

Louvain School of Management

Analyse du potentiel de la traduction pour réaliser une analyse de sentiment dans une autre langue que l'anglais

**Le cas du transfert entre l'anglais et le français lors
d'un apprentissage supervisé**

Auteur : Massagé Alexis

Promoteur(s) : Fouss François

Année académique 2021-2022

Master en Business Analytics

Remerciements

En préambule de ce mémoire, je tiens à exprimer ma gratitude à toutes les personnes qui m'ont aidé et soutenu lors de la réalisation de ce mémoire et sans qui ce travail n'aurait probablement jamais vu le jour. Je remercie tout d'abord le promoteur de ce mémoire, le Professeur François Fouss, pour sa confiance, sa disponibilité et ses conseils durant la rédaction de ce mémoire.

Je remercie également tous les membres académiques qui m'ont permis d'acquérir les bases théoriques et expériences nécessaires à l'accomplissement de ce mémoire ainsi que la communauté en ligne de data scientist pour le partage de leurs connaissances.

Ensuite, je tiens à remercier ma compagne pour sa confiance, sa rigueur et sa patience. Et enfin, mes parents et mes amis pour leurs conseils, soutiens, encouragements et leur participation à la relecture de ce mémoire.

Résumé

L'analyse de sentiment est l'application du Natural Language Processing (NLP) s'attachant à l'étude et la détection de la polarité du langage humain. Dans un contexte d'expression de plus en plus forte sur internet et face à l'importance de comprendre le sentiment des utilisateurs d'internet, cette application s'est manifestement développée au cours des dernières années. Néanmoins, la majorité de l'attention s'est tournée vers l'anglais, laissant les autres langues en manque de ressources pour pratiquer d'aussi bonnes analyses.

Nous proposons, dans ce mémoire, d'évaluer la possibilité de recourir à la traduction automatique pour procéder à un transfert entre une langue riche en ressources comme l'anglais, vers une langue plus pauvre comme le français. Nous comparons deux types de transfert. Le premier (a) est un *transfert d'environnement*, amenant le français dans l'environnement riche de l'anglais, qui se manifeste concrètement par la traduction des observations à prédire, vers l'anglais, pour profiter des modèles existants. Le deuxième (b) est un *transfert de ressources* de l'anglais vers le français, qui se manifeste concrètement par la traduction de datasets de l'anglais vers les autres langues, pour créer des modèles monolingues dans les autres langues. Nous appliquons une méthodologie en trois étapes, dont la première est empruntée à la littérature, reposant sur du machine learning supervisé pour déterminer le transfert permettant de restituer le sentiment le plus efficacement possible. Notre travail montre, dans la continuité de la littérature, que la traduction détruit peu l'information relative au sentiment. Cependant, les modèles entraînés sur des données originales prédisant la polarité d'observations traduites et ceux entraînés sur des données traduites prédisant la polarité d'observations originales, subissent une baisse de performance non négligeable. Toutefois, si un type de transfert devait être préconisé, ce serait celui de la traduction de datasets de l'anglais, vers les autres langues, permettant de développer des modèles dans ces autres langues. Enfin, si des données étiquetées existent dans les autres langues, même en nombre limité, leur mélange avec les données d'entraînement venant de la traduction pourrait permettre de corriger les problèmes de performances.

Table des matières

Introduction	1
Méthodologie	2
Partie 1 : Revue de la littérature	3
1. Natural Language Processing	3
1.1. Contextualisation.....	3
1.2. Historique	3
2. Analyse de sentiment.....	5
2.1. Natural Language Processing versus Natural Language Understanding	5
2.2. Définition et applications	5
2.3. Types d'analyse de sentiment.....	6
2.4. Méthodes lexicales (lexicon-based)	7
2.5. Méthodes d'apprentissage (Machine learning-based).....	8
2.6. Machine learning versus lexique	9
3. Prétraitement.....	11
3.1. Tokenisation	11
3.2. Normalisation	12
3.3. Vectorisation	13
4. Inégalité entre les langues.....	15
4.1. Engouement pour l'anglais.....	15
4.2. Types de transfert, traduction et positionnement dans la littérature	17
5. Conclusion partie I	19
Partie 2 : Étude quantitative	20
1. Objectif.....	20
2. Méthodologie.....	21
2.1. Méthodologie en 3 étapes.....	21
2.2. Définition des datasets et traduction automatique	25
2.3. Prétraitement	27
2.4. Différents classifieurs et paramètres	28
3. Analyse des résultats	30
3.1. Étape 1.....	30
3.2. Prétraitement	32
3.3. Étape 2.....	33

3.4. Étape 3.....	35
4. Discussion.....	38
5. Conclusion partie II	39
<i>Conclusion.....</i>	<i>41</i>
<i>Recul critique et recommandations</i>	<i>42</i>
<i>Bibliographie</i>	<i>43</i>
<i>Annexes.....</i>	<i>47</i>

Liste des figures

Figure 1 : Le NLP et son sous-ensemble le NLU. (Olivier, 2021)	5
Figure 2 : <i>Approches existantes pour l'analyse de Sentiment (Furqan et al., 2019)</i>	6
Figure 3 : Fonctionnement des approches lexicales et d'apprentissages (Taboada, 2016).....	7
Figure 4 : Nombre d'articles basés sur les différentes approches entre 2010 et 2013. (Medhat et al., 2014).....	10
Figure 5 : Nombre de publications relatives à l'analyse de sentiment en anglais et en d'autres langues (Medhat et al., 2014)	15
Figure 6 : Nombre de publications relatives à l'analyse de sentiment en anglais (à gauche) et multilingue (à droite) (Dashtipour et al., 2016)	15
Figure 7 : Schéma de l'étape 1 de notre méthodologie	21
Figure 8 : Types de transferts	22
Figure 9 : Schéma de l'étape 2a de notre méthodologie	22
Figure 10 : Schéma de l'étape 3a de notre méthodologie	23
Figure 11 : Schéma de l'étape 2b de notre méthodologie	24
Figure 12 : Schéma de l'étape 3b de notre méthodologie	24
Figure 13 : Distribution des datasets de films (à gauche) et produits (à droite) labellisés en deux et cinq classes	26
Figure 14 : Nuages de mots des mots les plus fréquents de chaque datasets	26
Figure 15 : Amélioration selon le type de n-gramme choisi, à gauche en deux classes et à droite en cinq classes	32
Figure 16 : Exemple de matrices déséquilibrées entre les faux positifs et faux négatifs	37

Liste des tableaux

Tableau 1 : Accuracy des modèles sur des observations originales anglophones.....	30
Tableau 2 : Accuracy des modèles sur des observations originales francophones	30
Tableau 3 : Accuracy selon le type de dataset et la langue	31
Tableau 4 : Accuracy selon le type de classifieur et la langue	31
Tableau 5 : Accuracy selon le type de n-gramme et de dataset	32
Tableau 6 : Accuracy selon le type de dataset et le type de normalisation	33
Tableau 7 : Accuracy selon le type de dataset et le type de vectorisation	33
Tableau 8 : Comparaison des accuracy lors de l'étape 2a.....	34
Tableau 9 : Comparaison des accuracy lors de l'étape 2b	34
Tableau 10 : Comparaison des accuracy lors de l'étape 3a.....	35
Tableau 11 : Comparaison des accuracy lors de l'étape 3b	36
Tableau 12 : Synthèse des écarts de performance entre les différentes étapes	37

Introduction

Depuis que l'écriture existe, les opinions de l'auteur sont souvent reflétées dans ce qu'il écrit. À l'heure d'internet et des réseaux sociaux, où il y a presque une injonction à partager son avis, il semble pertinent de s'intéresser à ceux-ci et plus particulièrement au sentiment caché derrière ces avis. Sur de nombreux sites, il nous est demandé de décrire la performance perçue de n'importe quel produit et il est maintenant rare de ne pas pouvoir évaluer quelque chose que nous avons acheté ou pas d'ailleurs. Les avis peuvent être exprimés de manière directe (avec une notation) ou de manière indirecte (avec un commentaire), et parfois les deux sont présents.

Une analyse de sentiment permet de déterminer la polarité d'un avis lorsqu'il a été exprimé de manière indirecte, lorsque seul le texte est disponible. Cette analyse peut catégoriser le sentiment de plusieurs façons, un sentiment peut être positif, négatif, voir neutre, mais il peut aussi être catégorisé d'une autre manière, comme par une palette d'émotions ou une notation. L'analyse de sentiment rassemble les techniques tentant de déterminer les sentiments exprimés dans le langage humain. Elle permet de détecter le sentiment vis-à-vis d'une publication sur les réseaux sociaux, le mécontentement relatif à un produit, voire même de détecter les propos haineux et le harcèlement. Cependant, la majorité de l'intérêt de la recherche et du public, ainsi que les efforts de développement, est tourné vers l'analyse de l'anglais. Qu'en est-il alors des autres langues ? Petit à petit, des ressources et modèles sont bâtis pour les autres langues, mais l'environnement n'est pas encore aussi optimal que celui de l'anglais. La proposition que ce mémoire explore, déjà étudiée dans certains articles, vise à procéder à un transfert entre une langue riche comme l'anglais vers une langue plus pauvre comme le français et via la traduction automatique. Soit en transférant le français dans l'environnement de l'anglais, soit en transférant les ressources de l'anglais vers le français.

Dans ce travail, nous évaluerons la maturité de la traduction automatique pour être utilisée comme outil de transfert entre langue riche et pauvre. Nous souhaitons plus particulièrement analyser l'ampleur de la perte d'information - si elle existe - après traduction, son impact sur les résultats d'une classification impliquant des observations traduites et non-traduites et déterminer le type de transfert permettant le moins de perte d'accuracy. Enfin, cette étude est exploratoire et a pour but de proposer des pistes pour utiliser la traduction automatique pour procéder à une analyse de sentiment dans d'autres langues.

Méthodologie

Afin de pouvoir répondre à notre problématique, nous commencerons par une revue de la littérature abordant ce qu'est le natural language processing (NLP) ou, en français, le traitement automatique des langues (TAL). Ensuite, nous nous attarderons sur ce qu'est une analyse de sentiment, pourquoi nous avons recours à ce genre d'analyse et les différentes façons de procéder. Nous nous intéresserons aux approches dites d'apprentissage (machine learning) et celles lexicales (basées sur un lexique de polarité) utilisées lors d'une analyse de sentiment.

Nous nous pencherons par la suite sur le prétraitement à faire en amont d'une telle analyse. Ce sont les techniques permettant de préparer les données avant de procéder à notre analyse de sentiment. Nous y trouvons, entre autres, la tokenisation, la vectorisation, la lemmatisation, la suppression des stop-words, qui sont autant de techniques avec différentes possibilités.

Nous discuterons aussi de la position de l'anglais et des autres langues dans le champ de l'analyse de sentiment ainsi que des différentes façons de traiter les autres langues. Nous nous intéresserons aussi à la traduction, les manières de l'utiliser pour opérer un transfert entre l'anglais et le français et son impact sur l'analyse de sentiment.

Dans la deuxième partie de ce travail, nous adopterons une approche quantitative, dans laquelle, nous créerons des modèles dont le but est de prédire la note d'un utilisateur à partir du commentaire de ce dernier. Cette analyse se fera à partir de commentaires de film et de produits vendus sur Amazon. Une méthodologie en 3 étapes sera mise en place, une première visera à tester la capacité d'un modèle entraîné sur des observations traduites à prédire d'autres observations traduites pour vérifier la qualité de la traduction dans une analyse de sentiment. Les deuxièmes et troisièmes étapes tentent de déterminer le meilleur type de transfert. Soit (a) le *transfert d'environnement* en testant la capacité d'un modèle entraîné sur des observations originales à prédire la polarité sur base de observations traduites. Soit (b) le *transfert de ressources* en testant la capacité d'un modèle entraîné sur des observations traduites à prédire la polarité sur base de observations originales. Nous mettrons en place différents classifieurs et techniques de prétraitement. Nous évaluerons les performances de l'analyse de sentiment entre les différents modèles selon leur accuracy. Finalement, nous confronterons nos résultats pour tenter de déterminer si l'utilisation de observations traduites affecte négativement – et dans quelle mesure – l'accuracy des prédictions et le type de transfert à privilégier pour maximiser l'accuracy des prédictions.

Partie 1 : Revue de la littérature

1. Natural Language Processing

1.1. Contextualisation

L'intelligence artificielle (IA) fait de plus en plus son nid dans notre quotidien, elle nous accompagne dans notre vie professionnelle jusqu'à dans notre vie privée. Elle est présente à la maison, au travail, dans la voiture, mais aussi à des endroits où nous ne la voyons pas, pour des applications dont nous n'avons pas idée. Inexorablement, l'IA est devenue et deviendra encore plus incontournable dans de nombreux domaines : le transport, la santé, le commerce, l'assistant personnel, l'industrie, l'environnement, la finance, la défense, etc. (L'OBS, 2021).

Dans bon nombre de ces domaines, il est devenu impératif que les machines puissent traiter le langage humain. Non pas seulement pour interagir avec l'humain, mais aussi pour analyser son langage et créer à partir de celui-ci. Les domaines d'application pratique sont nombreux : correction orthographique, assistants vocaux et chatbots, traduction automatique, génération de texte, etc. (Analytics & Insights, 2021).

Tout cela peut être réuni dans une discipline à la croisée des chemins entre la linguistique, l'informatique et l'intelligence artificielle : le natural language processing (NLP) ou en français, traitement automatique des langues (TAL). Nous parlons de langage naturel pour qualifier les langages humains, possédant une syntaxe, une grammaire et des ambiguïtés, à l'inverse des langages formels comme les mathématiques ou les langages informatiques, ne laissant aucune place à l'ambiguïté (Inbenta, 2021).

