'amplification et le modèle CamemBERT ont montré des performances bien meilleures que celles observées avec le modèle FlauBERT et IxBERT. À titre de comparaison, la meilleure valeur obtenue avec FlauBERT est égale à 0,714 d'exactitude pour le corpusIx avec la méthode *InjM*, 0,583 avec IxBERT/*InjL* comparé au 0,780 obtenu avec le meilleur modèle de CamemBERT. et La substitution lexicale par synonymes. Il s'agit ainsi d'une amélioration nette de +0,07 et de +0,20. Dans une troisième série d'expériences, nous examinons les performances du modèle multilingue de BERT sur notre tâche et l'apport des différentes méthodes d'amplification sur le processus d'apprentissage. Dans la section suivante, nous présentons les résultats obtenus.

7.2.3 Modèle mBERT

Dans cette section, nous présentons aux tableaux 7.14 et 7.15 les résultats obtenus avec les modèles multilingues de BERT sur notre tâche de classification en

terme d'exactitude et de F1-macro pour le corpusIx et Le corpusY.

A Résultat général

Nous avons obtenu la meilleure valeur d'exactitude de 0,667 (+0,18) et 0,612 en F1-macro pour le corpusIx avec le modèle mBERT BC. Nous avons également obtenu une meilleure valeur d'exactitude de 0,783 (+0,12) et de 0,660 de F1-macro pour le corpusY avec le modèle mBERT BU par rapport à la base de référence. Ces résultats ont été obtenus avec la même méthode d'amplification : l'injection de bruit au niveau du mot.

B Effet de l'amplification des données : corpusIx

TAD	CorpusIx - test			
	mBERT BC		mBERT BU	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,488	0,231	0,548	0,403
1 - Rétrotraduction	0,607 +0,12	0,559 +0,33	0,607 +0,06	0,569 +0,17
2 - Subs - Syn	0,583 +0,10	0,511 +0,28	0,613 +0,07	0,591 +0,19
3 - Subs - hype	0,613 +0,13	0,555 +0,32	0,565 +0,02	0,481 +0,08
4 - Subs - hypo	0,583 +0,10	0,425 +0,19	0,589 +0,04	0,516 +0,11
5 - Subs - plongement	0,631 +0,14	0,583 +0,35	0,595 +0,05	0,540 +0,14
6 - Subs - modèle de langue	0,577 +0,09	0,481 +0,25	0,637 +0,09	0,586 +0,18
7 - InjM	**0,667** **+0,18**	**0,612** **+0,38**	0,565 +0,02	0,496 +0,09
8 - InjL	0,542 +0,05	0,493 +0,26	0,655 +0,11	0,610 +0,21
9 - AllInj (7+8)	0,625 +0,14	0,569 +0,34	0,613 +0,07	0,567 +0,16
10 - AllSubs (2+3+4)	0,524 +0,04	0,478 +0,25	**0,661** **+0,11**	**0,608** **+0,21**
11 - 1+2+3+4+5+6+7+8	0,601 +0,11	0,553 +0,32	0,595 +0,05	0,543 +0,14

TABLE 7.14 – Résultats d'évaluation et gains obtenus sur le corpusIx pour les différentes architectures de mBERT avec les différentes méthodes d'amplification.

Toutes les méthodes d'amplification de données augmentent la performance des modèles multilingues de BERT (voir tableau 7.14). La fourchette d'amélioration de

l'exactitude pour mBERT BC se situe entre +0,04 (min. d'E : 0,524 et F1 : 0,478 pour la méthode *AllSubs*) et +0,18 (max. d'E : 0,667 et F1 : 0,612 pour la méthode *InjM*). S'agissant de mBERT BU, elle se situe entre +0,02 (min. d'E : 0,565 et F1 : 0,481 pour la substitution lexicale par hyperonymes) et +0,11 (max. d'E : 0,661 et F1 : 0,608 pour la méthode *AllSubs*).

C Effet de l'amplification des données : corpusY

Les différentes méthodes d'amplification n'apportent pas d'améliorations significatives pour l'exactitude comme cela est observé avec les modèles FlauBERT ou CamemBERT. La fourchette d'amélioration de l'exactitude pour mBERT BC se situe entre +0,02 (pour la rétrotraduction et la méthode *InjL*) à +0,09 (substitution par hyperonymes). Pour le modèle mBERT BU, seule la technique d'injection de bruit au niveau du mot apporte un gain supérieur à 0,10, soit +0,12 par rapport à la référence.

TAD	CorpusY - test			
	mBERT BC		mBERT BU	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,674	0,269	0,667	0,418
1 - Rétrotraduction	0,698 +0,02	0,577 +0,31	0,682 +0,02	0,550 +0,13
2 - Subs - Syn	0,729 +0,05	0,587 +0,32	0,605 -0,06	0,524 +0,11
3 - Subs - hype	**0,760** **+0,09**	**0,644** **+0,37**	0,744 +0,08	0,544 +0,13
4 - Subs - hypo	0,752 +0,08	0,550 +0,28	0,667 0,00	0,518 +0,10
5 - Subs - plongement	0,729 +0,05	0,617 +0,35	0,698 +0,03	0,542 +0,12
6 - Subs - modèle de langue	0,705 +0,03	0,499 +0,23	0,729 +0,06	0,583 +0,16
7 - InjM	0,721 +0,05	0,591 +0,32	**0,783** **+0,12**	**0,660** **+0,24**
8 - InjL	0,698 +0,02	0,561 +0,29	0,767 +0,10	0,651 +0,23
9 - AllInj (7+8)	0,698 +0,02	0,556 +0,29	0,659 -0,01	0,559 +0,14
10 - AllSubs (2+3+4)	0,752 +0,08	0,634 +0,36	0,690 +0,02	0,522 +0,10
11 - 1+2+3+4+5+6+7+8	0,674 0,00	0,540 +0,27	0,736 +0,07	0,578 +0,1

TABLE 7.15 – Résultats d'évaluation sur le corpusY de posts pour les différentes architectures de mBERT avec les différentes méthodes d'amplification.

D Résultats pour les meilleures méthodes d'amplification pour les deux corpus de test

Après avoir examiné l'impact des meilleures méthodes d'amplification pour l'architecture multilingue de mBERT, nous détaillons les meilleurs résultats obtenus pour chacun des corpus de test dans les tableaux 7.16 et 7.17.

TAD	CorpusIx - test		CorpusY - test	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,488	0,231	0,674	0,269
7 - InjM + mBERT BC	**0,667 (+0,18)**	**0,612 (+0,38)**	0,721 (+0,05)	0,591 (+0,32)

TABLE 7.16 – Meilleur résultat obtenu pour le corpusIx avec mBERT.

TAD	CorpusY - test		Corpusix- test	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,667	0,418	0,548	0,403
7 - InjM + mBERT BU	**0,783 (+0,12)**	**0,660(+0,24)**	0,565 (+0,02)	0,496 (+0,09)

TABLE 7.17 – Meilleur résultat obtenu pour le corpusY avec mBERT.

7.2.4 Discussions

De manière globale, nous avons constaté qu'une grande majorité des méthodes d'amplification testées, ont eu un impact positif sur notre tâche de classification en augmentant l'exactitude et par ricochet le score de F1. Les modèles développés avec CamemBERT sont ceux dont la performance demeure supérieure à celle des autres modèles (FlauBERT, mBERT). En termes de techniques d'amplification, l'injection de bruit au niveau du mot et la substitution par synonymes sont celles qui ont mieux performé pour le corpusIx. Nous avons obtenu la meilleure valeur d'exactitude avec CamemBERT B pré-entraîné sur 4 GB de CCNet (0,780) suivi de FlauBERT BU (0,715) et mBERT BC (0,667) pour le corpusIx (voir tableau 7.18). Les résultats globaux montrent que les méthodes appliquées séparément (méthode 7, 2) performent mieux que celles intégrant plusieurs techniques d'amplification. Les résultats présentés dans cette section sont très encourageants et montrent clairement l'impact des méthodes d'amplification sur les architectures de type Transformer pour notre tâche de classification que ce soit sur les données de verbatim ou de posts.

TAD	CorpusIx - test		CorpusY - test	
	Exactitude	F1-macro	Exactitude	F1-macro
7 - InjM + FlauBERT BU	**0,714 (+0,21)**	**0,683 (+0,42)**	0,791 (+0,13)	0,703 (+0,44)
9 + Allinj (7+8) + CamemBERT L 135 GB (CCNet)	**0,774 (+0,28)**	**0,739 (+0,42)**	0,705 (+0,01)	0,586 (+0,24)
2 - Subs-Syn + CamemBERT B entraîné sur 4GB (CCNet)	**0,780 (+0,19)**	**0,742 (+0,32)**	0,798 (+0,07)	0,694 (+0,24)
7 - InjM + mBERT BC	**0,667 (+0,18)**	**0,612 (+0,38)**	0,721 (+0,05)	0,591 (+0,32)

TABLE 7.18 – Meilleur résultat obtenu pour le corpusIx avec chaque architecture.

7.3 Expériences et résultats : Données de posts

Dans la section précédente, nous avons détaillé les résultats obtenus à partir des modèles appris sur les corpus de verbatim. Nous avons observé que les méthodes d'amplification des données se sont avérées bénéfiques pour notre tâche de classification. Dans cette section, nous avons mené une nouvelle série d'expériences sur un autre jeu de données, le corpus de posts initial et ses versions amplifiées afin de valider notre méthodologie et de vérifier nos résultats. Nous détaillons dans la suite les résultats obtenus avec les modèles développés à partir des corpus amplifiés de posts. Le protocole d'évaluation mis en œuvre est le même que celui détaillé à la section 7.1.3.

7.3.1 Modèle FlauBert

Nous détaillons dans les tableaux 7.19 et 7.20 les résultats obtenus sur les deux corpus de test (corpusIx et corpusY). Les métriques utilisées restent l'exactitude et la F1-macro.

A Résultat général

Globalement, les résultats présentés viennent confirmer que les méthodes d'amplification des données ont un effet positif sur la détection de freins, motivations et conditions pour le corpusY. Nous avons obtenu la meilleure valeur d'exactitude de 0,837 (+0,17) et 0,766 (+0,50) en F1-macro pour le corpusY avec le modèle FlauBERT BU et la substitution par hyponymes par rapport à la base de référence. Étant donné que les modèles détaillés dans cette section sont appris sur le corpus de posts, nous rajoutons également les résultats obtenus par ces derniers sur le corpu-

sIx. Pour le corpusIx, nous avons obtenu la meilleure exactitude de 0,673 (+0,16) avec l'injection de bruit au niveau du mot et FlauBERT L.

B Effet de l'amplification des données : corpusY

Les résultats détaillés dans la suite sont présentés au tableau 7.19. La méthode 11 (addition de plusieurs jeux de données) est la méthode la plus performante avec FlauBERT BC et IxBERT. La substitution lexicale par hyponymes est celle qui performe le mieux avec FlauBERT L. Pour FlauBERT BU, c'est la méthode de substitution par plongement de mots qui s'impose. Si on considère uniquement les améliorations de l'exactitude apportées par les méthodes d'AD, celles-ci se situent dans une fourchette d'amélioration allant de **+0,01** à **+0,17** pour l'ensemble des architectures.

Nous avons obtenu une fourchette d'amélioration de l'exactitude pour FlauBERT BC allant de +0,02 (min. E : 0,690 et F1 : 0,551 pour la méthode de substitution par hyponymes) à +0,11 (max. E : 0,775 et F1 : 0,684 pour la méthode 11).

TAD	CorpusY - test							
	FlauBERT BC		FlauBERT BU		IxBERT		FlauBERT L	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,667	0,269	0,674	0,269	0,682	0,466	0,667	0,267
1 - Retrotraduction	0,698 +0,03	0,514 +0,24	0,791 +0,12	0,660 +0,39	0,682 0,00	0,554 +0,09	0,822 +0,15	0,750 +0,48
2 - Subs - Syn	0,713 +0,05	0,582 +0,31	0,752 +0,08	0,621 +0,35	0,674 -0,01	0,451 -0,01	0,829 +0,16	0,733 +0,47
3 - Subs - hype	0,705 +0,04	0,466 +0,20	0,698 +0,02	0,450 +0,18	0,651 -0,03	0,500 +0,03	0,791 +0,12	0,735 +0,47
4 - Subs - hypo	0,690 +0,02	0,551 +0,28	0,713 +0,04	0,511 +0,24	0,690 +0,01	0,548 +0,08	**0,837** **+0,17**	**0,766** **+0,50**
5 - Subs - plongement	0,713 +0,05	0,517 +0,25	**0,814** **+0,14**	**0,713** **+0,44**	0,682 0,00	0,565 +0,10	0,783 +0,12	0,694 +0,43
6 - Subs - modèle de langue	0,767 +0,10	0,641 +0,37	0,736 +0,06	0,459 +0,19	0,705 +0,02	0,552 +0,09	0,721 +0,05	0,612 +0,34
7 - InjM	0,736 +0,07	0,648 +0,38	0,721 +0,05	0,515 +0,25	0,690 +0,01	0,567 +0,10	0,814 +0,15	0,723 +0,46
8 - InjL	0,651 -0,02	0,484 +0,22	0,798 +0,12	0,714 +0,44	0,721 +0,04	0,497 +,03	0,829 +0,16	0,729 +0,46
9 - AllInj (7+8)	0,775 +0,11	0,637 +0,37	0,720 +0,21	0,595 +0,57	0,721 +0,04	0,599 +0,13	0,791 +0,12	0,686 +0,42
10 - AllSubs (2+3+4)	0,736 +0,07	0,599 +0,33	0,744 +0,07	0,674 +0,40	0,713 +0,03	0,506 +0,04	0,682 +0,02	0,303 +0,04
11 - 1+2+3+4+5+6+7+8	**0,775** **+0,11**	**0,684** **+0,41**	0,744 +0,07	0,603 +0,33	**0,729** **+0,05**	**0,619** **+0,15**	0,791 +0,12	0,712 +0,44

TABLE 7.19 – Résultats d'évaluation sur le corpusY pour chaque architecture de FlauBERT avec les différentes méthodes d'amplification.