1.2. Historique

L'histoire du NLP commence en même temps que celle de l'informatique, c'est-à-dire en pleine deuxième guerre mondiale, lors de l'invention des premiers ordinateurs. Elle suit aussi l'histoire de l'intelligence artificielle dont les bases ont été posées par Turing et Von Neumann en 1950. La suite découle encore du domaine militaire avec l'expérience Georgetown-IBM en 1954, ayant pour but de traduire une soixantaine de phrases russes vers l'anglais. Malgré un choix des phrases à traduire relevant du cherry picking, les auteurs prétendaient à l'époque que 3 à 5 ans plus tard, c'est-à-dire avant 1960, la traduction automatique ne serait plus un problème. L'idée étant que la machine réussirait prochainement à comprendre le langage humain puis à le traduire. Il n'en sera finalement rien (Conseil de l'Europe, 2021) (Poibeau, 2014).

En 1950, Alan Turing avait d'ailleurs défini dans son ouvrage « Computing machinery and intelligence », un critère, dit de Turing, pour mesurer la maturité d'une IA : celui de se faire passer pour un humain dans un échange de message écrit (aussi appelé : the imitation game) (de Séguin, 2019; Olivier, 2021). En 1966 apparaît ELIZA, un automate conversationnel simulant un psychothérapeute (le premier chatbot). En réalité, il ne faisait que reformuler les phrases du patient pour avancer dans les questions (- J'ai mal à la tête. - Pourquoi avez-vous mal à la tête ?). La même année, le rapport ALPAC constate que les progrès ne furent pas à la hauteur des espérances et que la complexité des langues avait été sous-estimée, réduisant l'engouement de la recherche. Pendant les années 60-70, d'autres chatbots verront le jour, mais leurs capacités resteront très limitées. À la deuxième moitié des années 70, c'est tout le domaine de l'intelligence artificielle qui vécut son premier hiver.

L'espoir reprend avec la montée des systèmes experts et l'introduction des algorithmes de machine learning. Mais à cette époque, c'est une nouvelle déception, les résultats n'étaient pas satisfaisants compte tenu de la complexité et du coût de développement (Conseil de l'Europe, 2021). À la fin des années 80, l'ambition de faire comprendre à la machine le langage humain est revue à la baisse et l'objectif devient plutôt de détecter et exploiter l'information pertinente. (Poibeau, 2014).

À partir de la moitié des années 90, c'est l'augmentation de la puissance de calcul et l'accès à de grandes quantités de données qui permettent véritablement au machine learning de tenir ses promesses. Après 2010 et en partie grâce au deep learning, nous observons des IA battre des humains dans des épreuves de lecture et compréhension, des assistants vocaux et des traducteurs performants ou enfin le lancement de GPT-3 par OpenAI (Fenner, 2018). Cependant, malgré les reproches faites au test de Turing d'être irréalisable et inutile, à l'heure actuelle, aucune IA n'a été capable de le passer, car même GPT-3 peut répondre à une question comme « qui était le président des États-Unis en 1600 » : la reine Elizabeth I (Joarson, 2020). En effet, GPT-3 cherche avant tout à répondre à la question et ne comprend pas le piège. Voilà où nous en sommes, un traitement automatique des langues encore très imparfait, mais néanmoins déjà très utile.

2. Analyse de sentiment

2.1. Natural Language Processing versus Natural Language Understanding

Certains auteurs font la différence entre natural language processing (NLP) et natural language understanding (NLU), où le NLU est un sous-ensemble du NLP. Là où le NLP permet globalement de transformer des données non structurées du langage humain vers des données structurées et utilisables, le NLU cherche en plus à en inférer le sens (Exper.ai., 2020). Le NLP nous permet de traduire du texte, de reconnaître les classes grammaticales ou les entités nommées. La détection d'entités nommées consiste à reconnaître les noms propres comme les organisations et les lieux, mais aussi les quantités, distances ou dates dans un texte. Toutefois, en s'intéressant au sens du langage humain, le NLU nous permet d'échanger verbalement ou par écrit avec une IA, résumer du texte ou - ce qui nous intéressera tout particulièrement – réaliser une analyse de sentiment. (Olivier, 2021).

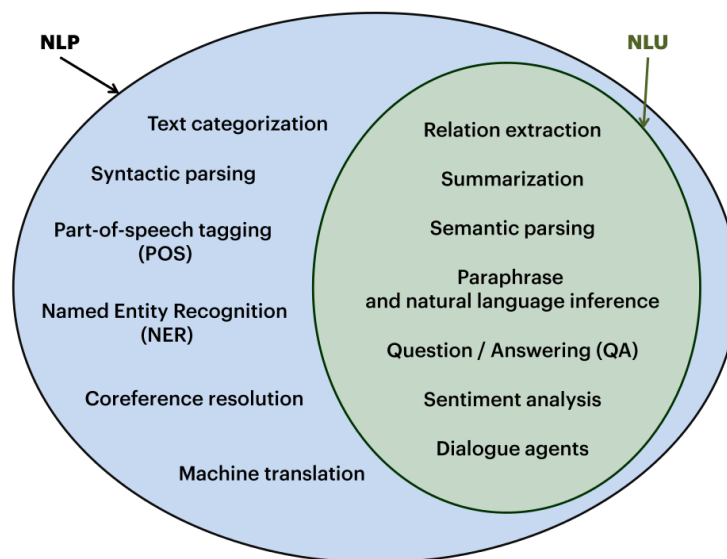


Figure 1 : Le NLP et son sous-ensemble le NLU. (Olivier, 2021)

2.2. Définition et applications

En NLP, une analyse de sentiment peut être considérée comme l'étude et la détection de la polarité du langage humain. Il est alors possible de classer des mots, phrases ou textes selon leur polarité : positif, négatif et parfois neutre. Il est par exemple aussi possible de déterminer une émotion comme la colère, la tristesse, la peur, la joie. Il arrive aussi que la plage de polarité soit plus étendue, comme lorsqu'un commentaire est accompagné d'un certain nombre d'étoiles (souvent 5 ou 10 niveaux). Dans ce cas, le commentaire peut être converti en notation avec une analyse de sentiment.

Avec l'explosion des moyens d'expression sur le web et les incitations à communiquer son avis, les quantités de textes disponibles sont énormes. Ce sont autant d'opportunités pour de nombreux acteurs de chercher à déterminer la polarité d'un texte : connaître l'opinion générale des internautes vis-à-vis d'une marque ou d'un produit selon les commentaires laissés sur les réseaux sociaux, reconnaître le soutien à certaines causes ou projets politiques, repérer les commentaires très mécontents pour mieux focaliser le service après-vente ou repérer les urgences à traiter, vérifier la cohérence d'un commentaire et d'une note (notamment utile pour les systèmes de recommandations), mais aussi détecter les propos haineux et le harcèlement (comme le projet Creep par exemple) (Koppe, 2019) (Towards AI, 2021). Une autre application à développer est la mise en avant de contenu selon leur polarité sur les réseaux sociaux. Actuellement, Facebook cherche à mettre en avant les contenus qui font beaucoup réagir, plutôt représenté par du contenu guidé par des sentiments négatifs comme l'indignation ou la colère. Par exemple les messages recueillant des émojis « colère » auraient 5 fois plus de chance d'être mis en avant par Facebook. Facebook pourrait utiliser l'analyse de sentiment pour définir les contenus à mettre en avant. Évidemment, il serait plus sain d'utiliser cette analyse pour plutôt mettre en avant les contenus positifs pour tenter d'amener de la diversité et s'attaquer au problème de bulle de filtre et de polarisation (Nguyễn Hoang, 2021).

2.3. Types d'analyse de sentiment

Selon de nombreux chercheurs, dont Furqan et al. (2019) Medhat, Hassan & Korashy (2014), Dashtipour et al. (2016) et Gupta & Agrawal (2020), nous pouvons diviser les techniques permettant de procéder à une analyse de sentiment en deux catégories principales : celles basées sur le lexique et celles basées sur le machine learning, ainsi que celles hybrides.

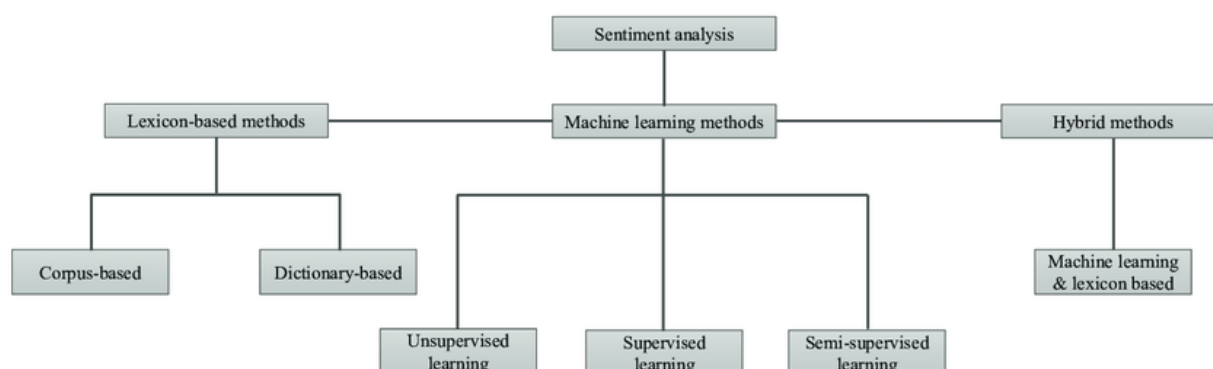


Figure 2 : Approches existantes pour l'analyse de Sentiment (Furqan et al., 2019)

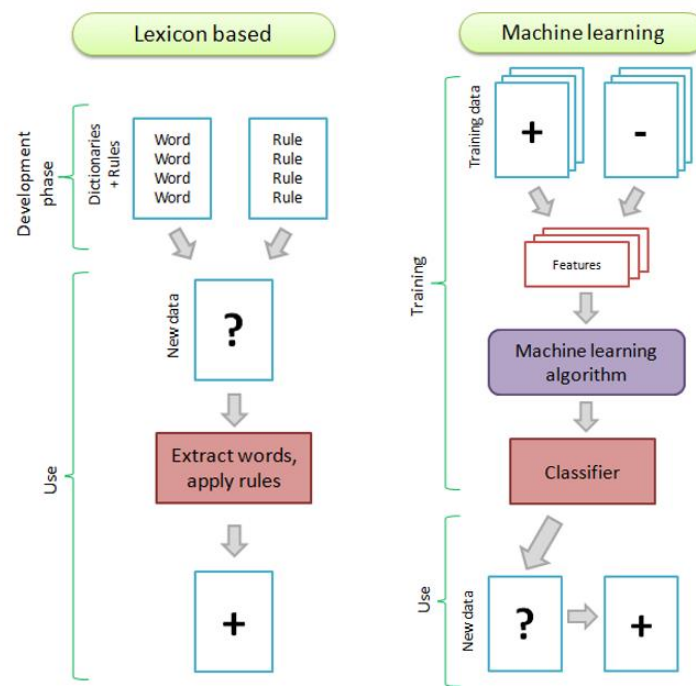


Figure 3 : Fonctionnement des approches lexicales et d'apprentissages (Taboada, 2016)

Comme le montre la figure 3 ci-dessus, l'analyse de sentiment basée sur le lexique détermine le sentiment global d'un texte à partir de l'orientation sémantique des mots ou combinaisons de mots du texte. Autrement dit, chaque mot ou phrase est associé à une polarité (pouvant être plus ou moins forte) dans un lexique de polarité, et la somme (ou une autre opération) de ces polarités donne le sentiment global.

L'analyse de sentiment basée sur le machine learning repose sur les algorithmes de machine learning pour résoudre, généralement, un problème de classification à partir des features (ou caractéristiques) extraites du texte.

2.4. Méthodes lexicales (lexicon-based)

Dans les méthodes lexicales, nous retrouvons deux méthodes, l'approche basée sur un dictionnaire et celle basée sur un corpus. Pour la première, un dictionnaire est créé à partir de quelques mots étiquetés manuellement dont l'orientation est connue. Ensuite, un lexique préexistant est utilisé pour remplir notre dictionnaire avec les synonymes et antonymes des mots présents, et ce à chaque nouvel ajout dans le dictionnaire, jusqu'à ce qu'on ne puisse plus ajouter de mots. Cette méthode ne permet pas d'appréhender le contexte dans lequel les mots sont utilisés, réservant cette technique à des modèles plus généraux, mais néanmoins polyvalents.

La deuxième méthode détermine l'orientation des mots à partir d'une multitude de corpus, ce qui permet de mieux les appréhender dans leurs différents contextes (permettant des applications plus spécialisées), soit avec une approche statistique ou sémantique. Dans l'approche statistique, les mots sont considérés comme positif s'ils apparaissent majoritairement dans des textes positifs et inversement. Alors que dans l'approche sémantique, les mots sémantiquement proches auront la même polarité. Dans ce cas, un mot sera considéré comme positif si la majorité de ses synonymes sont considérés comme positif et inversement (des liens sémantiques autres que la synonymie peuvent être utilisés) (Gupta & Agrawal, 2020).

2.5. Méthodes d'apprentissage (Machine learning-based)

Lorsqu'on utilise le machine learning pour une analyse de sentiment, nous retrouvons souvent la branche supervisée du machine learning, et plus particulièrement la classification. Le machine learning supervisé permet de déterminer des tendances qu'il est soumis à des données d'entraînement (il peut repérer les caractéristiques similaires ou au contraire dissimilaires entre les observations) et utiliser ces informations pour faire des prédictions. La classification s'attache à prédire des valeurs discrètes alors que la régression peut prédire une valeur continue.