Pour le modèle FlauBERT BU, la fourchette d'amélioration va de +0,02 (min. E : 0,698 et F1 : 0,450 pour la méthode de substitution lexicale par hyperonymes) à +0,14 (max. E : 0,814 et F1 : 0,713 pour la méthode de substitution par plongement de mots). S'agissant du modèle IxBERT, cette fourchette se situe entre +0,01 (min. E : 0,690 et F1 : 0,548 pour la substitution par hyponymes) et +0,05 (max. E : 0,729 et F1 : 0,619 pour la méthode d'addition de tous les jeux de données). La fourchette d'amélioration de l'exactitude pour FlauBERT L se situe entre +0,02 (min. E : 0,682 et F1 : 0,303 pour la méthode *AllSubs*) et +0,17 (max. E : 0,837 et F1 : 0,766 pour la méthode de substitution par hyponymes). Nous avons également des dégradations dans la performance de certains modèles dont IxBERT mais ces dernières restent très faibles avec -0,01 pour la substitution lexicale par synonymes à -0,03 pour la substitution par hyperonymes par rapport à la base de référence.

C Effet de l'amplification des données : corpusIx

Nous avons obtenu la meilleure valeur d'exactitude de 0,673 avec la méthode *InjM* et le modèle FlauBERT L pour le corpusIx par rapport à la base de référence.

TAD	CorpusIx - test							
	FlauBERT BC		FlauBERT BU		IxBert		FlauBERT L	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,464	0,335	0,482	0,217	0,458	0,435	0,512	0,318
1 - Retrotraduction	0,494 +0,03	0,448 +0,11	0,536 +0,05	0,483 +0,27	0,494 +0,04	0,479 +0,04	0,542 +0,03	0,521 +0,20
2 - Subs - Syn	0,571 +0,11	0,535 +0,20	0,476 -0,01	0,432 +0,22	0,482 +0,02	0,388 -0,05	0,643 +0,13	0,625 +0,31
3 - Subs - hype	0,357 -0,11	0,281 -0,05	0,393 -0,09	0,306 +0,09	0,429 -0,03	0,409 -0,03	0,536 +0,02	0,531 +0,21
4 - Subs - hypo	0,530 +0,07	0,501 +0,17	0,375 -0,11	0,335 +0,12	0,470 +0,01	0,440 0,00	0,631 +0,12	0,615 +0,30
5 - Subs - plongement	0,524 +0,06	0,500 +0,17	0,518 +0,04	0,482 +0,26	0,429 -0,03	0,420 -0,01	0,613 +0,10	0,589 +0,27
6 - Subs - modèle de langue	0,482 +0,02	0,461 +0,13	0,429 -0,05	0,321 +0,10	0,482 +0,02	0,453 +0,02	0,452 -0,06	0,447 +0,13
7 - InjM	0,470 +0,01	0,467 +0,13	0,417 -0,07	0,367 0,15	0,417 -0,04	0,404 -0,03	**0,673** **+0,16**	**0,633** **+0,31**
8 - InjL	0,292 -0,17	0,282 -0,05	0,571 +0,09	0,563 +0,35	0,464 +0,01	0,386 -0,05	0,529 +0,02	0,512 +0,19
9 - AllInj (7+8)	0,482 0,02	0,438 0,10	0,440 -0,04	0,366 -0,14	0,446 -0,01	0,414 -0,02	0,500 -0,01	0,496 +0,18
10 - AllSubs (2+3+4)	0,452 -0,01	0,371 +0,04	0,464 -0,02	0,468 +0,25	0,470 +0,01	0,401 -0,03	0,488 -0.02	0,231 -0.09
11 - 1+2+3+4+5+6+7+8	0,464 0,00	0,459 +0,12	0,536 +0,05	0,524 +0,31	0,452 +0,02	0,441 +0,01	0,560 0,05	0,544 +0,23

TABLE 7.20 – Résultats d'évaluation sur le corpusIx pour chaque architecture de FlauBERT avec les différentes méthodes d'amplification.

7.3.2 Modèle CamemBERT

Dans cette section, nous présentons les résultats obtenus avec les différentes tailles du modèle CamemBERT et les différentes méthodes d'AD dans les tableaux 7.21 à 7.24 pour les deux corpus de test.

A Résultat général

Nous avons obtenu la meilleure valeur d'exactitude de 0,868 (+0,19) et 0,801 (+0,53) en F1-macro pour le corpusY avec le modèle CamemBERT L et la substitution par synonymes par rapport à la base de référence sans amplification. En évaluant le même modèle sur le corpusIx, nous avons obtenu une exactitude 0,655 et une valeur de F1 de 0,644 par rapport à la référence (0,482, F1 de 0,217).

B Effet de l'amplification des données : corpusY

TAD	CorpusY - test					
	CamemBERT B 138 GB (OSCAR)		CamemBERT B 135 GB (CCNet)		CamemBERT L - 135 GB (CCNet)	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,674	0,269	0,752	0,484	0,674	0,269
1 - Retrotraduction	**0,783** **+0,11**	**0,68** **+0,41**	0,814 +0,06	0,755 +0,27	0,822 +0,15	0,771 +0,50
2 - Subs - Syn	0,767 +0,09	0,672 +0,40	0,845 +0,09	0,759 +0,27	**0,868** **+0,19**	**0,801** **+0,53**
3 - Subs - hype	0,752 +0,08	0,578 +0,31	0,736 -0,02	0,592 +0,11	0,837 0,16	0,738 0,47
4 - Subs - hypo	0,721 +0,05	0,513 +0,24	0,814 +0,06	0,715 +0,23	0,845 +0,17	0,781 +0,51
5 - Subs - plongement	0,760 +0,09	0,679 +0,41	0,822 +0,07	0,731 +0,25	0,829 +0,16	0,761 +0,49
6 - Subs - modèle de langue	0,767 +0,09	0,652 +0,38	0,698 -0,05	0,610 +0,13	0,775 +0,10	0,675 +0,41
7 - InjM	0,744 +0,07	0,666 +0,40	**0,853** **+0,10**	**0,780** **+0,30**	0,775 +0,10	0,708 +0,44
8 - InjL	0,744 +0,07	0,580 +0,31	0,806 +0,05	0,731 +0,25	0,806 +0,13	0,624 +0,36
9 - AllInj (7+8)	0,713 +0,04	0,629 +0,36	0,814 +0,06	0,741 +0,26	0,822 +0,15	0,733 +0,46
10 - AllSubs (2+3+4)	0,736 +0,06	0,570 +0,30	0,822 +0,07	0,740 +0,26	0,853 +0,18	0,787 +0,52
11 - 1+2+3+4+5+6+7+8	0,752 +0,08	0,609 +0,34	0,783 +0,03	0,716 +0,23	0,791 +0,12	0,651 +0,38

TABLE 7.21 – Résultats d'évaluation sur le corpus de test de posts pour les différentes architectures de CamemBERT avec différentes méthodes d'amplification. Les modèles utilisés sont ceux entraînés sur l'ensemble du corpus d'apprentissage (138/135 GB).

Modèles 138/135 GB. Nous avons obtenu une fourchette d'amélioration (voir le

tableau 7.21) de l'exactitude allant de +0,04 (min. E : 0,713 et F1 : 0,629 pour la méthode *Allinj*) à +0,11 (min. E : 0,783 et F1 : 0,68 pour la méthode de rétrotraduction) pour CamemBERT B pré-entraîné sur 138 GB d'OSCAR. Le modèle CamemBERT B pré-entraîné sur 135 de GB de CCNet enregistre une fourchette d'amélioration de l'exactitude qui se situe entre +0,03 (min E : 0,783 et F1 : 0,716 pour la méthode 11) et +0,10 (max. E : 0,853 et F1 : 0780 pour la *InjM*). La fourchette d'amélioration de l'exactitude pour CamemBERT L pré-entraîné sur 135 de GB de CCNet se situe entre +0,10 (0,775 et F1 de 0,675 pour la substitution lexicale à partir d'un modèle de langue) et +0,19 (0,868 et F1 de 0,801 pour la méthode de substitution par synonymes).

TAD	CorpusY - test					
	CamemBERT B 4 GB (CCNet)		CamemBERT B 4 GB (OSCAR)		CamemBERT B 4 GB (Wiki)	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,674	0,269	0,674	0,269	0,674	0,269
1 - Retrotraduction	0,845 +0,17	0,767 +0,50	0,767 +0,09	0,694 +0,43	0,736 +0,06	0,640 +0,37
2 - Subs - Syn	0,806 +0,13	0,684 +0,42	0,822 +0,15	0,755 +0,49	0,698 +0,02	0,548 +0,28
3 - Subs - hype	0,806 +0,13	0,715 +0,45	0,767 +0,09	0,664 +0,40	0,736 +0,06	0,465 +0,20
4 - Subs - hypo	0,829 +0,16	0,735 +0,47	0,845 +0,17	0,786 +0,52	0,744 +0,07	0,588 +0,32
5 - Subs - plongement	**0,860** **+0,19**	**0,775** **+0,51**	0,760 +0,09	0,645 +0,38	0,775 +0,10	0,607 +0,34
6 - Subs - modèle de langue	0,791 +0,12	0,686 +0,42	0,829 +0,16	0,764 +0,50	**0,791** **+0,12**	**0,656** **+0,39**
7 - InjM	0,783 +0,11	0,699 +0,43	0,798 +0,12	0,699 +0,43	0,698 +0,02	0,510 +0,24
8 - InjL	0,806 +0,13	0,714 +0,45	**0,860** **+0,19**	**0,808** **+0,54**	0,705 +0,03	0,453 +0,18
9 - AllInj (7+8)	0,814 +0,14	0,713 +0,44	0,806 +0,13	0,743 +0,47	0,736 +0,06	0,456 +0,19
10 - AllSubs (2+3+4)	0,806 +0,13	0,695 +0,43	0,845 +0,17	0,765 +0,50	0,729 +0,05	0,453 +0,18
11 - 1+2+3+4+5+6+7+8	0,783 +0,11	0,679 +0,41	0,775 +0,10	0,674 +0,40	0,713 +0,04	0,530 +0,26

TABLE 7.22 – Résultats d'évaluation sur le corpus de test de posts pour les différentes architectures de CamemBERT avec différentes méthodes d'amplification. Les modèles utilisés sont ceux entraînés sur une version réduite du corpus d'apprentissage (4 GB).

Modèles 4 GB. Toutes les méthodes d'amplification améliorent l'exactitude pour l'ensemble des architectures de CamemBERT pré-entraîné sur 4 GB. Les résultats sont illustrés au tableau 7.22. Pour avec CamemBERT pré-entraîné sur 4 GB de CCNet, la fourchette d'amélioration de l'exactitude se situe de +0,11 (min. E : 0,783 et

F1 : 0,679 pour la méthode 11) à +0,19 (max. E : 0,860 et F1 : 0,775 pour la méthode substitution lexicale par plongement de mots). Par contre, pour CamemBERT pré-entraîné sur 4 GB d'OSCAR, la fourchette d'amélioration de l'exactitude se situe de +0,09 (min. E : 0,760 et F1 : 0,645 pour la méthode de substitution lexicale par plongement de mots) à +0,19 (max. E : 0,860 et F1 : 0,808 pour la méthode InjL). Nous avons obtenu une fourchette d'amélioration de l'exactitude allant de +0,02 (min. E : 0,698 et F1 : 0,510 pour la méthode *InjM*) à +0,12 (max. E : 0,791 et F1 : 0,656 pour la méthode de substitution lexicale à partir d'un modèle de langue) avec CamemBERT pré-entraîné sur 4 GB de données issues Wikipédia.

B Effet de l'amplification des données : corpusIx

TAD	CorpusIx - test					
	CamemBERT B 138 GB (OSCAR)		CamemBERT B 135 GB (CCNet)		CamemBERT L - 135 GB (CCNet)	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,482	0,217	0,458	0,351	0,482	0,217
1 - Retrotraduction	0,595 +0,11	0,572 +0,36	0,595 +0,14	0,598 +0,25	0,667 +0,18	0,664 +0,45
2 - Subs - Syn	0,518 +0,04	0,506 +0,29	0,589 +0,13	0,545 +0,19	0,655 +0,17	0,644 +0,43
3 - Subs - hype	0,494 +0,01	0,432 +0,21	0,476 +0,02	0,431 +0,08	0,649 +0,17	0,625 +0,41
4 - Subs - hypo	0,482 0,00	0,426 +0,21	0,655 +0,20	0,641 +0,29	0,607 +0,13	0,592 +0,38
5 - Subs - plongement	0,524 +0,04	0,514 +0,30	0,583 +0,13	0,560 +0,21	0,708 +0,23	0,691 +0,47
6 - Subs - modèle de langue	0,554 +0,07	0,543 +0,33	0,488 +0,03	0,479 +0,13	0,554 +0,07	0,550 +0,33
7 - InjM	0,518 +0,04	0,492 +0,28	0,601 +0,14	0,574 +0,22	0,548 +0,07	0,544 +0,33
8 - InjL	0,512 +0,03	0,488 +0,27	**0,613** **+0,16**	**0,610** **+0,26**	0,548 +0,07	0,460 +0,24
9 - AllInj (7+8)	**0,631** **+0,15**	**0,629** **+0,41**	0,554 +0,10	0,546 +0,19	**0,720** **+0,24**	**0,701** **+0,48**
10 - AllSubs (2+3+4)	0,542 +0,06	0,499 +0,28	0,554 +0,10	0,536 +0,19	0,643 +0,16	0,641 +0,42
11 - 1+2+3+4+5+6+7+8	0,589 +0,11	0,573 +0,36	0,595 +0,11	0,584 +0,37	0,565 +0,08	0,449 +0,23

TABLE 7.23 – Résultats d'évaluation et gains obtenus sur le corpus de test de verbatim pour les différentes architectures de CamemBERT avec différentes méthodes d'amplification. Les modèles utilisés sont ceux entraînés sur l'ensemble du corpus d'apprentissage, à savoir le corpus d'apprentissage de 135GB.