Dans le cadre de l'apprentissage supervisé, il est donc nécessaire de fournir un dataset contenant des observations étiquetées selon leur polarité pour permettre au modèle d'apprendre. À l'aide de ces exemples, les mots ou groupes de mots les plus discriminants pour déterminer le sentiment seront déterminés. Des techniques semi-supervisées, hybrides et non supervisées permettent aussi de traiter un problème d'analyse de sentiment, souvent pour faire face à l'absence d'un lexique de polarité ou d'un dataset étiqueté suffisamment complet.

Le semi-supervisé peut utiliser peu de données étiquetées avec un grand nombre de données non étiquetées. Dans ce cas, le modèle est entraîné comme un modèle supervisé sur le peu de données étiquetées, puis il est utilisé pour prédire une partie des données inconnues. Lorsque la prédiction d'un échantillon semble sûre (lorsque la probabilité d'appartenir à une classe dépasse 0.95%, par exemple), il est alors étiqueté selon la prédiction. Ces étapes doivent être répétées en ajoutant à chaque fois plus de données non supervisées, jusqu'à ce que toutes les données non labélisées le deviennent (Ardi, 2020).

Nous n'aborderons pas ici les techniques de deep learning ou celles reposant sur l'apprentissage non supervisé. Les méthodes supervisées ont fait leurs preuves depuis bien longtemps et semblent être les plus utilisées, car elles sont plus facilement interprétables, mises en place et profitent souvent d'un cout computationnel moindre. De plus, la domination du deep learning n'est pas formellement établie et dépend d'un problème à l'autre. Un classifieur comme le SVM, semble encore régulièrement se montrer aussi bon ou voire dépasser le deep learning comme nous le montrent Sullivan (2018) et AL-Smadi, Qawasmeh, Al-Ayyoub, Jararweh, & Gupta (2017).

Nous pouvons enfin brièvement évoquer les techniques hybrides, au sens d'un mélange entre les approches lexicales et d'apprentissages. Elles nécessitent qu'une petite proportion des données soient étiquetées, permettant une couche d'apprentissage, mais un lexique de sentiment est souvent au cœur de la méthode Medhat et al. (2014).

2.6. Machine learning versus lexique

Lorsque nous nous intéressons à ce que nous dit la littérature, il semble raisonnable, malgré l'absence d'un consensus, de croire que les techniques basées sur machine learning semblent plus efficaces que celles basées sur le lexique dans le cadre d'une analyse de sentiment, du moins lors d'applications plus spécialisées.

Selon Nikil & Aloysius (2019), Devika, Sunitha & Ganesh (2016) et Kolchyna, Souza, Treleaven & Aste (2015), les techniques d'apprentissages sont plus efficaces, selon des performances mesurées par l'accuracy. Can, Ezen-Can, & Can (2018) affirment même que le machine learning performe mieux avec des données traduites que les méthodes lexicales avec des données originales. L'article de Dhaoui, Webster & Tan (2017) est plus prudent et conclut que les deux techniques sont semblables en mentionnant toutefois que le machine learning est plus efficace dans le cas de la prédiction de commentaires négatifs. Enfin, le papier de Mukhtar, Khan & Chiragh (2018) affirme la dominance de l'approche lexicale, mais dans le cadre de la langue urdu, là où les autres études traitent de l'anglais.

De plus, Devika (2016) nous rappelle que les techniques basées sur le lexique sont plus polyvalentes, là où celles basées sur le machine learning sont généralement plus spécialisées. En effet, les lexiques sont préconstruits et ne peuvent, sans de gros efforts continus d'adaptations, prendre en compte tous les mots d'argot, les abréviations ou mots empruntés à une autre langue. Néanmoins, ils jouissent d'une base très solide leur permettant de juger n'importe quelle phrase simple.

À l'inverse, les modèles de machine learning apprennent à partir d'un dataset étiqueté et ne sont pas dépendants de lexiques (sauf pour certaines étapes optionnelles de la normalisation, que nous verrons ensuite). Leur gros désavantage est d'être complètement démuni lorsque les modèles sont face à des mots et expressions auxquels ils n'ont jamais été confrontés, leur assurant une domination sur les méthodes lexicales uniquement lorsqu'ils sont utilisés dans des contextes similaires à ceux de leur apprentissage. C'est ce que nous pouvons probablement aussi conclure de l'étude de Kolchyna et al. (2015). Malgré l'ajout d'un lexique spécifique aux émoticônes et aux mots d'argot, améliorant significativement les performances de leur approche lexicale, les modèles de machine learning ont été plus performants dans le contexte de commentaires twitter.

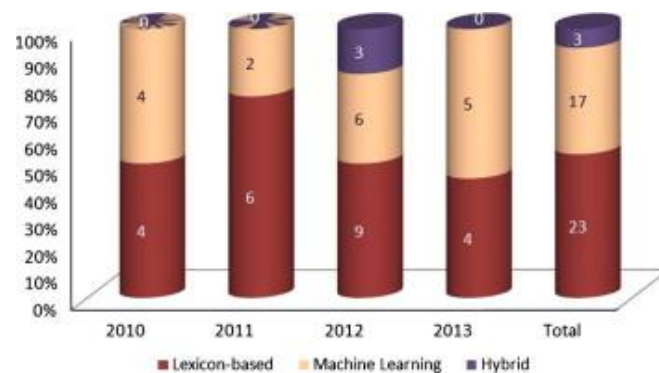


Figure 4 : Nombre d'articles basés sur les différentes approches entre 2010 et 2013. (Medhat et al., 2014)

Quant aux méthodes hybrides, celles-ci ont été peu utilisées, du moins jusqu'en 2013, comme nous le montrent Medhat et al. (2014) sur la figure 4, à cause d'une plus grande complexité algorithmique et de mise en place. Dhaoui et al. (2017) ainsi que Dashtipour et al. (2016), dans leur étude comparative, ajoutent que les méthodes hybrides ne sont pas nécessairement plus efficaces que le machine learning seul.

3. Prétraitement

Le prétraitement est une étape cruciale à faire en amont lors d'une analyse de sentiment. Comme l'énonce Poibeau, « Un texte ne saurait être assimilé à une masse de connaissances directement exploitable par la machine. Il faut dans un premier temps prévoir des traitements complexes pour identifier l'information pertinente, la normaliser, la catégoriser et éventuellement la mettre en contexte. » (Poibeau, 2014, para. 2).

La tokenisation, la normalisation et la vectorisation peuvent être considérées comme les trois grandes étapes du prétraitement. Néanmoins, les techniques basées sur le lexique et celles basées sur le machine learning ne suivent pas exactement les mêmes étapes.

La tokenisation est commune aux deux techniques et a pour but de rendre les chaînes de caractères exploitables. Ensuite, la normalisation fait office de filtre et supprime les mots jugés inutiles dans les deux cas. Les techniques basées sur le lexique se passent de la troisième étape, alors qu'elle est indispensable au machine learning. La vectorisation permet de représenter les mots restants dans l'espace des features, autrement dit, pour les rendre utilisables par les algorithmes de machine learning. Nous allons aborder ici les 3 grandes étapes plutôt sous l'angle du machine learning supervisé.

3.1. Tokenisation

Les machines sont incapables d'exploiter un texte brut et la tokenisation est la première étape permettant de rendre le langage humain utilisable par une machine. L'objectif est de séparer et d'isoler les mots, aussi appelés tokens, pour les identifier dans les chaînes de caractères que sont les phrases. Dans une analyse de sentiment, les tokens permettent la création des features une fois transformés par la vectorisation. Chaque token supplémentaire utilisé dans le corpus de commentaires ajoute une dimension supplémentaire à l'espace des features.

Pour ce faire, deux règles générales sont appliquées (du moins pour les langues basées sur l'alphabet latin et elles peuvent différer pour les autres systèmes d'écriture). La première est de supposer que chaque espace sépare des tokens et la deuxième est que la ponctuation sépare aussi des tokens. Quelques exceptions doivent toutefois être prises en compte pour la deuxième règle. Par exemple, les mots composés (séparés par un tiret, comme loup-garou) ou les adresses mail à devoir considérer comme un token unique (Francq, 2020).

3.2. Normalisation

Nous pouvons considérer cette étape comme un filtrage ou un nettoyage du corpus. Son objectif est de conserver les tokens les plus pertinents, ainsi que de réduire leur nombre, c'est donc une forme de sélection de features.

Une première étape est de mettre en minuscule tous les caractères, car nos algorithmes sont sensibles à la casse. En effet, un mot recevant une majuscule, car il est en début de phrase n'est pas différent du même mot n'en ayant pas. Cependant, il peut être utile de conserver les majuscules de certains mots pour les différencier, comme les noms propres (« des paris » et « Paris »). Pour ce faire, il est possible de conserver les majuscules des mots en milieu de phrases ou de procéder à une détection d'entités nommées.

Une deuxième étape est la suppression des stop-words ou mots vides, ce sont des mots très courants qui n'apportent peu ou pas d'information sur le sens, et en l'occurrence sur la polarité du texte. Il est possible de créer ou d'utiliser un dictionnaire préconstruit réunissant les stop-words à supprimer. En l'absence de dictionnaire, des méthodes statistiques peuvent être utilisées pour détecter les mots présents dans la majorité des commentaires.

Enfin, pour standardiser nos tokens, deux techniques permettent de transformer les tokens restant pour ne garder que leur radical, la racine du mot. En effet, un mot (ou lemme) est composé d'un radical auquel peuvent venir s'ajouter un préfixe ou un suffixe. Autrement dit, ces techniques permettent de revenir à une forme plus simple du mot.

La racinisation s'attache à repérer les affixes (l'ensemble des préfixes et suffixes) connues d'une langue et à les supprimer. Cela nécessite d'avoir à sa disposition une liste d'affixes et de les supprimer systématiquement au début ou à la fin du mot pour ne laisser que le radical. Malheureusement, il peut arriver que le radical lui-même soit amputé d'une partie de ses lettres ou bien que l'entièreté de l'affixe ne soit pas supprimée, renvoyant des radicaux incorrects et incompréhensibles (par exemple, « amies » devient « ami », mais « jouaient » devient « jou »).

La lemmatisation fonctionne en remplaçant les mots par leur radical à l'aide d'un dictionnaire (« amies » devient « ami », mais « jouaient » devient « jouer »). Cela nécessite un dictionnaire complet, car un mot non répertorié ne sera pas modifié. En outre, toutes les méthodes utilisant des lexiques sont très sensibles aux erreurs dans les observations, telles que des fautes de frappe, car elles ne trouveront aucune correspondance en plus d'ajouter du bruit aux données. Le problème peut être corrigé grâce à un correcteur automatique, même si la solution est loin d'être

infaillible. Il est aussi possible d'appliquer la racinisation pour les mots dont aucune correspondance n'a été trouvée dans le dictionnaire de lemmatisation. Si un lexique de lemmatisation est disponible et qu'un cout computationnel plus élevé est permis, la lemmatisation est à favoriser (Chaouche, 2021).

La normalisation a pour but de réduire la complexité algorithmique en réduisant la dimensionnalité, car chaque token unique ajoute une dimension. Cette réduction de la dimensionnalité permet d'alléger le modèle et de gagner du temps de calcul, mais au prix d'une suppression d'information. La suppression de certains tokens peut donc entraîner des conséquences néfastes, nulles voir bénéfiques pour le modèle, d'où l'importance de déterminer la sélection de feature la plus optimale.

Par exemple, l'impact de la lemmatisation sur les performances ne semble pas évident et peut fluctuer, selon les cas, comme le montrent Abdaoui et al (2015). Ils ont procédé à différentes analyses de sentiments à partir de tweets en français en appliquant différents prétraitements. Lorsqu'ils ont tenté de prédire la polarité et l'émotion des tweets (colère, joie, tristesse ...), les performances ont été améliorées de respectivement 2,62% et 1,78% après utilisation de la lemmatisation. Mais lorsqu'ils ont prédit le type de contenu du tweet (information, opinion, sentiment et émotion), les performances ont chuté de 0,98%. De plus, ils ont agrégé la lemmatisation ainsi que le traitement des liens, mails et tags dans la même catégorie, rendant difficile l'évaluation de l'impact de la lemmatisation.

Une fois que les tokens ont été normalisés, il est nécessaire de quitter le langage des humains pour les rendre exploitables pour la machine.

3.3. Vectorisation

Pour pouvoir utiliser des algorithmes de classification, nous devons passer par l'étape de vectorisation. Elle permet de transformer les tokens restants en features utilisables par l'algorithme conformément à l'apprentissage supervisé. Nous distinguons deux grands types de vectorisation dans le cadre du machine learning supervisé, les méthodes de comptage direct et celles de réduction de la dimensionnalité (Chaouche, 2021; Francq, 2020).

Pour le comptage direct, l'objectif est de représenter le texte via un modèle statistique basé sur les fréquences des tokens. Le plus connu est probablement le bag of words ou Bow. Nous représentons le document par l'ensemble des mots qu'il contient à l'aide d'un vecteur des fréquences d'apparition des mots rencontrés dans les documents considérés. Le vecteur aura

une taille égale au nombre de tokens différents et chaque élément du vecteur donne la fréquence de chacun des tokens correspondants.