Modèles 138/135 GB. Globalement, nous avons obtenu la meilleure valeur d'exactitude de 0,720 pour 0,701 de F1 avec la méthode *AllInj* et CamemBERT L pré-entraîné sur 135 GB de CCNet (voir tableau 7.23).

Modèles 4 GB. Nous avons obtenu la meilleure valeur d'exactitude de 0,696 pour 0,638 de F1 avec la méthode de substitution par hyponymes et CamemBERT B pré-entraîné sur 4 GB de CCNet. CamemBERT B pré-entraîné sur 4 GB de données issues de Wikipédia est le modèle qui performe le moins bien parmi toutes les architectures (voir tableau 7.24).

TAD	CorpusIx- test					
	CamemBERT B 4 GB (CCNet)		CamemBERT B 4 GB (OSCAR)		CamemBERT B 4 GB (Wiki)	
	Exactitude	F1-macro	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,482	0,217	0,482	0,217	0,482	0,217
1 - Retrotraduction	0,530 +0,05	0,524 +0,31	0,506 +0,02	0,501 +0,28	**0,571** **+0,09**	**0,562** **+0,34**
2 - Subs - Syn	0,649 +0,17	0,618 +0,40	0,607 +0,13	0,590 +0,37	0,399 -0,08	0,381 +0,16
3 - Subs - hype	0,595 +0,11	0,571 +0,35	0,571 +0,09	0,564 +0,35	0,440 -0,04	0,330 +0,11
4 - Subs - hypo	**0,696** **+0,21**	**0,638** **+0,42**	0,601 +0,12	0,571 +0,35	0,411 -0,07	0,391 +0,17
5 - Subs - plongement	0,673 +0,19	0,644 +0,43	**0,649** **+0,17**	**0,610** **+0,39**	0,530 +0,05	0,441 +0,22
6 - Subs - modèle de langue	0,512 +0,03	0,494 +0,28	0,565 +0,08	0,559 +0,34	0,435 -0,05	0,427 +0,21
7 - InjM	0,560 +0,08	0,552 + 0,34	0,655 +0,17	0,624 +0,41	0,464 -0,02	0,305 +0,09
8 - InjL	0,673 +0,19	0,643 +0,43	0,619 +0,14	0,596 +0,38	0,405 -0,08	0,317 +0,10
9 - AllInj (7+8)	0,565 +0,08	0,560 +0,34	0,512 +0,03	0,502 +0,29	0,464 -0,02	0,332 +0,11
10 - AllSubs (2+3+4)	0,482 0,00	0,437 +0,22	0,571 +0,09	0,561 +0,34	0,452 -0,03	0,332 +0,12
11 - 1+2+3+4+5+6+7+8	0,613 +0,13	0,570 +0,35	0,583 +0,10	0,561 +0,34	0,512 +0,03	0,455 +0,24

TABLE 7.24 – Résultats d'évaluation et gains obtenus sur le corpus de test de verbatim pour les différentes architectures de CamemBERT avec différentes méthodes d'amplification. Les modèles utilisés sont ceux entraînés sur une version réduite du corpus d'apprentissage (4 GB).

7.3.3 Modèle mBERT

Dans cette partie, nous présentons les résultats obtenus avec le modèle mBERT pour les deux corpus de test.

A Résultat général

Nous avons obtenu la meilleure valeur d'exactitude de 0,798 (+0,12) et 0,704 (+0,43) en F1-macro pour le corpusY avec le modèle mBERT BC et la méthode AllSubs par rapport à la base de référence sans amplification. En évaluant le même modèle sur le corpusIx, nous avons obtenu une exactitude 0,542 et une valeur de F1 de 0,525 par rapport à la référence. La meilleure valeur d'exactitude de 0,583 pour le corpusIx est obtenue avec la méthode *AllInj* et mBERT BC. Ces résultats sont détaillés dans les tableaux 7.25 et 7.26.

TAD	CorpusY - test			
	mBERT BC		mBERT BU	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,674	0,269	0,690	0,342
1 - Retrotraduction	0,791 0,12	0,683 +0,41	0,752 +0,06	0,628 +0,29
2 - Subs - Syn	0,682 +0,01	0,423 +0,15	0,674 -0,02	0,584 +0,24
3 - Subs - hype	0,760 +0,09	0,616 +0,35	0,767 +0,08	0,598 +0,26
4 - Subs - hypo	0,674 0,00	0,509 +0,24	0,736 +0,05	0,540 +0,20
5 - Subs - plongement	0,744 +0,07	0,660 +0,39	0,744 +0,05	0,659 +0,32
6 - Subs - modèle de langue	0,767 +0,09	0,646 +0,38	0,736 +0,05	0,615 +0,27
7 - InjM	0,760 +0,09	0,649 +0,38	0,736 +0,05	0,638 +0,30
8 - InjL	0,721 +0,05	0,486 +0,22	**0,767** **+0,08**	**0,647** **+0,31**
9 - AllInj (7+8)	0,729 +0,05	0,610 +0,34	0,736 +0,05	0,616 +0,27
10 - AllSubs (2+3+4)	**0,798** **+0,12**	**0,704** **+0,43**	0,690 0,00	0,608 +0,27
11 - 1+2+3+4+5+6+7+8	0,744 +0,07	0,655 +0,39	0,752 +0,06	0,639 +0,30

TABLE 7.25 – Résultats d'évaluation et gains obtenus sur le corpusY pour les différentes architectures de mBERT avec différentes méthodes d'amplification.

B Effet de l'amplification des données : corpusY

Globalement, toutes les méthodes d'amplification améliorent l'exactitude pour mBERT BC dans une fourchette d'amélioration allant de +0,01 (min. E : 0,682 et

F1 : 0,423 pour la substitution lexicale par synonymes) à +0,12 (meilleur modèle avec la méthode *AllSubs*). Pour l'architecture mBERT BU, seule la substitution lexicale par synonymes dégrade les performances du modèle, mais cette dégradation est peu significative (-0,02). La meilleure valeur d'exactitude a été obtenue avec l'injection de bruit au niveau du caractère avec une valeur de 0,767 pour 0,647 de F1-macro. Ces résultats sont détaillés dans le tableau 7.25.

TAD	CorpusIx - test			
	mBERT BC		mBERT BU	
	Exactitude	F1-macro	Exactitude	F1-macro
0 - Base de référence (sans AD)	0,482	0,217	0,446	0,315
1 - Retrotraduction	0,476 -0,01	0,462 +0,24	0,512 +0,07	0,475 +0,16
2 - Subs - Syn	0,435 -0,05	0,439 +0,22	0,440 -0,04	0,341 +0,03
3 - Subs - hype	0,542 +0,06	0,494 +0,28	**0,530** **+0,08**	**0,471** **+0,16**
4 - Subs - hypo	0,470 -0,01	0,433 +0,22	0,542 +0,10	0,481 +0,17
5 - Subs - plongement	0,500 +0,02	0,467 +0,25	0,488 +0,04	0,485 +0,17
6 - Subs - modèle de langue	0,506 +0,02	0,464 +0,25	0,435 -0,01	0,428 +0,11
7 - InjM	0,500 +0,02	0,492 +0,27	0,488 +0,04	0,480 +0,17
8 - InjL	0,482 0,00	0,381 +0,16	0,458 +0,01	0,452 +0,14
9 - AllInj (7+8)	**0,583** **+0,10**	**0,558** **+0,34**	0,464 +0,02	0,424 +0,11
10 - AllSubs (2+3+4)	0,542 +0,06	0,525 +0,31	0,482 +0,04	0,481 +0,17
11 - 1+2+3+4+5+6+7+8	0,500 +0,02	0,488 +0,27	0,506 +0,06	0,484 +0,17

TABLE 7.26 – Résultats d'évaluation et gains obtenus sur le corpus de test de verbatim pour les différentes architectures de mBERT avec différentes méthodes d'amplification.

7.3.4 Discussions

Dans cette section, nous présentons les résultats obtenus par la meilleure technique d'AD pour chaque modèle de langue. Les résultats sont détaillés dans le ta-

bleau 7.27. Nous avons obtenu la meilleure valeur d'exactitude avec la substitution lexicale par synonymes avec CamemBERT L pré-entraîné sur CCNet (135GB) pour le corpusY (0,868). La deuxième meilleure valeur est obtenue toujours avec un modèle CamemBERT mais pré-entraîné sur 4 GB du corpus OSCAR avec la méthode par injection au niveau du caractère.

TAD	CorpusY - test		CorpusIx - test	
	Exactitude	F1-macro	Exactitude	F1-macro
4 - Subs - hypo + FlauBERT L	0,837 (+0,17)	0,766 (+0,50)	0,631 (+0,12)	0,615 (+0,30)
2 - Subs - Syn + CamemBERT L pré-entraîné sur 135 GB (CCNet)	**0,868 (+0,19)**	**0,801 (+0,53)**	0,655 (+0,17)	0,644 (+0,43)
8- InjL + CamemBERT B pré-entraîné sur 4 GB (OSCAR)	0,860 (+0,19)	0,808 (+0,54)	0,619 (+0,14)	0,596 (+0,38)
10- AllSubs + mBERT BC	0,798 (+0,12)	0,704 (+0,43)	0,542 (+0,06)	0,525 (+0,31)

TABLE 7.27 – Meilleur résultat obtenu pour le corpusIx avec chaque architecture. Les améliorations par rapport à la base de référence sont notées par +/-.

7.3.5 Conclusion

De manière globale, nous avons constaté qu'une grande majorité des méthodes d'amplification étudiées ont eu un impact positif sur notre tâche de classification en améliorant l'exactitude et la mesure de F1-macro pour le corpusY et dans une moindre mesure le corpusIx. Les modèles monolingues sont ceux qui performent le mieux comparé au modèle multilingue de BERT. En termes de techniques d'AD, les méthodes d'injection de bruit au niveau du mot/caractère, les diverses méthodes de substitution lexicale ont permis d'obtenir des performances bien supérieures à la base de référence. Les résultats globaux sont présentés au tableau 9. Le meilleur modèle pour le corpusIx est obtenu en combinant CamemBERT B entraîné sur 4 GB (CCNet) et la méthode de substitution lexicale par synonymes. Pour le corpus Y, la meilleure performance a été obtenue utilisant CamemBERT B entraîné sur 135 GB (CCNet) et la méthode de substitution par synonymes.

TAD	Modèles générés à partir du corpus de Verbatim			
	CorpusIx - test		CorpusY - test	
	Exactitude	F1-macro	Exactitude	F1-macro
7 - InjM + FlauBERT BU	**0,714 (+0,21)**	**0,683 (+0,42)**	0,791 (+0,13)	0,703 (+0,44)
9 + Allinj (7+8) + CamemBERT L 135 GB (CCNet)	**0,774 (+0,28)**	**0,739 (+0,42)**	0,705 (+0,01)	0,586 (+0,24)
2 - Subs-Syn + CamemBERT B entraîné sur 4GB (CCNet)	**0,780 (+0,19)**	**0,742 (+0,32)**	0,798 (+0,07)	0,694 (+0,24)
7 - InjM + mBERT BC	**0,667 (+0,18)**	**0,612 (+0,38)**	0,721 (+0,05)	0,591 (+0,32)
TAD	Modèles générés à partir du corpus de post			
	CorpusY - test		CorpusIx - test	
	Exactitude	F1-macro	Exactitude	F1-macro
4 - Subs - hypo + FlauBERT L	0,837 (+0,17)	0,766 (+0,50)	0,631 (+0,12)	0,615 (+0,30)
2 - Subs - Syn + CamemBERT L pré-entraîné sur 135 GB (CCNet)	**0,868 (+0,19)**	**0,801 (+0,53)**	0,655 (+0,17)	0,644 (+0,43)
8- InjL + CamemBERT B pré-entraîné sur 4 GB (OSCAR)	0,860 (+0,19)	0,808 (+0,54)	0,619 (+0,14)	0,596 (+0,38)
10- AllSubs + mBERT BC	0,798 (+0,12)	0,704 (+0,43)	0,542 (+0,06)	0,525 (+0,31)

TABLE 7.28 – Meilleurs résultats obtenus pour le corpusIx et corpusY dans l'ensemble.

7.4 Plateforme de classification

L'objectif premier de ce travail était de proposer un modèle de détection en FMC pour des textes transcrits et à plus long terme des commentaires issus d'une plateforme communautaire. Ainsi, une plateforme dédiée à l'analyse de ces commentaires a été mise en place [8] et le meilleur modèle [9] développé y a été intégré. Cette plateforme a été créée pour permettre aux chargés d'études d'Ixiade de lancer le modèle d'analyse en FMC sur des documents textuels. Pour ce faire, ils déposent

8. La plateforme a été développé en collaboration avec un développeur. Pour le moment, elle est encore en cours de développement.

9. Dans un premier temps, nous avons utilisé uniquement le meilleur modèle pour le corpusIx (CamemBERT B pré-entraîné sur 4 GB de CCNet + substitution par synonymes) puisqu'il permettait également d'obtenir de bonnes performances (voir le tableau) pour le corpusY même s'il n'avait pas été entraîné sur ce type de données.

directement sur la plateforme les fichiers à analyser (format Word pour des transcriptions ou Excel pour les posts issus de Yoomaneo). Étant donné que la plateforme d'analyse et celle de Yoomaneo ne sont pas interconnectées l'une vers l'autre, le chargé d'études est obligé d'extraire le fichier de commentaires du projet spécifique dont il souhaite évaluer en FMC et le déposer sur la plateforme d'analyse. La plateforme dispose d'un encadré rectangulaire qui permet de visualiser les phrases classifiées avec leur catégorie FMC et CAUTIC comme observé sur la figure 7.1.

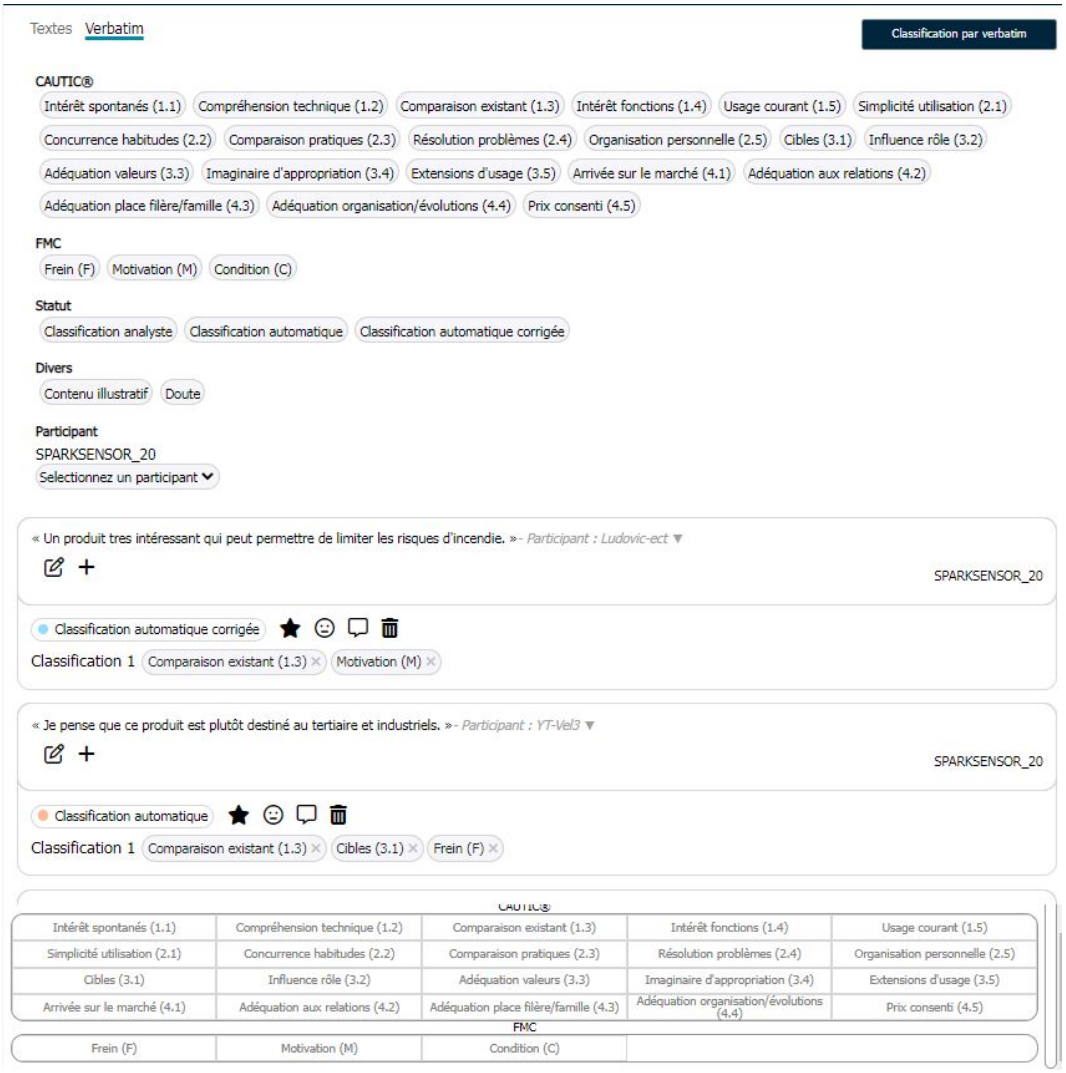


FIGURE 7.1 – Interface de la plateforme d'analyse : verbatim filtrés

7.5 Conclusion

Dans ce chapitre, nous avons présenté des études expérimentales pour découvrir les meilleures méthodes pour la classification en FMC en langue française. Les modèles générés ont été appris sur des corpus amplifiés via des différentes techniques issues de l'état de l'art et ont été évalués sur des corpus de natures différentes : un provenant de l'oral transcrit et constitué de verbatim et l'autre un corpus de posts/-commentaires ou plutôt des réponses écrites à une question posée dans le cadre d'un projet d'études. Confronté à un manque de données dès le départ, nous avons proposé d'amplifier notre jeu de données manuellement annoté par des experts grâce à des méthodes simples et rapides d'implémentation. Une fois, le corpus amplifié en fonction de chacune des méthodes, nous avons également proposé de combiner certaines techniques pour produire de plus grand jeu de données. L'objectif était de pouvoir examiner leur effet sur la tâche de classification. Les résultats montrent que l'amplification des données a permis d'accroître l'exactitude et la F1-macro pour la majorité des modèles utilisés dans une fourchette d'amélioration de +0,01 à +0,28 pour l'ensemble des résultats présentés dans ce chapitre. Pour l'heure, les résultats semblent satisfaisants et ont permis de proposer un modèle de classification qui a été intégré à une plateforme d'analyse. Toutefois, des améliorations restent à faire pour rendre le modèle plus efficient.

Chapitre 8

Conclusion générale

Dans ce travail de thèse, nous nous sommes intéressés à l'analyse de verbatim dans un contexte peu doté où nous avons proposé une méthodologie de construction d'un ensemble de trois corpus distincts d'apprentissage pour notre tâche de classification en FMC. Le processus, bien que long et fastidieux, a permis de disposer d'un jeu de données pour nous permettre d'entreprendre des expériences préliminaires. Nous avons ensuite étudié l'apport d'une méthode d'extraction de freins, de motivations et conditions pour un des trois corpus constitués (le corpus Amazon) dans l'optique de catégoriser nos classes et de pouvoir les extraire dans d'autres corpus. L'objectif poursuivi était d'agrandir nos faibles jeux de données. Confronté à un manque de résultats significatifs au travers de cette méthodologie, nous avons décidé d'amplifier deux de nos trois corpus préalablement constitués : un corpus de verbatim issu de transcriptions d'entretiens semi-directifs et de tables rondes et un corpus d'un ensemble de commentaires issus de la plateforme Yoomaneo. Nous avons proposé d'amplifier ces deux jeux de données à l'aide de différentes techniques d'amplification issues de l'état de l'art. Ces dernières avaient été massivement appliquées à la langue anglaise pour amplifier dans la grande majorité des cas des corpus de commentaires, de tweets et de critiques de films pour différentes tâches telles que l'analyse de sentiment, l'analyse de question-réponses, etc. Nos travaux sont les premiers à recourir à ces méthodes pour l'amplification de données de verbatim et de posts pour une tâche de détection de freins, motivations et conditions. Pour la partie classification, nous avons proposé de comparer l'impact de ces différentes méthodes d'amplification pour différents variants de BERT pour le français. Nous avons également proposé de comparer leurs résultats avec le modèle multilingue de BERT. Pour chaque modèle, nous avons étudié l'apport de chaque architecture. S'agissant des méthodes d'amplification utilisées, nous résumons dans le paragraphe suivant celles qui ont été utilisées.

1. l'injection de bruit au niveau du caractère et du mot qui consiste à altérer de manière aléatoire les mots ou caractères de mots présents dans une phrase pour produire de nouvelles phrases intégrant des erreurs de frappe ou d'orthographe ;
2. la substitution lexicale consiste à remplacer des mots dans une phrase à partir d'une ressource lexicale ou thésaurus par des équivalents similaires tels que des synonymes, hyponymes ou hyperonymes, à partir de plongements de mots tels que Word2vec ou encore FasTtext ou à partir d'un modèle de langue contextualisée ;
3. la rétrotraduction consiste à traduire dans un premier temps une phrase d'une langue A à B puis de retraduire la phrase traduite en langue B dans la langue d'origine A ;
4. À ces méthodes, nous avons également décidé d'étudier l'impact de l'addition de jeux données amplifiées à partir de différentes techniques d'AD (combinaison des méthodes de substitution lexicale à partir d'une ressource lexicale, injection de bruit au niveau du caractère et mot et addition des jeux de données de toutes les méthodes).

Ces méthodes d'AD ont permis d'amplifier nos jeux de données et de les utiliser pour générer des modèles à partir d'architectures de type Transformer.

A Apport des méthodes d'AD sur le processus d'apprentissage des modèles appris sur les corpus amplifiés

Apprentissage sur corpus de verbatim. Dans une première série d'expériences menées sur le corpus de verbatim, nous avons montré que l'ensemble des méthodes d'amplification utilisées dans ce travail de recherche ont permis d'accroître les résultats sur notre tâche de détection de FMC sur le corpus test de verbatim et de post pour l'ensemble des architectures. Avec 78% (0,780) d'exactitude pour le corpusIx, le modèle CamemBERT B pré-entraîné sur le corpus CCNET de 4 GB avec la méthode de substitution lexicale par synonymes est l'architecture la plus performante. Par contre, nous avons également observé une dégradation des performances pour certaines techniques telles que l'injection de bruit au niveau du caractère et la substitution par hyperonymes pour les architectures de FlauBERT L et CamemBERT pré-entraîné sur des données issues de Wikipédia (4 GB). Néanmoins, ces dégradations ne sont pas aussi significatives que nous le pensions. Nous avons aussi constaté que l'addition d'un très grand nombre de jeux de données issues de différentes techniques d'amplification (méthode 11) n'a pas permis d'améliorer significativement

les résultats comme nous le pensions. En effet, ce sont les méthodes individuelles qui se sont avérées plus performantes. En outre, les techniques d'injection de bruit se sont avérées dans l'ensemble plus efficace pour le corpus de verbatim et de posts. Enfin, les méthodes par substitution (hyponymes, plongements de mots, modèle de langue et syno-,hypo- et hyperonymes) et la rétrotraduction ont également permis d'obtenir de bien meilleures performances dans l'ensemble.

Les modèles de Transformers tous confondus utilisés dans cette thèse ont montré des résultats satisfaisants et encourageants pour la tâche de détection de FMC. Globalement, le modèle mBERT quelle que soit la taille d'architecture est celui qui performe le moins bien par rapport aux deux modèles français (FlauBERT et CamemBERT). Une des raisons serait liée au fait qu'il ait été entraîné sur plus d'une centaine de langues contrairement aux variants français qui ont été exclusivement entraîné sur des corpus en langue française. Cette observation a également été souligné dans les travaux de Chenais et al. (2021). L'utilisation et la comparaison des différentes tailles de modèles de CamemBERT a démontré que les modèles pré-entraînés sur un plus petit jeu de données (4 GB) pouvaient produire des résultats similaires à ceux des modèles pré-entraînés sur de grosses quantités de données (135 GB).

Apprentissage sur corpus de post. Dans une seconde série d'expériences, nous avons appris de nouveaux modèles sur les versions amplifiées du corpus de post. Pour le corpusY, l'architecture la plus performante est le modèle CamemBERT L pré-entraîné sur 135 GB de CCNet en utilisant la méthode de substitution lexicale par synonymes avec 87% (0,868) d'exactitude. Si nous nous focalisons sur la F1-macro, le gain observé est encore plus important avec +0,54 en utilisant CamemBERT B pré-entraîné sur 4 GB d'OSCAR avec la méthode d'injection de bruit au niveau du caractère. Les techniques de substitution lexicale toutes confondues ont permis d'accroitre l'exactitude pour toutes les architectures de Camembert à l'exception du modèle entraîné sur 4 GB de corpus issu de Wikipédia. Une raison à cela pourrait être expliqué par le type de données sur lequel le modèle de langue a été appris : 4 GB de données textuelles issues de Wikipédia.

Ces deux séries d'expérimentations ont montré l'apport des méthodes d'amplification pour notre tâche de classification dans un contexte où les données d'apprentissage étaient insuffisantes pour apprendre des modèles de classification générés à partir d'architecture Transformer. Les résultats obtenus montrent des améliorations tant au niveau de l'exactitude que du score F1. Dans la suite, nous partageons quelques pistes d'améliorations.

Perspectives

Les travaux effectués dans cette thèse peuvent être poursuivis selon plusieurs axes que nous détaillons dans la suite.