Il est aussi utile de pouvoir observer un token dans son contexte et nous pouvons l'exploiter à travers les n-grammes. Le n-gramme consiste à prendre toutes les séquences de tokens de taille 1 à n dans un texte, où n sera la taille maximale de la séquence de tokens que nous voulons considérer. Dans « Détester réussir sans mérite », en plus des tokens {(Détester), (réussir), (sans), (mérite)}, nous pouvons aussi extraire les bigrammes suivants : {(Détester, réussir), (réussir, sans) et (sans, mérite)} et les trigrammes suivants : {(Détester, réussir, sans) et (réussir, sans, mérite)}. Utilisés seuls, ces quatre tokens expriment à priori des polarités différentes, mais lorsqu'ils sont organisés en bigrammes et trigrammes, la polarité de l'expression semble plus claire. En réalité, le bag of words n'est que le cas particulier du n-gramme où n est égal à 1. À l'inverse de la normalisation, l'utilisation de ces n-grammes ajoute des features au modèle, lui permettant de déterminer les tokens qui, utilisés ensemble, sont discriminants.

Il est aussi possible d'introduire une pondération des tokens, pour tenter de déterminer les plus pertinents grâce à TF-IDF. Ici nous n'allons plus uniquement utiliser les fréquences d'apparition dans notre texte, mais nous allons introduire une hypothèse : un mot est pertinent s'il est beaucoup représenté dans un document, mais peu dans les autres. Avant, nous soulignions l'importance d'un mot qui apparaissait beaucoup, mais grâce à TF-IDF, un mot aura un score important s'il apparaît beaucoup dans un document, mais peu dans les autres.

Pour les méthodes de réduction de la dimensionnalité, l'objectif est de projeter les documents dans un espace réduit, formé à partir de dimensions synthétiques censées condenser l'essentiel de l'information. Peuvent-être utilisées, par exemple, les méthodes utilisant une décomposition en valeurs singulières (SVD), comme la LSA ou LDA, ou celles basées sur le plongement lexical avec un de ces représentants le plus connus, word2vec.

La différence entre ces deux méthodes est l'input fourni aux algorithmes. Avec celles de comptage, nous obtenons des matrices creuses contenant beaucoup de zéros, car la majorité des mots ne sont pas présents dans les observations. Dans le cas des matrices denses, l'information est synthétisée, fournissant des vecteurs de moins grandes tailles (Chaouche, 2021; Francq, 2020).

4. Inégalité entre les langues

Dans cette partie, nous allons nous intéresser à la place de l'anglais et des autres langues dans le NLP et l'analyse de sentiment. Nous évoquerons aussi le potentiel de la traduction pour répondre à notre problématique, et opérer un transfert entre l'anglais et le français ainsi que ce qui a déjà été fait dans la littérature à ce sujet.

4.1. Engouement pour l'anglais

Commençons par dire que, comme dans de nombreux domaines, l'anglais a une présence écrasante, étant la langue principale de la recherche. Mais dans ces nombreux domaines, l'anglais n'est que le moyen d'expression des objets étudiés, alors qu'en ce qui concerne l'analyse de sentiment, l'anglais en est aussi souvent l'objet.

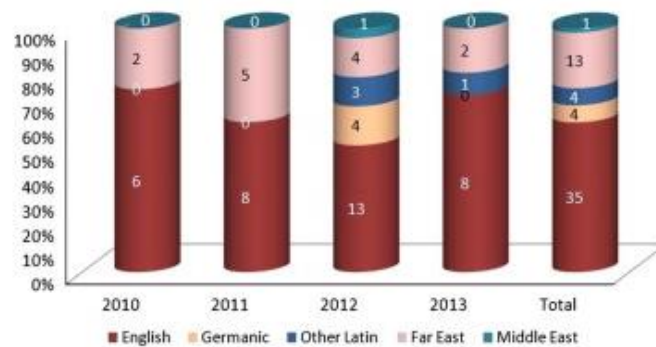


Figure 5 : Nombre de publications relatives à l'analyse de sentiment en anglais et en d'autres langues (Medhat et al., 2014)

Les chercheurs qui étudient l'analyse de sentiment, et de manière plus globale le NLP, s'y intéressent surtout sous le prisme de l'anglais. En effet, comme le montrent Medhat et al. (2014) sur la figure 5, la majorité de l'attention académique autour de l'analyse de sentiment est tournée vers l'anglais.

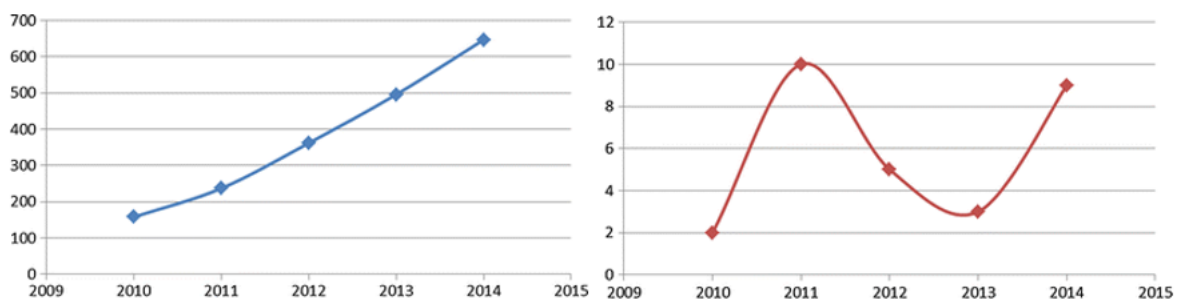


Figure 6 : Nombre de publications relatives à l'analyse de sentiment en anglais (à gauche) et multilingue (à droite) (Dashtipour et al., 2016)

Il est aussi possible de s'intéresser aux autres langues avec des modèles multilingues, traitant plusieurs langues à la fois, mais le constat reste le même selon Dashtipour et al. (2016), celui d'une pauvreté des approches s'intéressant aux autres langues, comme montré sur la figure 6.

Ce favoritisme pour l'anglais mène à la création de nombreuses ressources pour les analyses sur cette langue, comme des outils pour le prétraitement, mais surtout des datasets étiquetés et des modèles entraînés, souvent disponibles publiquement. Cela rend l'anglais plus accessible et attirant, créant un cercle non vertueux pour les autres langues.

Jusque récemment, il était plus compliqué de réaliser une analyse de sentiment pour les autres langues, car les outils nécessaires au prétraitement n'étaient pas courants pour ces autres langues. NLTK (Natural Language Toolkit) est probablement l'outil historique le plus connu sur python, mais il était dédié à l'anglais. Entretemps, d'autres outils sont venus combler le vide pour les autres langues, comme Spacy. Spacy propose, pour de nombreuses autres langues, des bibliothèques contenant les outils nécessaires au prétraitement.

Les techniques basées sur le lexique sont très dépendantes des ressources existantes (lexique de normalisation, de synonyme, de polarité) et donc particulièrement orientées vers l'anglais. Pour procéder à une analyse de sentiment basée sur le lexique en une langue qui ne possède pas de lexique de polarité, il est nécessaire de passer par la traduction complète du lexique pour pouvoir l'utiliser dans cette autre langue, ce qui encore aujourd'hui est considéré comme délicat sans intervention humaine. Cela constitue d'ailleurs une raison supplémentaire pour préférer l'approche du machine learning plutôt que l'approche lexical afin de se passer des lexiques de polarité.

Les techniques basées sur l'apprentissage dépendent plutôt d'un autre apport de la recherche et du public, les datasets. Ces datasets étiquetés, nécessitant beaucoup de travail, sont sur quoi repose le machine learning supervisé. Sans exemple, pas d'apprentissage supervisé possible et c'est leur absence qui limite souvent cette approche. Heureusement, dans le cadre d'une analyse de sentiment basé sur des évaluations de films et de produits, il est possible d'extraire facilement du web des datasets étiquetés, d'où leur utilisation dans ce mémoire. En effet, les sites que nous avons utilisés permettent d'ajouter un commentaire après avoir donné une note. Cela nous fournit donc un commentaire ainsi qu'une note, nous permettent de construire un dataset étiqueté.

4.2. Types de transfert, traduction et positionnement dans la littérature

Dans l'article de Dashtipour et al. (2016), la conclusion est que le problème principal de l'analyse de sentiment en d'autres langues que l'anglais est le manque de ressources lexicales. Ils ajoutent qu'une solution est d'opérer un transfert entre une langue bien fournie, comme l'anglais, vers les autres.

La traduction automatique pourrait répondre à ce manque en permettant le *transfert de ressources*. Il existe plusieurs façons dont la traduction pourrait ouvrir l'analyse de sentiment aux autres langues : (a) la traduction des observations à prédire vers l'anglais pour profiter des modèles existants, (b) la traduction de dataset de l'anglais vers les autres langues pour créer des modèles monolingues ainsi que (c) la création de modèles multilingues.

Néanmoins, construire un modèle multilingue semble ardu pour le moment, selon Ghafoor et al (2021), chaque ajout d'une langue supplémentaire dans un modèle multilingue réduit la capacité de prédiction de toutes les autres. Shalunts, Backfried et Commeignes (2016), de leur côté, envisagent la traduction pour créer des modèles monolingues comme une bonne alternative au multilingue. Mais pour que la traduction soit envisageable dans ces cas, il faut qu'elle réponde à une condition, celle d'être capable de restituer le sentiment, détectable par un algorithme, dans la langue traduite. Évidemment, la traduction pose aussi un certain nombre de problèmes. En plus de son coût computationnel, il est possible que la traduction commette des erreurs ou altère malencontreusement le sentiment contenu dans un commentaire. Elle peut aussi parfois se retrouver démunie face à l'argot ou l'ironie. En effet, dans le cas d'une analyse de sentiment, le moindre mot peut modifier le sentiment général.

Heureusement la traduction s'est considérablement améliorée ces dernières années. Selon plusieurs études s'intéressant à la traduction en tant qu'outil de transfert entre les langues et étudiant la conservation du sentiment après traduction, la qualité de celle-ci est suffisante pour l'utiliser dans une approche de machine learning supervisé. Les auteurs suivantes sont partis de données en anglais qu'ils ont traduits puis ils ont comparé l'accuracy des prédictions de des modèles d'analyse de sentiment entraînés sur des données originales et traduites. Tout d'abord, Brooke, Tofiloski & Taboada (2009) accusaient une perte de performance entre 3 et 5% entre les prédictions des observations originales en anglais et celles traduites vers l'espagnol. Ensuite, Balahur & Turchi (2012) ont procédé à la même analyse vers l'espagnol, l'allemand et le français et accusaient une dégradation des performances de maximum 8%. Ils réitérèrent l'expérience en 2014 et obtiennent de nouveau de bons résultats. Shalunts, Backfried et

Commeignes (2016) attestent de pertes contenues, de maximum 5% après traduction pour l'allemand, le russe et l'espagnol. Al-Shabi, Adel, Omar & Al-Moslmi (2017), concluent que la traduction est mûre pour être utilisée de cette façon, du moins pour la langue arabe, et ajoutent qu'un bon modèle de classification peut être obtenu avec des données traduites malgré le bruit induit par la traduction automatique. Enfin, Ghafoor et al. (2021) constatent une différence moyenne de 2-3% entre l'anglais original et l'Urdu et l'Hindi traduit et jusqu'à 1% pour l'allemand traduit.

En ce qui concerne les outils de traduction automatique, Balahur & Turchi (2012) suggère que Google traduction est l'outil le plus efficace comparé à Bing et Moses et plus particulièrement pour le Français, ce qu'ils reconfirment en 2014. Cette conclusion est basée sur l'utilisation du Bleu score comme métrique, il renvoie un indice entre 0 et 1, pour lequel un indice proche de 1 signifie que les textes en langues différentes sont similaires. Dans leur étude, Can, Ezen-Can, & Can (2018) ont aussi utilisés Google traduction.

5. Conclusion partie I

Nous avons passé en revue l'objectif et l'utilité d'une analyse de sentiment et nous avons compris que dans un monde d'expression, nous devons être capables de traiter le langage naturel et de catégoriser la polarité de son contenu. Nous avons aussi mentionné qu'il y avait deux grandes méthodes pour procéder à une analyse de sentiment, celle basée sur le lexique et celle basée sur le machine learning. Elles ont toutes les deux besoin d'inputs différents et ont chacune leurs utilités, mais le machine learning semble être plus prometteur, surtout dans des tâches plus spécialisées.

Nous avons aussi montré qu'il y avait jusqu'à 3 étapes essentielles pour le prétraitement et que les deux méthodes ne bénéficiaient pas exactement des mêmes étapes. La première permet de découper le texte et la seconde de ne garder que les morceaux qui nous intéressent. Cette dernière est très sensible, car il existe beaucoup de techniques de normalisation et elles sont à définir au cas par cas. La dernière étape, réservée au machine learning, permet de rendre le tout exploitable par ses algorithmes.

Nous avons mis en évidence que l'anglais était favorisé par la recherche, menant à une création de ressources plus importantes pour cette langue et à un intérêt général plus prononcé. Nous avons vu que la traduction pouvait être une piste solide pour procéder à une analyse de sentiment dans une langue moins bien fournie. Nous avons aussi évoqué que la proposition de créer des modèles multilingues n'était pas la plus favorable et nous avons retenu deux types de transfert reposant sur la traduction, (a) la traduction des observations à prédire vers l'anglais pour profiter des modèles existants ainsi que (b) la traduction de datasets de l'anglais vers les autres langues pour créer des modèles monolingues.

Nous avons aussi évoqué la qualité de la traduction et que certains chercheurs la trouvent suffisamment mature pour être utilisée dans le cadre d'une analyse de sentiment basé sur une approche de machine learning supervisé. Le travail des chercheurs mentionné à la fin de la section précédente servira de base à notre méthodologie. Enfin, Google traduction semble être un outil de confiance pour procéder à cette analyse.