Axes à très court terme :

Système d'ensemble [1]. Le meilleur modèle sélectionné a été intégré à une plateforme d'analyse. Pour renforcer la précision au niveau des classifications générées par le modèle, nous envisageons d'utiliser également le meilleur modèle développé avec le corpus de post uniquement sur les données issues de Yoomaneo. Ceci nous conduirait à disposer de deux modèles distincts : le meilleur modèle sur le corpus de verbatim serait utilisé uniquement sur les données de verbatim et le meilleur modèle sur le corpus de posts pour les données Yoomaneo. Potentiellement, nous pourrions envisager de jumeler les deux modèles pour créer un modèle d'ensemble qui nous permettrait de renforcer la précision de la classification sur les deux types de données. Les deux modèles pourraient être couplés avec un autre algorithme classique de classification comme cela a été fait dans Mercadier (2020) ou ce dernier propose un système de vote pour l'analyse de sentiment pour des textes médicaux en combinant trois modèles de classification : deux architectures de Transformer couplées avec un algorithme classique de régression logistique.

Injection de connaissances extérieures. Nous envisageons également de pouvoir entraîner à nouveau les modèles en injectant des connaissances extérieures issues d'une ontologie développée en modélisant de manière experte le paradigme de l'entretien qualitatif d'évaluation de l'acceptabilité (Simonnet, 2022).

Méthodes d'amplification. Nous mettons à disposition de la communauté les scripts utilisés pour amplifier nos données.

Axes à moyen terme :

Améliorations à faire au niveau des Transformers. Dans ce travail de recherche, nous avons seulement utilisé une couche de classification en sortie de nos modèles préentraînés. Une piste à examiner serait d'étudier l'apport de plusieurs couches

1. Un système d'ensemble en apprentissage automatique est une méthode qui consiste à utiliser différents algorithmes d'apprentissage automatique en les regroupant pour obtenir de meilleures prédictions.

de classification pour le modèle retenue. Ceci dans l'optique d'étudier les résultats obtenus, et ce, en fonction des différentes méthodes d'amplification. Il serait intéressant d'examiner l'apport de plusieurs couches de classification notamment sur le jeu de données construit en additionnant des données amplifiées à partir de différentes techniques d'AD. C'est une tâche que nous n'avons pas eu l'occasion d'effectuer.

Méthode d'amplification. Dans ce travail de recherche, nous avons uniquement implémenté des méthodes d'amplification pour des textes en français et examiné de manière générale leur apport pour notre tâche. Il serait intéressant d'explorer pour chaque méthode d'amplification, les différents paramètres d'amplification choisis. Pour la méthode de substitution par exemple, nous avons uniquement remplacé 0,05% des mots d'une phrase en fonction de la longueur et n'avons pas testé les autres mesures telles 0,10% ou encore 0,15% de mots à remplacer comme dans (Feng et al., 2020).

Dans ce travail de recherche, nous avons pu démontrer qu'il était possible d'adapter dans un contexte où les données sont insuffisantes, différentes techniques initialement développées pour la langue anglaise à la langue française afin d'amplifier des données dans un contexte où ces dernières s'avèrent insuffisantes pour utiliser des architectures avec un nombre important de paramètres. En outre, les résultats obtenus ont montré des améliorations parfois importantes par rapport à la base de référence non amplifiée. Les résultats présentés dans leur globalité sont encourageants et ouvrent différentes perspectives de recherche.

Annexe A

Définitions

Méthode : Mobilisation cohérente d'outils visant la production d'une démonstration pour répondre à une problématique donnée.

Une innovation : Introduction d'un objet nouveau (technologie, produit, service, action de changement, process...) dans un milieu existant et impliquant un changement des manières de pratiquer et des savoir-faire.

Un changement : Processus de transformation d'un milieu existant après l'introduction d'une innovation.

Sens : Manière dont un individu comprend, juge, perçoit son environnement. On peut dire d'un projet pour lequel les critères CAUTIC sont validés, qu'il a « sens ».

Usage : Se servir de manière intentionnelle et volontaire d'un objet (produit ou service) dont on pense qu'il a des caractéristiques intéressantes (on fera l'effort de s'adapter à l'objet et/ou d'adapter l'objet).

Utilisateur : Celui qui fait usage au sens ci-dessus. On distingue 3 rangs d'usagers d'une innovation : l'utilisateur final, l'utilisateur final, l'utilisateur intermédiaire (souvent acheteur/prescripteur) et les porteurs de projets.

Signification d'usage : sens et valeur que l'utilisateur attribue à l'innovation qui lui est proposée.

Expériences utilisateurs : sens et valeur que l'utilisateur attribue à l'innovation qui lui est proposée.

Régression : prédiction d'une valeur ou d'un ensemble de valeur.

Classification : classifier les données en fonction de leurs caractéristiques.

Annexe B

Tableaux

Règles de Frein
Verb/Adj/Noun/Adv + "mais + Verb/Adj/Noun/Adv
*Verb/Adj/Noun/Adv + "mais" + **
*"mais" + Pron/ProPn * + Adv_neg + Verb*
"mais" + Adv_neg + Verb + L
*"mais" + Pron/ProPn + Adv_neg + Pron * + Verb*
"mais" + Adp/Det + Noun + Adv_neg + Pron + Verb*
"mais" + Adp + Adv_neg + Adv + Verb*
"mais" + Adp + L
*"mais" + Adj/Verb/Noun/Adv + Pron * + Adv_neg + Verb*

TABLE B.1 – Règles établies pour les freins.

L : Correspond à un adverbe ou une préposition, pronom qui s'utilise avec le négateur "ne" : pas, plus, aucun, rien, jamais, nullement, personne, etc.

***** : Placé devant une catégorie signifie que l'élément apparaît entre 0 et plusieurs fois.
? : Signifie que l'élément peut être présent une fois ou peut ne pas être présent.

Règles de motivation
*Verb/Adj/Noun + "mais" + * + Adv_neg + Pron/Noun * + Verb + L + * + Verb/Adj/Noun (1)*
*Verb/Adj/Noun + "mais" + Adv_neg + * + Adv_neg + Pron/Noun * + Verb + L + * + Verb/Adj/Noun (2)*
** + Adv_neg + Pron/Noun * + Verb + L + * +* *"mais" + * + Adv_neg + Noun/Pron * + Verb + L + * + Verb/Adj/Noun (8)*
** + Adv_neg + Noun * + Verb + L + * + "mais" + * + Adj/Verb/Noun (9)*
** + Adv_neg * + Noun * + Verb + L + * + "mais" + * + Adj/Verb/Noun (10)*
*"mais" + * Adj/Verb/Noun + Adv_neg + Noun/Pron * + Verb + L + * + Verb/Adj/Noun (11)*
*Adj/Verb/Noun + * + "mais" + * + Adj/Verb/Noun (3)*
*Adj/Verb/Noun + "mais" + * + Adj/Verb/Noun + Adv_neg + Pron/ Noun * + Verb + L + * + Adj/Verb/Noun (23)*
*Adj/Verb/Noun + "mais" + Adj/Verb/Noun/Adv + Adj/Verb/Noun/Adv * + Adj/Verb/Noun * + Adj/Verb/Noun (4)*
*Adj/Verb/Noun + "mais" + Adv * + Adj/Verb/Noun (5)*
*Adj + Noun + * + "mais" + Adv * + Adj/Verb/Noun (6)*
*"mais" + Pron + Verb + Adv * + Adj (12)*
*"mais" + * + Pron + Verb + Adv * + Adj (13)*
*"mais" + * + Det + Noun + Verb + Adv * + Adj (14)*
*"mais" + * + Det + Adj + Noun (15)*
*"mais" + * + Adv ? + Adj (16)*
*"mais" + * + Pron + Verb + Det * + Adv * + Adj (17)*
*"mais" + Pron + Verb + Det * + Adv * + Adj (18)*
*"mais" + * + Adv + Det * + Adj (19)*
"mais" + Pron + Verb + Det + Noun + ADP + Verb/Adj/Noun (20)
** + Adv_neg + Verb + L + * + "mais" + * + Adv * + Verb * + Verb/Adj/Noun (21)*
** + Adv_neg * + Verb + L + * + "mais" + Verb + Det + Adj + Noun (22)*
Det + Noun + Verb + Adj + "mais" + Adv ? + *Adj/Verb/ Noun (7)*

TABLE B.2 – Règles établies pour les motivations.

Annexe C

Bibliographie

Bibliographie

Amine Abdaoui, Jérôme Azé, Sandra Bringay, et Pascal Poncelet. Feel : a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3) : 833–855, 2017. 62, 114, 116

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, et Mahmoud Al-Ayyoub. Arabic sentiment analysis : Lexicon-based and corpus-based. In *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*, pages 1–6. IEEE, 2013. 62

Anne Abeillé et Danièle Godard. The grande grammaire du français project. In *Proceedings of LREC*, 2010. 64

Gavin Abercrombie et Riza Batista-Navarro. 'aye'or'no' ? speech-level sentiment analysis of hansard uk parliamentary debate transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018. 22, 42

Milam Aiken et Mina Park. The efficacy of round-trip translation for mt evaluation. *Translation Journal*, 14(1) :1–10, 2010. 148

Chedi Bechikh Ali, Halla Mulki, et Hatem Haddad. Impact du prétraitement linguistique sur l'analyse des sentiments du dialecte tunisien. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, page 383, 2018. 64

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, et Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv :1904.03323*, 2019. 93

Jean-Claude Anscombre et Oswald Ducrot. L'argumentation dans la langue. *Langages*, (42) :5–27, 1976. 50

Wissam Antoun, Fady Baly, et Hazem Hajj. Arabert : Transformer-based model for arabic language understanding. *arXiv preprint arXiv :2003.00104*, 2020. 93

Segun Taofeek Aroyehun et Alexander Gelbukh. Aggression detection in social media : Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97, 2018. 144

Nicholas Asher, Farah Benamara, et Yannick Mathieu. Distilling opinion in discourse : A preliminary study. In *Proceedings of COLING 2008 : Companion volume : Posters*, pages 7–10, 2008. 54

Magdalena Augustyn, Sabrina Ben Hamou, Gwendoline Bloquet, Vannina Goossens, Mathieu Loiseau, et Fanny Rinck. Lexique des affects : constitution de ressources pédagogiques numériques. In *Colloque International des étudiants-chercheurs en didactique des langues et linguistique.*, 2006. 63, 116

Jérome Azé et Mathieu Roche. Présentation de l'atelier deft'05. In *proceedings of TALN*, pages 99–111, 2005. 52

Dzmitry Bahdanau, Kyunghyun Cho, et Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*, 2014. 89

Rushlene Kaur Bakshi, Navneet Kaur, Ravneet Kaur, et Gurpreet Kaur. Opinion mining and sentiment analysis. In *2016 3rd international conference on computing for sustainable global development (INDIACom)*, pages 452–455. IEEE, 2016. 47

Michele Banko et Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pages 26–33, 2001. 129

Javier Barcenilla et Joseph Maurice Christian Bastien. L'acceptabilité des nouvelles technologies : quelles relations avec l'ergonomie, l'utilisabilité et l'expérience utilisateur ? *Le travail humain*, 72(4) :311–331, 2009. 33

Amira Barhoumi. *Une approche neuronale pour l'analyse d'opinions en arabe*. PhD thesis, Le Mans, 2020. 49, 59, 62

Amira Barhoumi, Nathalie Camelin, et Yannick Estève. Des représentations continues de mots pour l'analyse d'opinions en arabe : une étude qualitative. In

Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018, page 215, 2018. 64

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, et Dmitry Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. In *artificial intelligence and statistics*, pages 130–138. PMLR, 2016. 140

B Bathelot. L'encyclopédie illustrée du marketing. En ligne à l'adresse : http://www.definitions-marketing.com/definition/marketing, year=2015. 68

Markus Bayer, Marc-André Kaufhold, et Christian Reuter. A survey on data augmentation for text classification. *arXiv preprint arXiv :2107.03158*, 2021. 129, 130, 136

Yonatan Belinkov et Yonatan Bisk. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv :1711.02173*, 2017. 131, 132, 133, 135

Farah Benamara, Sylwia Ozdowska, et Laurent Mazuel. Actes de la 14ème conférence sur le traitement automatique des langues naturelles. In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. REncontres jeunes Chercheurs en Informatique pour le Traitement Automatique des Langues (Posters)*, 2007. 109

Farah Benamara, Véronique Moriceau, et Yvette Yannick Mathieu. Catégorisation sémantique fine des expressions d'opinion pour la détection de consensus. *Actes du dixième DÉfi Fouille de Textes*, page 43, 2014. 63

Farah Benamara, Cyril Grouin, Jihen Karoui, Véronique Moriceau, et Isabelle Robba. Analyse d'opinion et langage figuratif dans des tweets en français". *In Actes de l'atelier Défi fouille de textes-DEFT 2017 : 24éme Conférence TALN : Traitement Automatique des Langues naturelles*, 2017a. 52

Farah Benamara, Cyril Grouin, Jihen Karoui, Véronique Moriceau, et Isabelle Robba. Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du défi fouille de textes deft2017. *In Actes de l'atelier Défi fouille de textes-DEFT 2017 : 24éme Conférence TALN : Traitement Automatique des Langues naturelles*, 2017b. 66

Émile Benveniste. Problèmes de linguistique générale. *Gallimard*, 1966. 50

Delphine Bernhard et Anne-Laure Ligozat. Analyse automatique de la modalité et du niveau de certitude : application au domaine médical (automatic analysis of modality and level of certainty : application to the medical domain). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 352–363, 2011. 66, 72

Noémi Boubel. Extraction automatique de modifieurs de valence affective dans un texte. 2001. 66

Noémi Boubel. Construction automatique d'un lexique de modifieurs de polarité (automatic construction of a contextual valence shifters lexicon)[in french]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 3 : RECITAL*, pages 123–136, 2012. 66

Dominique Boullier et Audrey Lohard. *Opinion mining et Sentiment analysis : Méthodes et outils*. OpenEdition Press, 2012. 45, 49

Caroline Brun, Diana Nicoleta Popa, et Claude Roux. Xrce : Hybrid classification for aspect-based sentiment analysis. In *SemEval@ COLING*, pages 838–842. Citeseer, 2014. 51

Daphne Blunt Bugental. Acquisition of the algorithms of social life : A domain-based approach. *Psychological bulletin*, 126(2) :187, 2000. 72

Bernard Buisson et Philippe Silberzahn. Innovations de rupture : il n'y a pas de fatalité. *L'Expansion Management Review*, (1) :100–105, 2005. 29

Amélie Buttard. Emotion et acceptabilité : quelle influence des valeurs psychologiques ? 2018. 28

Frederik Cailliau et Ariane Cavet. Analyse des sentiments et transcription automatique : modélisation du déroulement de conversations téléphoniques. *Revue Traitement Automatique des Langues*, 51(3) :131–154, 2010. 22

Nathalie Camelin, Géraldine Damnati, Frédéric Béchet, et Renato De Mori. Détection automatique d'opinions dans des corpus de messages oraux. *Journées d'Etude sur la Parole (JEP'06), Dinard*, 2006. 22

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, et Jorge Pérez. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020 :2020, 2020. 93

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, et Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5) :301–310, 2001. 65

Patrick Charaudeau. Grammaire du sens et de l'expression, hachette. *Éducation*, pages 55–75, 1992. 56, 57

Baptiste Chardon. *Chaîne de traitement pour une approche discursive de l'analyse d'opinion*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2013. 13, 54, 65, 66

Karan Chawla, Ankit Ramteke, et Pushpak Bhattacharyya. Iitb-sentiment-analysts : Participation in sentiment analysis in twitter semeval 2013 task. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 495–500, 2013. 52

Hella Chemingui et Hajer Ben lallouna. Resistance, motivations, trust and intention to use mobile financial services. *International Journal of Bank Marketing*, 31(7) :574–592, 2013. 68, 70, 75

Gabrielle Chenais, Hélène Touchais, Marta Avalos, Loïck Bourdois, Philippe Revel, Cédric Gil-Jardiné, et Emmanuel Lagarde. Performance en classification de données textuelles des passages aux urgences des modèles bert pour le français. In *PFIA 2021-Journée Santé et IA*, 2021. 191

TC Chinsha et Shibily Joseph. Aspect based opinion mining from restaurant reviews. *International Journal of Computer Applications*, 975 :8887, 2014. 59

Nalini Chintalapudi, Gopi Battineni, Marzio Di Canio, Getu Gamo Sagaro, et Francesco Amenta. Text mining with sentiment analysis on seafarers' medical documents. *International Journal of Information Management Data Insights*, 1 (1) :100005, 2021. 49

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, et Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*, 2014. 89

Vincent Claveau, Antoine Chaffin, et Ewa Kijak. La génération de textes artificiels en substitution ou en complément de données d'apprentissage. In *Traitement Automatique des Langues Naturelles*, pages 124–136. ATALA, 2021. 146

Mathieu Cliche. Bb_twtr at semeval-2017 task 4 : Twitter sentiment analysis with cnns and lstms. *arXiv preprint arXiv :1704.06125*, 2017. 64

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1) :37–46, 1960. 109

Francis Corblin et Lucia Tovena. L'expression de la négation dans les langues romanes. *Les langues romanes : problèmes de la phrase simple*, 242 :279, 2003. 64, 116, 117

Claude Coulombe. *Techniques d'amplification des données textuelles pour l'apprentissage profond*. PhD thesis, Téluq-université, 2020. 127, 128, 133, 137, 140, 144, 145, 146

Julien Cusin. L'apprentissage par l'échec commercial. *Vie sciences de l'entreprise*, (1) :33–53, 2008. 30

Clément Dalloux. Détection de l'incertitude et de la négation : un état de l'art. In *RECITAL 2017-18ème Rencontre des Étudiants Chercheurs en Informatique en Traitement Automatique des Langues*, pages 1–14, 2017. 65

Guillaume Daval-Frerot, Abdessalam Bouchekif, et Anatole Moreau. Epita at semeval-2018 task 1 : Sentiment analysis using transfer learning approach. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 151–155, 2018. 52

Kushal Dave, Steve Lawrence, et 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*, pages 519–528, 2003. 48

Henriëtte de Swart. Negation in a cross-linguistic perspective. In *Expression and Interpretation of Negation*, pages 1–53. Springer, 2010. 64

Edward L Deci et Richard M Ryan. Overview of self-determination theory : An organismic dialectical perspective. *Handbook of self-determination research*, pages 3–33, 2002. 72

EL Deci et RM Ryan. Intrinsic motivation and self-determination in human behavior : Springer science & business media. 1985. 72

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018. 15, 87, 92, 93, 94, 95, 162

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi : 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423. 17, 161

Xiaowen Ding, Bing Liu, et Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM, 2008. 51

Damien Dupré. *L'influence de produits innovants sur l'émotion des utilisateurs : une approche multi-componentielle*. PhD thesis, Université Grenoble Alpes, 2016. 31, 33

Chris Dyer, Victor Chahuneau, et Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 644–648, 2013. 142

Sergey Edunov, Myle Ott, Michael Auli, et David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv :1808.09381*, 2018. 144

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, et Michael Auli. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv :1908.05204*, 2019. 144

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4) : 169–200, 1992. 51

Hicham El Boukkouri. Ré-entraîner ou entraîner soi-même ? stratégies de pré-entraînement de bert en domaine médical. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des*

Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 29–42. ATALA ; AFCP, 2020. 13, 93

Kawin Ethayarajh. How contextual are contextualized word representations ? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv :1909.00512*, 2019. 87

Marzieh Fadaee, Arianna Bisazza, et Christof Monz. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv :1705.00440*, 2017. 142

Noura Farra, Elie Challita, Rawad Abou Assi, et Hazem Hajj. Sentence-level and document-level sentiment mining for arabic texts. In *2010 IEEE international conference on data mining workshops*, pages 1114–1119. IEEE, 2010. 58

Steven Y Feng, Aaron W Li, et Jesse Hoey. Keep calm and switch on ! preserving sentiment and fluency in semantic text exchange. *arXiv preprint arXiv :1909.00088*, 2019. 143, 146

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, et Eduard Hovy. Genaug : Data augmentation for finetuning text generators. *arXiv preprint arXiv :2010.01794*, 2020. 132, 134, 138, 139, 143, 147, 149, 156, 193

Elliot M Fielstein, Steven H Brown, et Theodore Speroff. Algorithmic de-identification of VA medical exam text for hipaa privacy compliance : preliminary findings. *Medinfo*, 1590, 2004. 61

John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957. 82

Susan T Fiske, KE Rosenblum, et TMC Travis. Social beings : A core motives approach to social psychology. *Social psychology*, 2009. 72

Vasiliki Foufi, Tatsawan Timakum, Christophe Gaudet-Blavignac, Christian Lovis, Min Song, et al. Mining of textual health information from reddit : Analysis of chronic diseases with extracted entities and their relations. *Journal of medical Internet research*, 21(6) :e12876, 2019. 49

Jibril Frej. *Incorporation de Connaissances a priori pour la Recherche d'Information Textuelle Neuronale*. PhD thesis, Université Grenoble Alpes, 2021. 90

Elia Gabarron, Enrique Dorronzoro, Octavio Rivera-Romero, et Rolf Wynn. Diabetes on twitter : a sentiment analysis. *Journal of diabetes science and technology*, 13(3) :439–444, 2019. 49

Thomas Gaillat, Annanda Sousa, Manel Zarrouk, et Brian Davis. Finsentia : Sentiment analysis in english financial microblogs. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, page 271, 2018. 42, 64

Núria Gala et Caroline Brun. Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions. In *19ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, volume 2, pages 495–502, 2012. 63, 116

Murthy Ganapathibhotla et Bing Liu. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 241–248, 2008. 133

Aldo Gangemi, Valentina Presutti, et Diego Reforgiato Recupero. Frame-based detection of opinion holders and topics : a model and a tool. *IEEE Computational Intelligence Magazine*, 9(1) :20–30, 2014. 52

Rosanna Garcia et Roger Calantone. A critical look at technological innovation typology and innovativeness terminology : a literature review. *Journal of Product Innovation Management : An international publication of the product development & management association*, 19(2) :110–132, 2002. 29

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, et Antonio Reyes. Semeval-2015 task 11 : Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 470–478, 2015a. 52

Debanjan Ghosh, Weiwei Guo, et Smaranda Muresan. Sarcastic or not : Word embeddings to predict the literal or sarcastic meaning of words. In *proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1003–1012, 2015b. 66

Anastasia Giannakidou. 64. negative and positive polarity items. In *Volume 2*, pages 1660–1712. De Gruyter Mouton, 2011. 64

Sébastien Gillot. Fouille d'opinions. In *Traitement du texte et du document*. 2010. 48

Praveen Kumar Badimala Giridhara, Chinmaya Mishra, Reddy Kumar Modam Venkataramana, Syed Saqib Bukhari, et Andreas Dengel. A study of various text augmentation techniques for relation classification in free text. *In proceedings of ICPRAM*, 3 :5, 2019. 138, 140

Natacha Goreux et Anne-Cécile Jeandrain. Analyse des freins et motivations des consommateurs à la pratique du shopping dans les univers virtuels. En ligne à l'adresse : https://scholar.google.fr/scholar?hl=fr&as_sdt=0%2C5&q=Analyse+des+freins+et+motivations+des+consommateurs+%7B%5C%60a%7D+la+pratique+du+shopping+dans+les+univers+virtuels&btnG=. 75

David Gotteland et Christophe Haon. *Développer un nouveau produit : méthodes et outils*. Pearson Education France, 2005. 30

David Gotteland et Christophe Haon. Nouveaux produits : les clefs de la réussite. *L'Expansion Management Review*, (3) :26–32, 2007. 30

Natalia Grabar, Cyril Grouin, Thierry Hamon, et Vincent Claveau. Recherche et extraction d'information dans des cas cliniques. présentation de la campagne d'évaluation Deft 2019. 2019. 42

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, et Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 140

Xiaoting Guo, Wei Yu, et Xiaodong Wang. An overview on fine-grained text sentiment analysis : Survey and challenges. In *Journal of Physics : Conference Series*, volume 1757, page 012038. IOP Publishing, 2021. 59

Thierry Hamon, Amel Fraisse, Patrick Paroubek, Pierre Zweigenbaum, et Cyril Grouin. Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (deft). In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*, 2015. 52

Tanjim Ul Haque, Nudrat Nawal Saber, et Faisal Muhammad Shah. Sentiment analysis on large scale amazon product reviews. In *2018 IEEE international conference on innovative research and development (ICIRD)*, pages 1–6. IEEE, 2018. 42

Zellig S Harris. Distributional structure. *Word*, 10(2-3) :146–162, 1954. 82

Vasileios Hatzivassiloglou et Kathleen McKeown. Predicting the semantic orientation of adjectives. In *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*, pages 174–181, 1997. 62

Kaiming He, Xiangyu Zhang, Shaoqing Ren, et Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 128

Abdel Rahman Hedar et M Doss. Mining social networks arabic slang comments. In *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013. 64

Paul A Herbig et Ralph L Day. Customer acceptance : the key to successful introductions of innovations. *Marketing Intelligence & Planning*, 1992. 70

Nicolas Hernandez et Brigitte Grau. Analyse thématique du discours : segmentation, structuration, description et représentation. In *Conférence CIDE'05*, 2002. 36

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *In Proceedings of the national academy of sciences*, 79 (8) :2554–2558, 1982. 87

Laurence R Horn. A natural history of negation. reissued with new introduction (2001), david Hume series, 1989. 64

Jeremy Howard et Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv :1801.06146*, 2018. 87, 88

Minqing Hu et 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. 133

William Huitt. Motivation to learn : An overview. *Educational psychology interactive*, 12, 2001. 71

Thien Ho Huong et Vinh Truong Hoang. A data augmentation technique based on text for vietnamese sentiment analysis. In *Proceedings of the 11th International Conference on Advances in Information Technology*, pages 1–5, 2020. 140

Md Amirul Islam, Sajal Halder, Md Ashraf Uddin, Uzzal Kumar Acharjee, et al. An efficient sentiment mining approach on social media networks. In *Emerging Technologies in Data Mining and Information Security*, pages 451–461. Springer, 2019. 49

Jahanzeb Jabbar, Iqra Urooj, Wu JunSheng, et Naqash Azeem. Real-time sentiment analysis on e-commerce application. In *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, pages 391–396. IEEE, 2019. 49

Agata Jackiewicz. Etudes sur les discours évaluatifs et d'opinion. *Études sur les discours évaluatifs et d'opinion*, pages 1–269, 2016. 50

Jitwongnan Jarukan. L'analyse des adjectifs axiologiques dans les ouvrages touristiques sur la thaïlande. *Sciences de l'Homme et Société*, 2014. 56, 70

Verena Joachim, Patrick Spieth, et Sven Heidenreich. Active innovation resistance : An empirical study on functional and psychological barriers to innovation adoption in different contexts. *Industrial Marketing Management*, 71 :95–107, 2018. 69