Partie 2 : Étude quantitative

1. Objectif

L'analyse de sentiment peut être une tâche ardue dans d'autres langues que l'anglais à cause d'un manque d'outils, de données ou de modèles. Une réponse possible à ce problème est le transfert entre une langue riche, comme l'anglais, vers les autres langues comme le mentionne Dashtipour et al. (2016). Dans ce mémoire, nous allons étudier la possibilité d'utiliser la traduction pour opérer ce transfert. Dans ce cas, ce transfert pourrait se déployer de deux façons, (a) un *transfert d'environnement* par la traduction des observations à prédire vers l'anglais pour utiliser les modèles anglophones déjà développés ou (b) un *transfert de ressources* par la traduction des datasets anglophones vers les autres langues pour permettre de développer des modèles ayant été soumis à suffisamment de données.

L'objectif de ce travail est donc d'explorer l'utilisation de données traduites dans le cadre d'une analyse de sentiment par machine learning supervisé. Pour tester cette hypothèse, il convient d'analyser si la traduction automatique détruit une partie de l'information nécessaire pour mener à bien cette analyse. Nous testerons alors la capacité d'un modèle entraîné sur des données traduites à prédire la polarité d'autres observations traduites. Nous tenterons ensuite de déterminer le meilleur type de transfert, en testant (a) la capacité d'un modèle entraîné sur des données originales à prédire la polarité d'observations traduites et (b) la capacité d'un modèle entraîné sur des données traduites à prédire la polarité d'observations originales. Il est à noter que le but de ce travail est plutôt d'effectuer une comparaison pertinente que d'obtenir la meilleure accuracy possible pour nos modèles.

Au vu des nombreux tests qui seront faits dans ce mémoire, nous allons aussi en profiter pour essayer de répondre à des interrogations au sujet du prétraitement que nous avons rencontré dans la littérature. Elles concernent l'utilisation de bigrammes et trigrammes lors de la vectorisation ou l'intérêt de TF-IDF face à bag of words.

Dans ce travail, nous nous intéresserons uniquement à la traduction entre l'anglais et le français. Cependant, si nos résultats sont satisfaisants, nous pourrions envisager d'opérer le transfert entre l'anglais et n'importe quelle langue dont la qualité de la traduction est suffisante pour conserver le sentiment. Enfin, nous nous concentrons ici sur l'analyse de sentiment, mais la question de la langue touche toutes les disciplines du NLP et d'autres pourraient profiter de ce *transfert de ressources*. Notre analyse et ses résultats pourraient être un indice de l'intérêt de l'exploration de cette voie.

2. Méthodologie

Pour étudier l'impact de la traduction dans le cadre d'une analyse de sentiment, nous avons prévu une méthodologie en trois étapes, dont la première est similaire à celle proposée par Balahur (2012) Brooke, Tofiloski & Taboada (2009), Al-Shabi, Adel, Omar & Al-Moslmi (2017) ou Ghafoor et al. (2021) que nous avons évoquées à la fin de la partie précédente. Cependant, ces articles sont partis de données en anglais et ont uniquement évalué la traduction des données de l'anglais vers d'autres langues. Pour mieux attester de la potentielle destruction d'information après traduction, nous sommes également partis de données en français, que nous traduirons vers l'anglais et auxquelles nous appliquons la même méthodologie.

2.1. Méthodologie en 3 étapes

Lors de cette étape 1, ayant pour but de vérifier que la traduction permet de conserver l'information relative à la polarité, nous commençons par traduire notre dataset original d'une première langue L1 vers une seconde langue L2, puis nous séparons les datasets originaux et traduits entre train et test set (le train set contient les données d'apprentissage et le test set contient les données de test). Les modèles d'analyse de sentiment sont ensuite entraînés à partir de leur train set respectif et sont utilisés pour prédire d'observations de leur test set. Les prédictions sont évaluées selon leur accuracy en utilisant les véritables outputs des test sets. Finalement, nous comparons l'accuracy du modèle i sur le test set i avec l'accuracy du modèle ii sur le test set ii (figure 7). Dans ce cas, l'accuracy nous semble approprié comme métrique, car il n'y a, à priori, pas de différence d'importance entre les faux positifs et faux négatifs. Comme mentionné précédemment, contrairement à Balahur (2012), nous partons une fois de données originales (L1) en anglais et une fois de données originales (L1) en français.

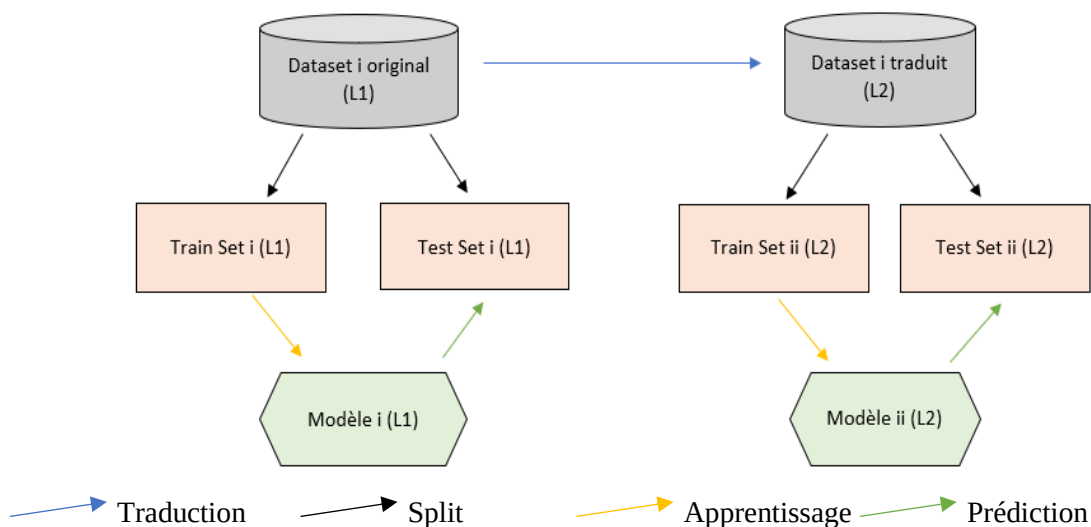


Figure 7 : Schéma de l'étape 1 de notre méthodologie

Malheureusement, l'étape 1 nous renseigne uniquement sur l'impact de la traduction lorsque des données traduites sont utilisées pour réentraîner un modèle, mais elle ne nous permet pas d'évaluer les transferts. C'est pourquoi nous avons d'autres étapes dans notre méthodologie. Les étapes 2a et 3a évaluent le premier type de transfert, (a) amener le français dans l'environnement de l'anglais en traduisant les observations à prédire du français vers l'anglais. Les étapes 2b et 3b évaluent le deuxième type de transfert, (b) transférer les ressources de l'anglais vers le français en traduisant les datasets d'entraînement de l'anglais vers le français. C'est en comparant les résultats des étapes 2a et 3a avec ceux des étapes 2b et 3b que nous pourrions conclure sur le type de transfert à privilégier.

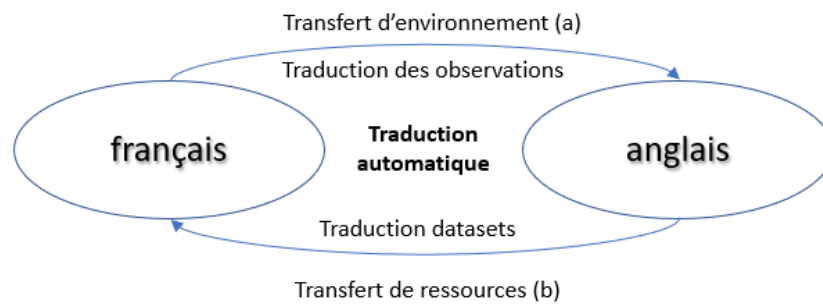


Figure 8 : Types de transferts

Pour choisir le type transfert a, il faudrait que les modèles anglophones (entraînés sur des données originales en anglais) soient capables de prédire efficacement la polarité d'observations traduites du français vers l'anglais. Pour choisir le type de transfert b, il faudrait que les modèles entraînés sur des données traduites de l'anglais vers le français soient capables de prédire efficacement la polarité d'observations francophones. Les prédictions seront considérées comme efficaces si les accuracy obtenues sur les deux test sets sont proches.

Nous commençons par étudier le premier type de transfert (a) en testant la capacité d'un modèle entraîné sur des données originales à prédire la polarité d'observations traduites (2a et 3a).

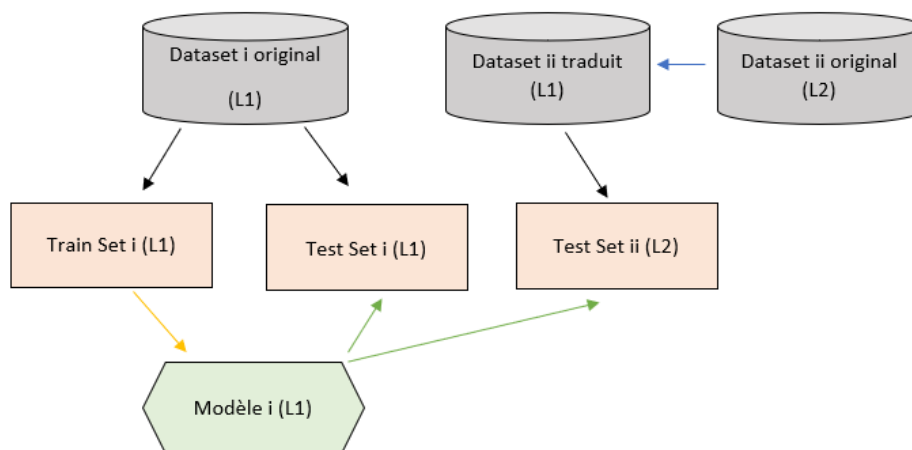


Figure 9 : Schéma de l'étape 2a de notre méthodologie

Dans l'étape 2a (figure 9), nous utilisons un modèle entraîné sur des données originales (train set i) pour prédire la polarité d'observations originales (test set i) et d'observations traduites (test set ii). Notre objectif reste de comparer l'accuracy des prédictions entre les test set i et ii et s'ils sont proches, les prédictions seront considérées comme bonne. Pour éviter un biais évident lors de la comparaison, les deux datasets originaux doivent être similaires, ce que nous développons dans la section suivante.

La séparation entre train et test set ainsi qu'une cross validation viennent assurer l'indépendance entre les données d'entrainements et celles devant être prédites. Mais malgré ces précautions, il reste un autre biais pouvant favoriser le test set dans notre étape 2a, en effet, le test set i partage son dataset d'origine avec le train set i.

C'est pourquoi dans l'étape 3a (figure 10), nous proposons de comparer l'accuracy des prédictions faites sur les test set ii et iii lorsque le modèle a été entraîné sur un dataset tiers i. En effet, entrainer le modèle sur un dataset tiers permet au modèle d'être plus neutre.

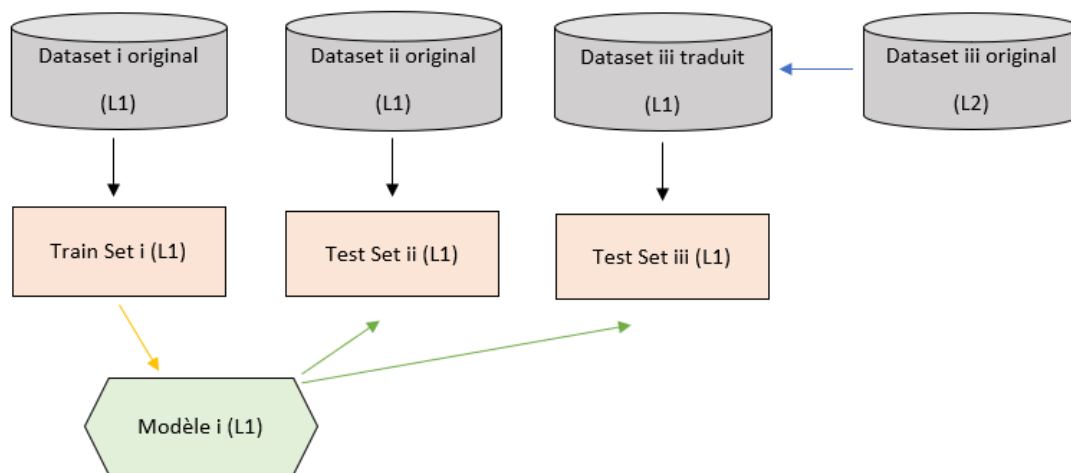


Figure 10 : Schéma de l'étape 3a de notre méthodologie

Cependant, nous n'avons à notre disposition que les quatre datasets déjà utilisés. Nous utilisons donc un des types de dataset que nous avons à notre disposition (de films ou produits) pour l'entrainement et l'autre type pour les prédictions. Il est donc nécessaire que tous nos datasets soient étiquetés de la même façon et possèdent le même nombre de classes. Pour ce faire, nous transformons les datasets de produits en cinq classes vers deux classes avec une règle semblable à la création de ceux de films, c'est-à-dire en supprimant la classe neutre (la classe 3) et en agrégeant les classes 1 et 2 dans une classe négative et les classes 4 et 5 en une classe positive.

Nous étudions maintenant le deuxième type de transfert (b) en testant la capacité d'un modèle entraîné sur des données traduites à prédire la polarité d'observations originales (étapes 2b et 3b). Les schémas expliquant la méthodologie sont légèrement modifiés avec cette variante et nous les représentons ci-dessous (figure 11 et 12).

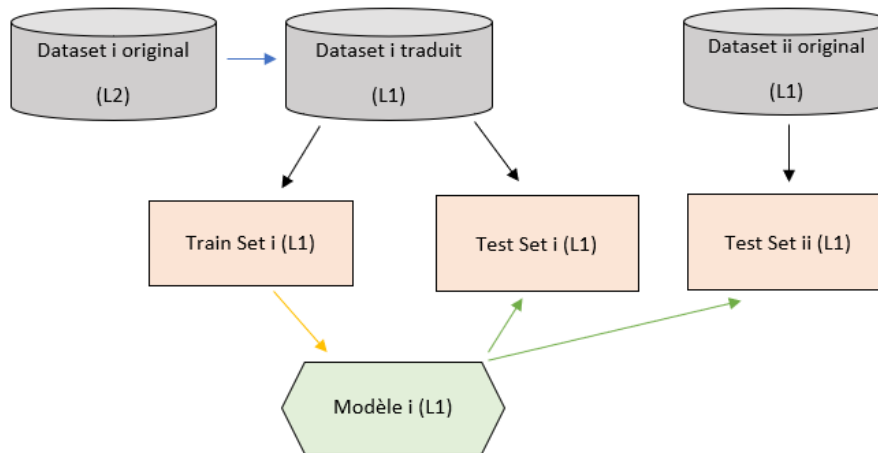


Figure 11 : Schéma de l'étape 2b de notre méthodologie

Dans l'étape 2b, nous utilisons un modèle entraîné sur des données traduites (train set i) pour prédire la polarité d'observations traduites (test set i) et d'observations originales (test set ii), comme montré sur la figure 11, puis nous comparons les accuracy du modèle sur les test set i et ii.

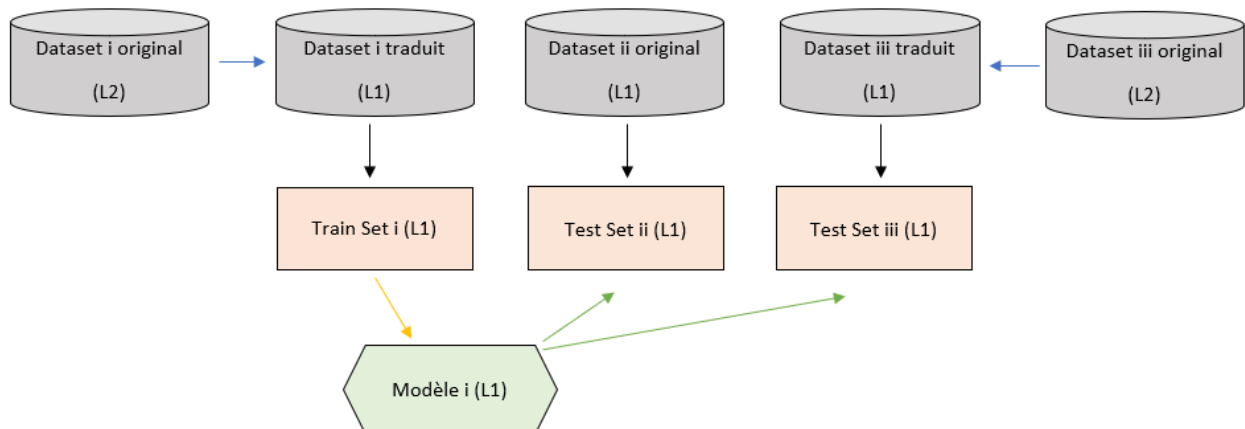


Figure 12 : Schéma de l'étape 3b de notre méthodologie

Dans l'étape 3b, un dataset tiers i sert à entraîner le modèle. Le modèle prédit ensuite la polarité d'observations originales (test set ii) et d'observations traduites (test set iii), comme expliqué sur la figure 12, puis nous comparons les accuracy du modèle sur les test set ii et iii.

2.2. Définition des datasets et traduction automatique

Pour ce travail, nous avons collecté des commentaires relatifs à des films et à des produits vendus sur Amazon, qui formeront nos différents datasets. Nous avons deux datasets pour chaque langue, l'un étiqueté à deux classes, l'autre à cinq classes. L'avantage de ce type de données est d'avoir à disposition les notes associées aux commentaires, permettant la construction d'un dataset étiqueté.

Nous avons donc 4 datasets comprenant au total 720.000 observations, composés comme tels :

Un premier dataset de commentaires de film provenant d'Allociné (site francophone de critique de film), constituée par Blard (2020), contenant 160 000 échantillons en français sur une échelle dichotomique. La polarité est considérée comme positive ou négative et les classes sont équilibrées (figure 13 à gauche).

Malgré nos recherches, aucun dataset similaire aussi complet ne semblait exister en anglais et nous en avons donc créé un. Nous avons fait en sorte de le construire selon les mêmes règles que le précédent. Tout d'abord, nous avons construit ce dataset à partir de commentaires de films déjà extraits du site IMDB par Aditya, Abhilash & Abhijit (2020), proposant une multitude de fichier csv contenant chacun les notes et les commentaires relatifs à un certain film (1150 films différents au total). IMDB est un site dont l'espace commentaire a la même fonction de critique de film qu'Allociné et tous deux visent un profil semblable d'amateurs de film. Nous avons reproduit des classes équilibrées et les avons construites de la même façon que le dataset d'Allociné. En partant d'une note entre 1 et 10, les observations dont la note est supérieure ou égale 8 ont été assignées comme positives et celles inférieures ou égales à 4 ont été assignées comme négatives. Les observations comprises entre cet intervalle ont été rejetées. Ensuite, seuls les commentaires de moins de 2000 caractères ont été gardés, conformément au premier dataset. Enfin, il était important que les observations pour un film donné, représentent autant de commentaires positifs que négatifs, sinon, le nom d'un acteur, d'un réalisateur, d'un personnage, etc. pourrait devenir une feature pertinente pour le modèle. Par exemple, si 100 commentaires du Seigneur des Anneaux étaient extraits, plus de 90 seraient positifs. Tous noms liés à ces films seraient associés dans 90% des cas à des commentaires positifs. Étant donné que les noms des acteurs ou personnages sont souvent utilisés dans les commentaires, l'algorithme pourrait plutôt utiliser ces noms pour déterminer une polarité positive, le rendant évidemment non généralisable. Nous avons terminé ce dataset en fusionnant les couples notes et commentaires dans un dataset final contenant aussi 160 000 observations. Les règles

Les deux datasets suivants ont été composés par Keung, Lu, Szarvas & Smith (2020) de commentaires de produits Amazon en français et en anglais, contenant 200 000 échantillons chacun sur une échelle de 1 à 5. Ils sont originaires de la plateforme Hugging Face et peuvent être obtenus via leur librairie python « datasets ». Dans ces datasets les classes sont également équilibrées (figure 13 à droite).



The image displays four word clouds arranged in a 2x2 grid, each representing a different theme or sentiment. The top-left cloud is in French and features words like 'mais', 'scène', 'film', 'histoire', 'même', 'celui', 'réalisateur', and 'bien'. The top-right cloud is also in French, with prominent words such as 'dommage', 'donc', 'qualité', 'prix', 'très', 'petit', 'temps', 'tout', 'bien', 'faire', 'prix', 'même', 'aussi', and 'pratique'. The bottom-left cloud is in English and includes words like 'film', 'great', 'one', 'love', 'use', 'well', 'actor', 'watch', 'make', 'need', 'buy', 'product', 'case', 'happy', 'keep', 'book', 'size', 'back', 'work', 'even', and 'really'. The bottom-right cloud is also in English, with words such as 'well', 'take', 'look', 'will', 'great', 'one', 'love', 'use', 'need', 'buy', 'product', 'case', 'happy', 'keep', 'book', 'size', 'back', 'work', 'even', and 'really'.

26

Dans le même ordre, les datasets de films en français ont en moyenne 105 mots par commentaire et ceux de produits 32. En ce qui concerne l’anglais, nous retrouvons en moyenne 172 mots dans les commentaires de films et 39 dans ceux des produits. Les commentaires en anglais ont donc en général plus de mots, mais nous remarquons surtout que les commentaires de film sont bien plus riches en contenu.

Pour mener ce travail à bien, nous avons dû ensuite traduire ces quatre datasets. Afin de réaliser la traduction des 720.000 observations cumulées, nous avons utilisé l’API de Google Traduction avec la librairie « pygooglettranslation ». Balahur & Turchi (2012, 2014) avaient déjà utilisé Google Traduction dans leur étude et avaient conclu que l’outil est plus efficace que Bing et Moses, surtout dans le cas du français. En 2012, Balahur et Turchi avait obtenu un Bleu de score de 0.274 pour Google Traduction de l’anglais vers le français, reflétant une traduction loin d’être parfaite, mais ils sont néanmoins arrivés à la conclusion que la qualité de la traduction automatique était suffisante pour permettre d’être utilisée pour une analyse de sentiment. En effet, la traduction n’a pas besoin d’être parfaite, seulement de permettre de détecter la polarité.

2.3. Prétraitement

La normalisation et la vectorisation sont des étapes importantes pour déterminer l’input à fournir aux classifieurs. Lors de la première étape de notre méthodologie, nous avons étudié différents prétraitements, afin d’essayer de répondre à des interrogations sur ceux-ci et d’identifier le prétraitement optimal. Nous avons plus particulièrement appliqué aux 1.440.000 observations cumulées après traduction de nos datasets 4 variantes de normalisation, 2 types de vectorisation et une plage de $n = 1$ à 3 pour les n -grammes, c’est-à-dire 24 versions par dataset et par classifieur. Cette étude du prétraitement a été faite exclusivement lors de la première étape de notre méthodologie et le prétraitement optimal a été appliqué lors des étapes 2 et 3.

La première variante de normalisation consiste simplement en la suppression des majuscules. Nous précisons aussi que les étapes de normalisation se cumulent entre chaque variante. Pour la deuxième variante, nous avons supprimé les potentiels liens vers des sites internet et les adresses mail. Nous avons ensuite supprimé tous les caractères spéciaux et les chiffres pour ne garder que les caractères compris entre a et z, mais nous avons conservé les accents pour le français. Nous avons aussi supprimé tous les tokens de 2 caractères et moins, à l’exception de ‘no’ et ‘if’ en anglais et ‘si’ et ‘ni’ en français, pour se débarrasser d’interjection et de nombreux pronoms et préposition rarement informative sur la polarité (tout en gardant l’information de

négation et condition). La différence entre la deuxième et la troisième variante est l'application de la lemmatisation. Elle a été effectuée sur base des lexiques fournis par Spacy dans ses modèles de langages : « en_core_web_sm » et « fr_core_news_sm ». Enfin, pour la quatrième et dernière variante, nous supprimons les stopwords, ou mots les plus courants, dont le lexique est aussi fourni par Spacy.

Quant à la vectorisation, Balahur et Turchi (2012) ont questionné à la fin de leur article l'intérêt d'utiliser TF-IDF au lieu de bag of words. Ils se sont répondu en 2014 en montrant que TF-IDF améliorait faiblement les résultats, mais ils ont soulevé une nouvelle interrogation, l'intérêt de l'utilisation des bigrammes. Al-Shabi et al., (2017) ont aussi rapporté l'intérêt de TF-IDF et ajoutent que l'utilisation de bigrammes est assez bénéfique, mais leur article étudiait la traduction vers la langue arabe, probablement assez différente du français. Néanmoins, ils ont aussi soulevé la question de l'utilisation des trigrammes. Nous profitons donc de ce travail pour tenter de répondre aux interrogations précédentes en testant l'utilisation des n-grammes où $n = 1, 2$ et 3 ainsi que bag of words et TF-IDF pour la vectorisation.

Enfin, lors de la vectorisation, nous pouvons paramétrer les seuils de fréquences minimaux et maximaux d'apparition des n-grammes dans le corpus et ne sélectionner que ceux entre ces seuils pour devenir nos features. Nous n'avons pris que les n-grammes qui apparaissent dans moins de 90% des cas (un n-gramme apparaissant dans chaque commentaire n'est pas informatif) et ceux qui apparaissent au moins 5 fois dans tout le corpus.

2.4. Différents classifieurs et paramètres

Pour entrainer nos modèles, nous avons utilisé un Bayes naïf (BN) et une régression logistique (LR). Un support vector machine (SVM) et un réseau de neurones (NN) ont été mis en place, mais leur longue durée computationnelle, induite par la grande dimensionnalité des données (jusqu'à presque un million), a entravé leur utilisation au vu des nombreux modèles à entrainer dans ce travail. Nous avons essayé de réduire la dimensionnalité pour réduire le temps d'apprentissage, mais cela avait un impact important sur leur accuracy, jusqu'à les rendre bien moins performants que les deux autres classifieurs retenus, tout en restant relativement longs. Pour ces classifieurs, nous avons utilisé les bibliothèques disponibles sur SKlearn. Les datasets ont été séparés en 5 folds et des prédictions ont été faites sur chacun d'entre eux à partir des quatre autres folds, dans la logique de la cross validation. Pour chacune des configurations, les modèles ont fait l'objet d'un grid search, permettant de tester tous les paramètres souhaités pour un classifieur et de déterminer les paramètres optimaux.

La régression logistique est un classifieur exprimant la probabilité d'une observation d'appartenir à une certaine classe avec la fonction sigmoïde et une combinaison linéaire des variables explicatives. Les coefficients des variables sont estimés durant la phase d'entraînement avec un modèle de maximum de vraisemblance. Les solvers sag et saga sont conseillés pour les grands datasets mais le solveur lbfgs était deux fois plus rapide dans nos tests préliminaires. Pour éviter le surapprentissage et pousser le modèle à rester simple, un terme de pénalité est ajouté ; nous avons implémenté le terme l2, car il est le seul compatible avec le solveur lbfgs. Nous avons aussi utilisé un gridsearch pour déterminer l'hyperparamètre « C » optimal pour chaque variante du prétraitement. Ce paramètre, dit de régularisation, définit l'importance du terme de pénalité (Satinet et Fouss, 2022)

Le Bayes Naïf est un classifieur utilisant la formule de Bayes pour prédire la probabilité d'un set de features d'appartenir à une classe. En NLP, l'input utilisé est un vecteur représentant le document vectorisé à partir d'une technique d'extraction de feature (ou vectorisation). Ce classifieur est dit naïf, car il part de l'hypothèse que les features définissant les classes sont indépendantes entre elles (Medhat, 2014). Nous avons utilisé un gridsearch pour déterminer l'hyperparamètre « alpha » optimal, utilisé dans le lissage de LaPlace. Ce lissage sert à empêcher l'apparition d'une probabilité égale à 0, survenant lorsqu'il est impossible d'évaluer la probabilité à priori d'apparition d'un mot lorsqu'il n'est pas présent et cet hyperparamètre détermine la manière d'opérer ce lissage (Vaibhav, 2020).