Alice Joiret et al. Mémoire de master en criminologie, à finalité spécialisée :" Étude des motivations et des freins à la consommation d'alcool et de cannabis chez les jeunes de 15 à 25 ans.". 2019. En ligne à l'adresse : https://matheo.uliege.be/bitstream/2268.2/7877/4/Etude%20des%20motivations%20et%20des%20freins%20%C3%A0%20la%20consommation%20d%27alcool%20et%20de%20cannabis%20chez%20les%20jeunes%20de%2015%20%C3%A0%2025%20ans.pdf. 75

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, et Tomas Mikolov. Fasttext.zip : Compressing text classification models. *arXiv preprint arXiv :1612.03651*, 2016. 86, 87, 92, 139

Jihen Karoui. *Détection automatique de l'ironie dans les contenus générés par les utilisateurs*. PhD thesis, Université de Toulouse 3 Paul Sabatier ; Faculté des Sciences Economiques, 2017. 53

Jihen Karoui, Farah Benamara, et Véronique Moriceau. *Détection automatique de l'ironie : Application à la fouille d'opinion dans les microblogs et les médias sociaux*. ISTE Group - Éditeur des Sciences, 2019. 67

Erick Kauffmann, David Gil, Jesús Peral, Antonio Ferrández, et Ricardo Sellers. A step further in sentiment analysis application in marketing decision-making. In *The International Research & Innovation Forum*, pages 211–221. Springer, 2019. 48

Emmanuelle Kelodjoue. Extraction d'opinions pour l'analyse multicritère à partir de corpus oraux transcrits : État de l'art. *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, pages 525–540, 2019. 62

Douglas T Kenrick, Vladas Griskevicius, Steven L Neuberg, et Mark Schaller. Renovating the pyramid of needs : Contemporary extensions built upon ancient foundations. *Perspectives on psychological science*, 5(3) :292–314, 2010. 72

Jacob Devlin Ming-Wei Chang Kenton et Lee Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 92

Catherine Kerbrat et al. L'enonciation de la subjectivité dans le langage. *A. Colin*, 1980. 50, 70

Adam Kilgarri. Senseval : An exercise in evaluating word sense disambiguation programs. In *proceedings of the first international conference on language resources and evaluation*, pages 581–588, 1998. 52

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, et Saif Mohammad. Nrc-canada-2014 : Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 437–442, 2014. 51

Mirella Kleijnen, Nick Lee, et Martin Wetzels. An exploration of consumer resistance to innovation and its antecedents. *Journal of economic psychology*, 30 (3) :344–357, 2009. 70, 71

Graham Klyne. Resource description framework (rdf) : Concepts and abstract syntax. *http ://www. w3. org/TR/2004/REC-rdf-concepts-20040210/*, 2004. 149

Sosuke Kobayashi. Contextual augmentation : Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv :1805.06201*, 2018. 129, 142, 143

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses : Open source toolkit for statistical machine

translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180, 2007. 94

Laurence Kohn et Wendy Christiaens. Les méthodes de recherches qualitatives dans la recherche en soins de santé : apports et croyances. *Reflets et perspectives de la vie économique*, 53(4) :67–82, 2014. 35

Oleksandr Kolomiyets, Steven Bethard, et Marie-Francine Moens. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics : human language technologies*, volume 2, pages 271–276. ACL ; East Stroudsburg, PA, 2011. En ligne à l'adresse : https://scholar.google.com/scholar?hl=fr&as_sdt=0%2C5&q=kolomiyets+synonym&btnG=. 137

Nihel Kooli et Erwan Pigneul. Analyse de sentiments à base d'aspects par combinaison de réseaux profonds : application à des avis en français. In *conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 1, pages 303–310, 2018. 64

Klaus Krippendorff. *The content analysis reader*. SAGE, 2009. 109

Anna Kruspe, Jens Kersten, Matti Wiegmann, Benno Stein, et Friederike Klan. Classification of incident-related tweets : Tackling imbalanced training data using hybrid cnns and translation-based data augmentation. In *Proceedings of the 27th Text REtrieval Conference (TREC 2018), Gaithersburg, Maryland, November 14*, volume 16, page 2018, 2018. 144

Tuire Kuisma, Tommi Laukkanen, et Mika Hiltunen. Mapping the reasons for resistance to internet banking : A means-end approach. *International Journal of Information Management*, 27(2) :75–85, 2007. 68, 69, 70

Mathieu Lafourcade, Nathalie Le Brun, et Alain Joubert. Vous aimez ?... ou pas ? likeit, un jeu pour construire une ressource lexicale de polarité. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, 2015. 62, 116

Saadi Lahlou. L'analyse lexicale. *Variances*, (3) :13–24, 1994. 36

Joseph Lark, Emmanuel Morin, et Sebastián Peña Saldarriaga. Canéphore : un corpus français pour la fouille d'opinion ciblée. In *Actes de la 22e conférence*

sur le Traitement Automatique des Langues Naturelles. Articles courts, pages 102–108, 2015. 51

Md Tahmid Rahman Laskar, Xiangji Huang, et Enamul Hoque. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5505–5514, 2020. 87

Pekka Laukkanen, Suvi Sinkkonen, et Tommi Laukkanen. Consumer resistance to internet banking : postponers, opponents and rejectors. *International journal of bank marketing*, 26(6) :440–455, 2008. 68, 70

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, et Didier Schwab. Flaubert : Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.302. 13, 15, 17, 88, 93, 94, 95, 96, 160, 162

Yann LeCun, Yoshua Bengio, et Geoffrey Hinton. Deep learning. *nature*, 521 (7553) :436–444, 2015. 87, 132

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, et Jaewoo Kang. Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) : 1234–1240, 2020. 93

Bohan Li, Yutai Hou, et Wanxiang Che. Data augmentation approaches in natural language processing : A survey. *arXiv preprint arXiv :2110.01852*, 2021a. 128

Min Li, Hui Zhao, Hao Su, YuRong Qian, et Ping Li. Emotion-cause span extraction : a new task to emotion cause identification in texts. *Applied Intelligence*, pages 1–13, 2021b. 48

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, et Hassan Sajjad. Findings of the first shared task on machine translation robustness. *arXiv preprint arXiv :1906.11943*, 2019. 94

Xin Li et Dan Roth. Learning question classifiers. In *COLING 2002 : The 19th International Conference on Computational Linguistics*, 2002. 134, 143

Zhenhao Li et Lucia Specia. Improving neural machine translation robustness via data augmentation : Beyond back translation. *arXiv preprint arXiv :1910.03009*, 2019. 144

Pingping Lin et Xudong Luo. A survey of the applications of sentiment analysis. *International Journal of Computer and Information Engineering*, 14(10) : 334–346, 2020. 50

Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1) :1–167, 2012. 48, 53, 58

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, et Veselin Stoyanov. Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*, 2019. 95, 162

Florian Loeser. *Modélisation probabiliste de l'influence des émotions sur l'acceptabilité des innovations*. PhD thesis, Université Grenoble Alpes, 2019. 30

Shayne Longpre, Yu Wang, et Christopher DuBois. How effective is task-agnostic data augmentation for pretrained transformers ? *arXiv preprint arXiv :2010.01764*, 2020. 134, 143

Cheng-Yu Lu, Jen-Shin Hong, et Samuel Cruz-Lara. Emotion detection in textual information by semantic role labeling and web mining techniques. In *Third Taiwanese-French Conference on Information Technology-TFIT*, 2006. 51

Edward Ma. Nlp augmentation. https ://github.com/makcedward/nlpaug, 2019. 154

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, et Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics : Human language technologies*, pages 142–150, 2011. 42

Craig Macdonald, Iadh Ounis, et Ian Soboroff. Overview of the trec 2007 blog track. In *TREC*, volume 7, pages 31–43, 2007. 52

William MacDougall. The energies of man. *New York : Scribners*, 1933. 72

Vivien Macketanz, Aljoscha Burchardt, et Hans Uszkoreit. Tq-autotest : Novel analytical quality measure confirms that deepl is better than google translate. *nd*, 2021. 148

Zied Mani et Inès Chouk. Drivers of consumers' resistance to smart products. *Journal of Marketing Management*, 33(1-2) :76–97, 2017. 68

Morgane Marchand. Fouille d'opinion : ces mots qui changent de polarité selon le domaine. In *Actes de CORIA*, pages 347–352, 2013. 66

Morgane Marchand. *Domaines et fouille d'opinion : une étude des marqueurs multi-polaires au niveau du texte*. PhD thesis, Université Paris Sud-Paris XI, 2015. 66

Vukosi Marivate et Tshephisho Sefara. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 385–399. Springer, 2020. 137, 138, 140, 144, 147

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, et Benoît Sagot. Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*, 2019. 15, 88, 93, 95

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, et Benoît Sagot. Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020a. 15, 17, 88, 93, 95, 161

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoan Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Benoît Sagot, et Djamé Seddah. Les modèles de langue contextuels camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement. In *JEP-TALN-RECITAL 2020-33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 54–65. ATALA ; AFCP, 2020b. 17, 161

Sigrid Maurel, Paolo Curtoni, et Luca Dini. A hybrid method for sentiment analysis. In *Actes d'INFORSID*, 2008. 64

Aurélien Max et Guillaume Wisniewski. Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010. 131

Diana G Maynard et Mark A Greenwood. Who cares about sarcastic tweets ? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of LREC 2014*. ELRA, 2014. 66

Bryan McCann, James Bradbury, Caiming Xiong, et Richard Socher. Learned in translation : Contextualized word vectors. *arXiv preprint arXiv :1708.00107*, 2017. 88

John McCrae, Dennis Spohr, et Philipp Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *Extended Semantic Web Conference*, pages 245–259. Springer, 2011. 14, 150

David McHugh, Sarah Shaw, Travis R Moore, Leafia Zi Ye, Philip Romero-Masters, et Richard Halverson. Uncovering themes in personalized learning : Using natural language processing to analyze school interviews. *Journal of Research on Technology in Education*, 52(3) :391–402, 2020. 22

Yves Mercadier. *Classification automatique de textes par réseaux de neurones profonds : application au domaine de la santé*. PhD thesis, Université de Montpellier, 2020. 13, 77, 144, 146, 192

Tomas Mikolov, Kai Chen, Greg Corrado, et Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*, 2013a. 13, 82, 83, 84, 139

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, et Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b. En ligne à l'adresse : https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf. 82, 83, 84, 87, 92

Tomáš Mikolov, Wen-tau Yih, et Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics : Human language technologies*, pages 746–751, 2013c. 83

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, et Armand Joulin. Advances in pre-training distributed word representations. *arXiv preprint arXiv :1712.09405*, 2017. 87

Ana Minanovic, Hrvoje Gabelica, et Živko Krstić. Big data and sentiment analysis using knime : Online reviews vs. social media. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1464–1468. IEEE, 2014. 50

Jacques Moeschler. The pragmatic aspects of linguistic negation : Speech act, argumentation and pragmatic inference. *Argumentation*, 6(1) :51–76, 1992. 64

M Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, et José M Perea-Ortega. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18) : 7250–7257, 2013. 62

Andrew Moore et Paul Rayson. Lancaster a at semeval-2017 task 5 : Evaluation metrics matter : predicting sentiment from financial news headlines. *arXiv preprint arXiv :1705.00571*, 2017. 42, 64

Ahmed Mourad et Kareem Darwish. Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 55–64, 2013. 62

Matthijs Mulder, Anton Nijholt, Marten Den Uyl, et Peter Terpstra. A lexical grammatical implementation of affect. In *International Conference on Text, Speech and Dialogue*, pages 171–177. Springer, 2004. 51

Hala Mulki, Hatem Haddad, et Mourad Gridach. Polarity analysis of non figurative tweets : Tw-star participation on actes de deft 2017. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, page 92, 2017. 51, 62, 64

Claude Muller. La négation en français. *Syntaxe, sémantique et éléments de comparaison avec les autres langues romanes, Genève : Droz*, 1991a. 64

Claude Muller. *La négation en français : syntaxe, sémantique et éléments de comparaison avec les autres langues romanes*. Librairie Droz, 1991b. 64

Henry A Murray. Explorations in personality : A clinical and experimental study of fifty men of college age, new york (oxford university press) 1938. 1938. 72

Tetsuya Nasukawa et Jeonghee Yi. Sentiment analysis : Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, 2003. 48

Julien Nelson. *Contribution à l'analyse prospective des usages dans les projets d'innovation*. PhD thesis, Arts et métiers ParisTech, 2011. 29

Mike Donald Tapi Nzali. *Analyse des médias sociaux de santé pour évaluer la qualité de vie des patientes atteintes d'un cancer du sein*. PhD thesis, 2017. 13, 77

OECD. Guidelines for collecting and interpreting innovation data. *A joint publication of OECD and Eurostat, Organization for Economic Co-Operation and Development. Statistical Office of the European Communities*, 2005. 28

Iadh Ounis, Craig Macdonald, et Ian Soboroff. Overview of the trec-2008 blog track. Technical report, Glasgow university in UK, 2008. 51, 52

Bo Pang et Lillian Lee. A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*, 2004. 134

Bo Pang, Lillian Lee, et Shivakumar Vaithyanathan. Thumbs up ? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002. 51, 65

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, et Quoc V Le. SpecAugment : A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv :1904.08779*, 2019. 128

Manojkumar Parmar, Bhanurekha Maturi, Jhuma Mallik Dutt, et Hrushikesh Phate. Sentiment analysis on interview transcripts : An application of nlp for quantitative analysis. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1063–1068. IEEE, 2018. 22