3. Analyse des résultats

3.1. Étape 1

Nous retrouvons, ci-dessous, les tableaux récapitulatifs de résultats de l'analyse faite dans le cadre de la première étape de notre méthodologie, exprimés selon leur accuracy. Dans le premier tableau, la langue originale est l'anglais alors que dans le deuxième, la langue originale est le français. Les résultats sont aussi regroupés selon le type de dataset (de films ou produits), le type d'observations (originales ou traduites), la variante de normalisation (V1 à V4), ainsi que le type de n-gramme, de vectorisation et enfin de classifieur.

			Anglais Original et Traduction en Français															
			Films (2 classes)								Produits (multiclasse)							
			Original				Traduction				Original				Traduction			
			V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
ngrams n = 3	Bow	Bayes	93,22	93,05	92,91	90,23	92,41	92,33	92,10	89,46	53,20	52,88	52,66	47,16	51,48	51,39	51,25	44,98
		LR	93,36	93,16	93,01	91,11	92,59	92,41	92,24	90,01	53,92	53,40	53,23	48,42	52,61	51,98	51,92	46,00
	Tfidf	Bayes	93,49	93,25	93,08	90,43	92,70	92,52	92,25	89,64	53,34	52,94	52,66	47,43	51,68	51,36	51,23	45,19
		LR	93,91	93,58	93,45	91,38	93,12	92,90	92,58	90,25	54,40	53,97	53,86	48,60	52,96	52,58	52,45	46,21
	features (k)		912	765	755	396	945	741	721	371	278	224	226	108	304	219	216	96
ngrams n = 2	Bow	Bayes	92,27	92,30	91,98	90,08	91,41	91,59	91,18	89,34	52,85	52,55	52,22	47,29	51,26	51,04	50,80	45,05
		LR	93,18	92,97	92,75	91,00	92,32	92,15	91,92	89,93	53,74	53,29	53,12	48,43	52,37	52,00	51,79	46,02
	Tfidf	Bayes	92,63	92,59	92,23	90,32	91,76	91,84	91,46	89,47	52,92	52,58	52,18	47,50	51,12	51,09	50,73	45,15
		LR	93,66	93,42	93,14	91,35	92,73	92,64	92,30	90,22	54,31	53,90	53,73	48,55	52,87	52,48	52,26	46,23
	features (k)		429	412	367	289	427	417	376	295	147	138	126	91	152	139	128	87
ngrams n = 1	Bow	Bayes	89,15	89,23	88,75	88,32	88,32	88,71	88,08	87,29	50,12	49,91	49,45	46,52	48,74	48,30	48,12	44,28
		LR	91,26	91,14	90,66	89,97	90,36	90,36	89,84	88,68	51,74	51,59	51,17	47,60	50,23	50,05	49,63	45,17
	Tfidf	Bayes	89,40	89,46	88,82	88,29	88,69	88,76	88,18	87,36	49,30	49,20	48,47	46,30	48,03	47,55	47,07	44,23
		LR	91,55	91,27	90,79	90,19	90,54	90,44	89,95	88,83	52,16	51,91	51,46	47,76	50,73	50,34	50,08	45,31
	features (k)		39	38	31	31	49	48	39	39	18	17	13	13	22	21	16	16

Tableau 1 : Accuracy des modèles sur des observations originales anglophones

			Français Original et Traduction en Anglais															
			Films (2 classes)								Produits (multiclasse)							
			Original				Traduction				Original				Traduction			
			V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
ngrams n = 3	Bow	Bayes	92,62	92,73	92,49	90,55	91,61	91,8	91,56	89,56	52,53	52,64	52,57	46,95	51,49	51,5	51,29	46,56
		LR	93,77	93,65	93,54	91,45	93,07	92,95	92,74	90,68	53,89	53,51	53,48	47,86	52,91	52,75	52,5	47,46
	Tfidf	Bayes	92,93	92,89	92,75	90,54	92,04	92	91,8	89,62	52,55	52,69	52,55	47,01	51,36	51,4	51,12	46,49
		LR	94,17	94,01	93,93	91,85	93,49	93,39	93,21	91,02	54,13	53,95	53,93	48,05	53,24	53,17	52,87	47,55
	features (k)		616	438	437	218	541	435	435	214	238	167	167	79	207	165	167	80
ngrams n = 2	Bow	Bayes	91,76	92,19	91,88	90,44	90,91	91,16	90,92	89,49	52,27	52,41	52,34	47,04	51,36	51,41	51,1	46,67
		LR	93,62	93,52	93,38	91,44	92,90	92,81	92,53	90,66	53,70	53,55	53,45	47,86	52,76	52,56	52,33	47,49
	Tfidf	Bayes	92,23	92,42	92,21	90,42	91,47	91,57	91,27	89,51	52,14	52,36	52,17	47,1	51,17	51,24	50,87	46,6
		LR	94,03	93,87	93,72	91,84	93,27	93,19	92,98	90,97	54,06	53,82	53,71	48,08	53,14	52,97	52,65	47,55
	features (k)		328	293	277	190	295	270	253	186	130	114	108	71	115	104	99	70
ngrams n = 1	Bow	Bayes	89,08	89,65	89,22	88,86	88,04	88,68	88,34	87,96	49,77	49,77	49,71	46,37	49,33	49,32	48,96	46,07
		LR	92,04	92	91,68	90,59	91,2	91,17	90,77	89,76	51,66	51,53	51,29	47,29	50,86	50,77	50,45	46,94
	Tfidf	Bayes	89,78	89,98	89,47	88,86	88,83	89,03	88,53	87,73	49,08	49,11	48,84	46,36	48,49	48,54	48,09	45,84
		LR	92,38	92,27	91,9	90,76	91,5	91,41	91,01	89,97	51,89	51,63	51,5	47,48	51,02	50,92	50,49	47,01
	features (k)		51	50	41	41	41	41	35	35	20	19	16	15	16	15	12	12

Tableau 2 : Accuracy des modèles sur des observations originales francophones

Par exemple, dans la première case du tableau 1, 93,22 est l'accuracy du modèle, entraîné sur des données originales de films en français et utilisant comme prétraitement la première variante V1, des n-grammes où n = 3 et bag of words (ou Bow) et obtenu par un Bayes naïf.

Nous pouvons aussi retrouver dans les tableaux 1 et 2 le nombre de features (donnant la dimensionnalité), en millier, et nous constatons que la dimensionnalité diminue lorsque la normalisation est plus forte (ou contient plus d'étapes, en passant de V1 à V4). Mais ce n'est pas une règle absolue, nous pouvons trouver dans nos tableaux deux cas où la dimensionnalité augmente après lemmatisation (par exemple, pour les produits en langues originales et n-grammes où $n = 3$ dans le tableau 1). En effet, avec les seuils d'apparition minimaux appliqués, un n-gramme doit apparaître au moins cinq fois et il a plus de chances d'apparaître après lemmatisation étant donné que les tokens sont standardisés après cette étape. En général, le nombre de dimensions augmente aussi après traduction vers le français, cela peut s'expliquer par les accords et la conjugaison française.

	Anglais vers Français		Français vers Anglais	
	Original	Traduction	Original	Traduction
Films (2 classes)	91,60	90,75	91,90	91,04
Produits (multiclasse)	51,16	49,49	50,99	50,18

Tableau 3 : Accuracy selon le type de dataset et la langue

Le tableau 3 agrège les résultats des tableaux 1 et 2, en faisant une moyenne de l'accuracy selon le type de dataset et les langues originales et traduites. Il était attendu que les résultats des prédictions des observations de produits soient inférieurs à ceux de film, car il y a cinq classes et de plus, les notes neutres (donc ambiguës) ont été supprimées. Sur ce tableau 3, nous voyons que les modèles entraînés avec des données traduites performant toujours moins bien et leur accuracy diminue en moyenne de 0,93% (ou 0,85 point de pourcentage) pour les films (deux classes) et de 2,44% (ou 1,25 point) pour les produits (cinq classes) ($p < 0.01$). Ceci est un premier résultat important, indiquant qu'après traduction l'accuracy diminue toujours, mais assez peu en deux classes.

	Films (2 classes)		Produits (multiclasse)	
	Original	Traduction	Original	Traduction
Bayes	91,07	90,21	50,37	49,11
LR	92,43	91,58	51,78	50,55

Tableau 4 : Accuracy selon le type de classifieur et la langue

La tableau 4 nous montre que la régression logistique performe toujours mieux que le Bayes naïf. Opter pour la régression logistique, s'accompagne d'un gain d'accuracy moyen de 1,5% pour deux classes et de 2,87% pour le multiclasse en contrepartie d'un temps computationnel bien plus élevé pour l'apprentissage (entre 50 et 100 fois plus long).

3.2. Prétraitement

Nous nous intéressons maintenant au prétraitement afin d'identifier le prétraitement optimal. En choisissant des n-grammes où $n = 2$ au lieu de $n = 1$, l'accuracy augmente en moyenne de 2,4% (ou 2,15 points) en deux classes et 4,16% (ou 2,04 points) en multiclasse ($p < 0.01$) pour en contrepartie une augmentation de 748% du nombre de features. Alors que passer de n-grammes où $n = 2$ à $n = 3$ permet de gagner 0,41% (ou 0,37 point) d'accuracy en deux classes et 0,30% (ou 0,16 point) en multiclasse ($p < 0.01$) pour en contrepartie, une augmentation de 162% du nombre de features (tableau 5).

n-grammes où $n =$	1	2	3
Films (2 classes)	89,77	91,92	92,29
Produits (multiclasse)	49,05	51,08	51,24
Nb de features moyen	28	216	371

Tableau 5 : Accuracy selon le type de n-gramme et de dataset

La version des n-grammes où $n = 3$ est sans étonnement la meilleure, mais le gain de performance n'est pas linéaire avec l'augmentation du nombre de features (figure 15). Si nous appliquons une simple règle de 3 au cas du deux classes, la première augmentation de feature de 748% augmente notre accuracy de 2,4%, dès lors, l'augmentation attendue de l'accuracy pour une nouvelle augmentation de feature de 162% devrait s'élever à 0,54%, mais elle n'est que de 0,30%. Le constat est encore pire en multiclasse, l'augmentation n'est que 0,16% contre une amélioration attendue de 0,93%. L'amélioration reste non négligeable, mais nous avons gardé la version $n = 2$ pour les n-grammes afin de diminuer le temps computationnel, étant donné que l'objectif de ce travail n'est pas d'obtenir la meilleure accuracy.

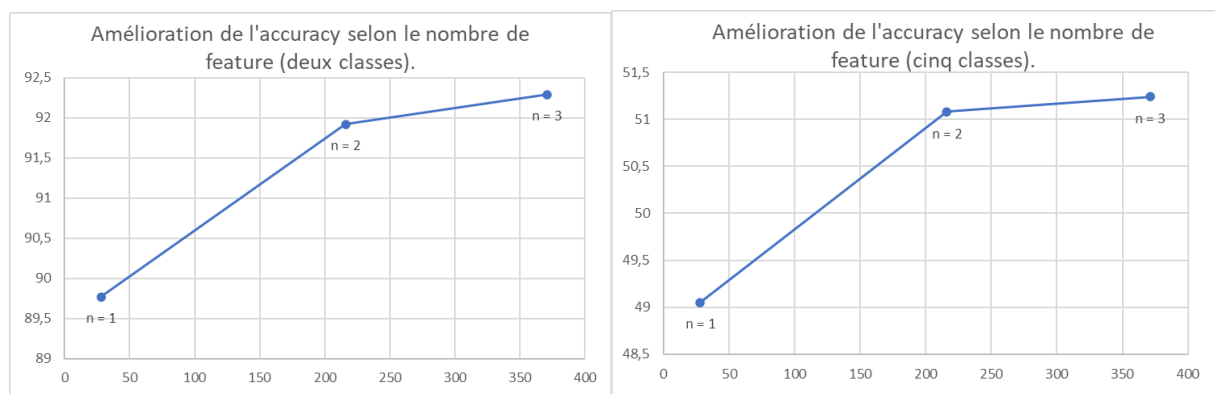


Figure 15 : Amélioration selon le type de n-gramme choisi, à gauche en deux classes et à droite en cinq classes

Variante du prétraitement =	V1	V2	V3	V4
Films (2 classes)	91,89	91,88	91,57	89,95
Produits (multiclasse)	51,89	51,7	51,45	46,77

Tableau 6 : Accuracy selon le type de dataset et le type de normalisation

En revanche, selon le tableau 6, la version 1 de la normalisation (V1), consistant à ne pas modifier les commentaires (à part la suppression des majuscules), fut celle permettant d'obtenir les meilleurs résultats. Nous avons donc conservé cette version pour les étapes 2 et 3, de plus cela permet d'éviter les étapes lourdes de la normalisation comme la lemmatisation. La lemmatisation souffre aussi de quelques ratés, par exemple, "cinéma" est transformé en "cinémer" comme si c'était un verbe conjugué au passé simple.