Patrick Paroubek, Cyril Grouin, Patrice Bellot, Vincent Claveau, Iris Eshkol-Taravella, Amel Fraisse, Agata Jackiewicz, Jihen Karoui, Laura Monceaux, et Juan-Manuel Torres-Moreno. Deft2018 : recherche d'information

et analyse de sentiments dans des tweets concernant les transports en île de france. *In Actes de DEFT*, 2018. 58

Hélène Marie Louise Pasquier. *Définir l'acceptabilité sociale dans les modèles d'usage : vers l'introduction de la valeur sociale dans la prédiction du comportement d'utilisation*. PhD thesis, Université Rennes 2, 2012. 33

Yifan Peng, Shankai Yan, et Zhiyong Lu. Transfer learning in biomedical natural language processing : an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv :1906.05474*, 2019. 87

Jeffrey Pennington, Richard Socher, et 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. 84, 85, 87, 92, 139

Marie-Paule Péry-Woodley, Nicholas Asher, Patrice Enjalbert, Farah Benamara, Myriam Bras, Cécile Fabre, Stéphane Ferrari, Lydia-Mai Ho-Dac, Anne Le Draoulec, Yann Mathet, et al. Annodis : une approche outillée de l'annotation de structures discursives. In *Actes de TALN 2009 (Conférence sur le Traitement Automatique des Langues Naturelles)*, page paper_TALN_52, 2009. 54

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, et Luke Zettlemoyer. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv :1802.05365*, 2018. 87, 88, 92

Michelle-Eve Pilon-Caron. Comprendre les freins et les motivations dans l'utilisation des services de l'entreprise airbnb. 2015. En ligne à l'adresse : https://core.ac.uk/display/77618256. 75

Pascal Pizelle, J Hoffmann, C Verchère, et M Aubouy. Innover par les usages. *Grenoble : Éditions d'Innovation*, 2014. 36, 37

Robert Plutchik et Henry Kellerman. Biological foundations of emotion. vol 3 of emotion : Theory, research, and experience, 1986. 51

Blake D Poland. Transcription quality as an aspect of rigor in qualitative research. *Qualitative inquiry*, 1(3) :290–310, 1995. 36

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, et Ion Androutsopoulos. Semeval-2015 task 12 : Aspect based sentiment analysis. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 486–495, 2015. 59

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5 : Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30, 2016. 59

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, et Rada Mihalcea. Beneath the tip of the iceberg : Current challenges and new directions in sentiment analysis research. *In proceedings of IEEE Transactions on Affective Computing*, 2020. 59

Timothée Poulain et Victor Connes. Deft 2021 : Évaluation automatique de réponses courtes, une approche basée sur la sélection de traits lexicaux et augmentation de données (deft 2021 : Automatic short answer grading, a lexical features selection and data augmentation based approach). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, pages 31–40, 2021. 140

Alec Radford, Karthik Narasimhan, Tim Salimans, et Ilya Sutskever. Improving language understanding by generative pre-training. 2018. En ligne à l'adresse : https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf. 88

Hichem Rahab, Abdelhafid Zitouni, et Mahieddine Djoudi. Sana : Sentiment analysis on newspapers comments in algeria. *Journal of King Saud University-Computer and Information Sciences*, 33(7) :899–907, 2019. 42

Callen Rain. Sentiment analysis in amazon reviews using probabilistic machine learning. *Swarthmore College*, 2013. 42

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, et Percy Liang. Squad : 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv :1606.05250*, 2016. 96

Sudha Ram. A model of innovation resistance. *ACR North American Advances*, 1987. 68, 69

Sundaresan Ram et Jagdish N Sheth. Consumer resistance to innovations : the marketing problem and its solutions. *Journal of consumer marketing*, 6(2) : 5–14, 1989. 68, 69, 70

Gaetan Raoul Raoul. Juicero ferme ses portes après l'échec de son presse-jus connecté, 2017. En ligne à l'adresse : https://www.objetconnecte.com/juicero-femerture-jus-0409. 13, 30, 31

Chetanya Rastogi, Nikka Mofid, et Fang-I Hsiao. Can we achieve more with less ? exploring data augmentation for toxic comment classification. *arXiv preprint arXiv :2007.00875*, 2020. 134, 135

Kumar Ravi et Vadlamani Ravi. A survey on opinion mining and sentiment analysis : tasks, approaches and applications. *Knowledge-based systems*, 89 : 14–46, 2015. 59

Eshrag Refaee et Verena Rieser. ilab-edinburgh at semeval-2016 task 7 : A hybrid approach for determining sentiment intensity of arabic twitter phrases. In *Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016)*, pages 474–480, 2016. 64

Véronique Rey et Núria Gala. Les mots de bouche à oreilles : le cas de polymots, 2011. En ligne à l'adresse : https://hal.archives-ouvertes.fr/hal-03198400. 63

Georgios Rizos, Konstantin Hemker, et Björn Schuller. Augment to prevent : short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 991–1000, 2019. 140

Everett M Rogers. Diffusion of innovations. *New York : The Free Press of Glencoe*, 1962. 31

Stuart Rose, Dave Engel, Nick Cramer, et Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining : applications and theory*, 1 : 1–20, 2010. 138

Sara Rosenthal, Noura Farra, et Preslav Nakov. SemEval-2017 task 4 : Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August 2017. Association for Computational Linguistics. 52

David E Rumelhart, Geoffrey E Hinton, et RJ Williams. Learning internal representations by error propagation. *Parallel distributed processing : Explorations in the microstructure of cognition, Volume 1 : Foundations*, 1986. 87

José Saias. Senti. ue : Tweet overall sentiment classification approach for semeval-2014 task 9. In Proceedings of the 8th International Workshop on Semantic E valuation (Semeval 2014), 2014. 52

Antoine Saporta, Corentin Dancette, et Matthieu Cord. Generative adversarial networks. 2021. En ligne à l'adresse : http://webia.lip6.fr/~dancette/deep-learning/assets/TP9-10.pdf. 131

P Sarrazin et D Trouilloud. Comment motiver les élèves à apprendre ? les apports de la théorie de l'autodétermination. *Comprendre les apprentissages, sciences cognitives et éducation*, 2 :123–141, 2006. 13, 73

Roger C Schank et Robert P Abelson. *Scripts, plans, goals, and understanding : An inquiry into human knowledge structures*. Psychology Press, 2013. 72

Karel Šebesta, Zuzanna Bedrichová, Katerina Šormová, Barbora Štindlová, Milan Hrdlicka, Tereza Hrdlicková, Jirı Hana, Vladimır Petkevic, Tomáš Jelınek, Svatava Škodová, et al. Czesl grammatical error correction dataset (czesl-gec). *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University*, 2017. 131

Dmitriy Selivanov. Glove word embeddings, 2020. En ligne à l'adresse : http://text2vec.org/glove.html. 85

Rico Sennrich, Barry Haddow, et Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv :1511.06709*, 2015a. 144

Rico Sennrich, Barry Haddow, et Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv :1508.07909*, 2015b. 94

Gilles Sérasset. Dbnary : Wiktionary as a lmf based multilingual rdf network. In *Proceedings of Language Resources and Evaluation Conference, LREC 2012*, 2012. 149, 150

Gilles Sérasset et Andon Tchechmedjiev. Dbnary : Wiktionary as linked data for 12 language editions with enhanced translation relations. In *3rd Workshop on Linked Data in Linguistics : Multilingual Knowledge Resources and Natural Language Processing*, pages 67–71, 2014. 149

Christophe Servan, Zied Elloumi, Hervé Blanchon, et Laurent Besacier. Word2vec vs dbnary ou comment (ré) concilier représentations distribuées et réseaux lexico-sémantiques ? le cas de l'évaluation en traduction automatique. In *Actes de TALN 2016*, 2016. 151

Sam Shleifer. Low resource text classification with ulmfit and backtranslation. *arXiv preprint arXiv :1903.09244*, 2019. 144

Amira Shoukry et Ahmed Rafea. Sentence-level arabic sentiment analysis. In *Proceedings of International Conference on Collaboration Technologies and Systems 2012 (CTS)*, pages 546–550. IEEE, 2012. 58

Philippe Silberzahn, Jean-Yves Prax, Bernard Buisson, et Vincent Sincholle. Innovations radicales : le pari de l'intrapreneuriat. *L'Expansion Management Review*, (2) :66–71, 2007. 29

Doriane Simonnet. *Vers un outil sémantique d'autocodage qualitatif pour l'évaluation de l'acceptabilité des innovations*. Theses, Université Grenoble Alpes [2020-....], March 2022. URL https://tel.archives-ouvertes.fr/tel-03689721. 192

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, et 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, 2013. 96, 133

Julien Soler, Florian Loeser, Charlotte Decorps, Doriane Simonnet, et Niklas Henke. Petit traité pour déjouer les 10 grands pièges de l'innovation. (6-17) :52. En ligne à l'adresse : https://leslivresblancs.fr/livre/entreprise/innovation/petit-traite-pour-dejouer-les-10-grands-pieges-de-linnovation. 27, 31, 32

Carlo Strapparava et Rada Mihalcea. Semeval-2007 task 14 : Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, 2007. 52

Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect : an affective extension of wordnet. In *Proceedings of LREC*, volume 4, pages 1083–1086. Citeseer, 2004. 137

Pedro Javier Ortiz Suárez, Benoît Sagot, et Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures.

In *Proceedings of 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 2019. 95

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, et Sivanesan Sangeetha. Ammus : A survey of transformer-based pretrained models in natural language processing. *arXiv e-prints*, pages arXiv–2108, 2021. 93

Ilya Sutskever, Oriol Vinyals, et Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 89

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, et Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37 (2) :267–307, 2011. 62

Jennifer R Talevich, Stephen J Read, David A Walsh, Ravi Iyer, et Gurveen Chopra. Toward a comprehensive taxonomy of human motives. *PloS one*, 12 (2) :e0172279, 2017. 72

Luke Taylor et Geoff Nitschke. Improving deep learning with generic data augmentation. In *Proceedings of IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1542–1547. IEEE, 2018. 128

Matt Thomas, Bo Pang, et Lillian Lee. Get out the vote : Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics, 2006. 22, 42

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Proceedings of LREC*, volume 2012, pages 2214–2218. Citeseer, 2012. 94

Juan Manuel Torres-Moreno, Marc El-Bèze, Fréderic Béchet, et Nathalie Camelin. Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ? application au défi deft 2007. *Actes du troisième DÉfi Fouille de Textes*, page 129, 2007. 52

Peter D Turney. Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*, 2002. 51

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, et Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 13, 89, 91

Matthieu Vernier. *Analyse à granularité fine de la subjectivité*. PhD thesis, Université de Nantes, 2011. 50, 51, 56, 70

Laurence Vidrascu. *Analyse et détection des émotions verbales dans les interactions orales*. PhD thesis, Université Paris Sud-Paris XI, 2007. 51

Dong Wang et Yang Liu. Opinion summarization on spontaneous conversations. *Computer Speech & Language*, 34(1) :61–82, 2015. 49

William Yang Wang et Diyi Yang. That's so annoying ! ! ! : A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, 2015. 139, 140, 145

Jason Wei et Kai Zou. Eda : Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv :1901.11196*, 2019. 16, 129, 133, 134, 135, 137, 143, 147, 152, 153, 157

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, et Edouard Grave. Ccnet : Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv :1911.00359*, 2019. 95

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, et Melanie Martin. Learning subjective language. *Computational linguistics*, 30(3) :277–308, 2004. 51

Janyce Wiebe et al. Learning subjective adjectives from corpora. *In Proceedings of the 70th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, 20 :735–740, 2000. 62

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, et Andrés Montoyo. A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 60–68, 2010. 65

Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adriane Boyd, Detmar Meurers, Andrea Abel, et Jirka Hana. Merlin : An online trilingual learner corpus empirically grounding the european reference levels in authentic learner data. In *Proceedings of ICT for Language Learning 2013, Florence, Italy. Libreria universitaria. it Edizioni*, 2013. 131

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, et Songlin Hu. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer, 2019. 143

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system : Bridging the gap between human and machine translation. *arXiv preprint arXiv :1609.08144*, 2016. 92

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, et Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv :1904.12848*, 2019. 135, 140

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, et Andrew Y Ng. Data noising as smoothing in neural network language models. *arXiv preprint arXiv :1703.02573*, 2017. 135, 137

Shweta Yadav, Asif Ekbal, Sriparna Saha, et Pushpak Bhattacharyya. Medical sentiment analysis using social media : towards building a patient assisted system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. 42

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, et Quoc V Le. Xlnet : Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 134

Ainur Yessenalina, Yisong Yue, et Claire Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1046–1056, 2010. 58

Torsten Zesch. Measuring contextual fitness using error contexts extracted from the wikipedia revision history. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 529–538, 2012. 131

Dongxu Zhang et Dong Wang. Relation classification via recurrent neural network. *arXiv preprint arXiv :1508.01006*, 2015. 138

Xiang Zhang, Junbo Zhao, et Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28 :

649–657, 2015. En ligne à l'adresse : https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf. 137

Tao Zhou. An empirical examination of initial trust in mobile banking. *Internet Research*, 21(5) :527–540, 2011. 70

Xiaodan Zhu et Gerald Penn. Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*, 2006. 49

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, et Sanja Fidler. Aligning books and movies : Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. 91

Hend Zouari. French AXA insurance word embeddings : Effects of fine-tuning bert and camembert on AXA france's data. En ligne à l'adresse : https://www.diva-portal.org/smash/get/diva2:1476590/FULLTEXT01.pdf. 95