Type de vectorisation =	Bow	Tf-idf
Films (2 classes)	91,18	91,47
Produits (multiclasse)	50,43	50,48

Tableau 7 : Accuracy selon le type de dataset et le type de vectorisation

Le tableau 7 nous montre que TF-IDF performe en moyenne légèrement mieux que bag of words, améliorant l'accuracy de 0,31% (ou 0,28 point) en deux classes et de 0,11% (0,05 point) en multiclasse. Néanmoins TF-IDF ne permet pas toujours d'obtenir les meilleurs résultats.

En résumé, nous conservons pour les étapes 2 et 3 de notre méthodologie, la première variante de la normalisation et pour la vectorisation, nous conservons les n-grammes où $n = 2$ ainsi que bag of words et TF-IDF.

3.3. Étape 2

Après avoir déterminé les meilleurs paramètres de prétraitement à partir de l'étape 1, nous les avons utilisés pour les étapes 2 et 3 de notre méthodologie. Pour rappel, l'objectif de la suite de la méthodologie est de comparer deux types de transfert. (a) Le *transfert d'environnement* du français vers l'anglais en testant la capacité d'un modèle entraîné sur des données originales à prédire la polarité d'observations traduites, en comparant les accuracy des prédictions du modèle sur des observations originales et traduites. (b) Le *transfert de ressources* de l'anglais vers le français en testant la capacité d'un modèle entraîné sur des données traduites à prédire la polarité d'observations originales, en comparant encore les accuracy des prédictions du modèle sur des observations traduites et originales.

Le tableau 8 dévoile les résultats de l'étape 2a dans laquelle nous avons utilisé un modèle entraîné sur des données originales pour prédire la polarité d'observations originales et traduites.

		Anglais Original et Français Traduit				Français Original et Anglais Traduit				Moyenne
		Bow		Tfidf		Bow		Tfidf		
		Bayes	LR	Bayes	LR	Bayes	LR	Bayes	LR	
Films (2 classes)	Original	92,27	93,18	92,63	93,66	91,76	93,62	92,23	94,03	92,92
	Traduction	83,08	85,02	83,38	81,59	79,48	84,36	81,31	80,38	82,33
Produits (multiclasse)	Original	52,85	53,74	52,92	54,31	52,27	53,70	52,14	54,06	53,25
	Traduction	45,98	46,93	46,03	43,70	45,36	46,23	44,09	41,85	45,02

Tableau 8 : Comparaison des accuracy lors de l'étape 2a

Par exemple, 92,27 est l'accuracy obtenu lorsqu'un modèle a été entraîné sur des données originales de films en anglais et qu'il prédit la polarité d'observations venant du même dataset. 83,08 est l'accuracy obtenu lorsque le même modèle, entraîné sur les mêmes données, prédit la polarité d'observations traduites du français vers l'anglais à partir d'un dataset de films équivalent. Lorsque nous comparons les accuracy des prédictions sur des observations en langues originales et traduites, nous observons, en moyenne, une différence de 11,4% (ou 10,6 points) en deux classes et 15,5% (ou 8,23 points) en multiclasse ($p < 0.01$).

La tableau 9 dévoile les résultats de l'étape 2b dans laquelle nous avons utilisé un modèle entraîné sur des données traduites pour prédire la polarité d'observations traduites et originales.

		Anglais Original et Français Traduit				Français Original et Anglais Traduit				Moyenne
		Bow		Tfidf		Bow		Tfidf		
		Bayes	LR	Bayes	LR	Bayes	LR	Bayes	LR	
Films (2 classes)	Traduction	90,91	92,90	91,47	93,27	91,41	92,32	91,76	92,73	92,10
	Original	79,28	85,46	80,83	82,31	84,39	85,75	84,66	82,22	83,11
Produits (multiclasse)	Traduction	51,36	52,76	51,17	53,14	51,26	52,37	51,12	52,87	52,01
	Original	46,46	46,99	45,12	43,16	46,54	47,40	46,69	43,06	45,68

Tableau 9 : Comparaison des accuracy lors de l'étape 2b

Lorsque nous comparons les accuracy des prédictions sur des observations en langues originales et traduites, nous observons, en moyenne, une différence de 9,8% (ou 8,98 points) en deux classes et 12,2% (ou 6,33 points) en multiclasse ($p < 0.01$). Les écarts sont assez importants dans cette étape 2, mais il est possible qu'ils soient exagérés en raison d'un biais de la source, c'est-à-dire lorsqu'un des deux test sets partage son origine avec le train set, mais pas l'autre. C'est pourquoi nous avons mis en place la troisième étape de la méthodologie. Néanmoins, si nous comparons les écarts de performance en pourcentage entre les étapes 2a et 2b, les résultats favorisent déjà le deuxième type de transfert (b).

3.4. Étape 3

Nous utilisons, lors de l'étape 3, un dataset tiers pour l'entraînement et deux test sets indépendants des données d'entraînements et venant de datasets similaires pour les prédictions. Par exemple, lorsque nous utilisons un dataset de produits en anglais pour l'apprentissage, le modèle prédit ensuite la polarité d'observations originales de films en anglais et la polarité d'observations traduites du français vers l'anglais de films. Les prédictions faites sur un certain type de données (film ou produit) étant réalisées à partir d'un modèle entraîné sur des données de l'autre type, il est attendu que l'accuracy soit plus faible que lors des étapes précédentes. Dans cette étape 3, tous les datasets ont été remis en deux classes donc nous ne faisons plus la différence entre le deux classes et le multiclasse, mais seulement entre les datasets de films et de produits.

Le tableau 10 renvoie les résultats de l'étape 3a dans laquelle nous avons utilisé un modèle entraîné sur des données originales tierces pour prédire la polarité d'observations originales et traduites.

		Anglais Original et Français Traduit				Français Original et Anglais Traduit				Moyenne
		Bow		Tfidf		Bow		Tfidf		
		Bayes	LR	Bayes	LR	Bayes	LR	Bayes	LR	
Films (2 classes)	Original	77,14	83,48	76,60	83,72	77,14	80,82	78,16	81,51	79,82
	Traduction	68,64	76,50	68,07	76,55	74,77	76,44	74,69	76,82	74,06
Produits (2 classes)	Original	75,83	77,32	75,27	77,72	75,29	76,29	74,91	77,38	76,25
	Traduction	73,01	74,93	71,76	74,80	67,79	70,26	67,84	71,15	71,44

Tableau 10 : Comparaison des accuracy lors de l'étape 3a

Ici, les différences sont plus contenues que lors de l'étape 2a. Nous observons, en moyenne, une différence entre les accuracy des prédictions sur des observations en langues originales et traduites de 7,22% (ou 5,76 points) pour les films et de 6,31% (ou 4,81 points) pour les produits ($p < 0.01$) ou une moyenne globale de 6,76%.

Le tableau 11 renvoie les résultats de l'étape 3b dans laquelle nous avons utilisé un modèle entraîné sur des données traduites tiers pour prédire la polarité d'observations traduites et originales.

		Anglais Original et Français Traduit				Français Original et Anglais Traduit				Moyenne
		Bow		Tfidf		Bow		Tfidf		
		Bayes	LR	Bayes	LR	Bayes	LR	Bayes	LR	
Films (2 classes)	Traduction	78,50	80,18	79,00	80,87	76,23	82,33	75,86	82,81	79,47
	Original	76,04	76,44	75,29	77,14	68,35	76,63	68,5	76,90	74,41
Produits (2 classes)	Traduction	72,58	74,70	71,91	75,36	74,15	76,43	74,28	76,88	74,54
	Original	66,58	71,32	66,37	71,55	75,00	75,85	74,65	76,49	72,23

Tableau 11 : Comparaison des accuracy lors de l'étape 3b

Les écarts sont à nouveau plus faibles qu'à l'étape 2b. Nous observons en moyenne, une différence entre les accuracy des prédictions sur des observations traduites et originales de 6,37% (ou 3,68 points) pour les films et de 3,1% (ou 2,31 points) pour les produits ($p < 0.01$) ou une moyenne globale de 4,73%.

Si nous acceptons l'hypothèse que les datasets soient similaires et comparables, les pourcentages cités lors de des étapes 3a et 3b (6,76% et 4,73%) pourraient révéler, une fois le biais de la source écarté, la perte d'accuracy d'un modèle entraîné uniquement sur un certain type de données (originales ou traduites) lors de prédictions de la polarité de données de l'autre type. Lors de cette étape 3, les écarts de performance en pourcentage entre les résultats des étapes 3a et 3b favorisent encore le deuxième type de transfert (b).

Un dernier élément que nous avons remarqué avec cette étape 3 est que même après avoir remis nos deux datasets en deux classes, l'accuracy des modèles prédisant la polarité de données de films est clairement meilleur que ceux prédisant la polarité de données de produits (figures 10 et 11). Cela peut facilement s'expliquer par un plus grand nombre de tokens moyen pour les données de films, fournissant plus d'information. Mais cela veut aussi dire que les différences que nous avons obtenues entre les données de films et de produits lors des étapes précédentes ne peuvent pas être imputées à la différence entre deux classes et cinq classes.

	Etape 2		Etape 3	
	Films (2 classes)	Produits (multiclasse)	Films (2 classes)	Produits (2 classes)
Transfert a	11,40%	15,50%	7,22%	6,31%
Transfert b	9,80%	12,20%	6,37%	3,10%

Tableau 12 : Synthèse des écarts de performance entre les différentes étapes

En résumé, lors des étapes 2 et 3, les écarts de performances que nous avons obtenus (tableau 12) ont montré que (b) les modèles entraînés sur des données traduites prédisant la polarité d'observations originales semblent faire moins d'erreurs que (a) les modèles entraînés sur des données originales prédisant la polarité d'observations traduites. Autrement dit, le transfert de ressources consistant en la traduction des données d'entraînement de l'anglais vers le français pour développer un modèle semble préférable pour effectuer des prédictions sur de nouvelles observations francophones.

Enfin, un dernier élément mis en lumière par les étapes 2 et 3 est à mentionner. Lors de l'étape 1, les faux positifs et faux négatifs sont restés équilibrés, mais lors l'étape 2 et 3, un fort déséquilibre est apparu, montrant souvent des recall et precision assez éloignés. Par exemple, pour les matrices de confusion (figure 15), le déséquilibre est évident.

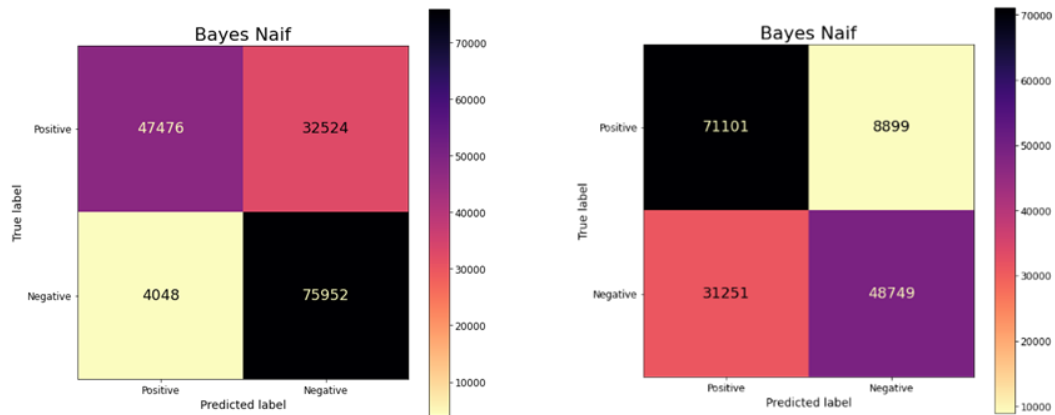


Figure 16 : Exemple de matrices déséquilibrées entre les faux positifs et faux négatifs

La première révèle un recall de 59,34% et une precision de 92,14%, quant à la deuxième, un recall de 88,87% et une precision de 69,47%. La régression logistique est moins affectée par cette problématique, mais présente quand même 50% de matrice déséquilibrée comparé au Bayes naïf qui en présente plus de 90%. Dans notre cas, nous avons défini une matrice déséquilibrée comme une matrice dont l'une des erreurs est au moins 50% plus présente que l'autre. Les faux positifs affectent plutôt les prédictions de produits et les faux négatifs les films, mais le faible nombre de dataset nous empêche de conclure par rapport à cette tendance.

4. Discussion

En regardant les écarts de performance lors des prédictions pour les modèles entraînés sur deux classes aux étapes 2 et 3, nous avons conclu qu'il est préférable d'opter pour le transfert de ressources et donc de traduire les datasets d'entraînement de l'anglais vers le français afin d'utiliser un modèle entraîné sur ces données pour effectuer des prédictions pour de nouvelles observations francophones. Néanmoins, une baisse de 4,8% entre les prédictions de ce modèle sur des observations traduites et originalement dans la langue du modèle est loin d'être négligeable et rend difficile de promouvoir cette utilisation en l'état. Effectivement, idéalement ce modèle devrait être capable de prédire des observations originales aussi bien qu'il le fait pour les observations traduites (le modèle ayant été entraîné sur des données traduites).

Ensuite, la baisse de 4,8% pourrait, dans la réalité, être bien différente. Dans notre méthodologie, nous imaginons un modèle entraîné avec des données traduites utilisé pour prédire d'observations originales, mais en réalité, il existe probablement aussi déjà des données étiquetées dans cette langue (même si elles ne sont pas en quantité suffisante) et nous pourrions envisager d'ajouter ces données traduites aux données originales disponibles pour avoir suffisamment de données d'entraînement. Dans cette optique, la partie des données traduites sera probablement de légèrement moins bonne qualité, mais elles n'impliqueront pas nécessairement une telle baisse d'accuracy lorsqu'elles sont mélangées avec des données originales.

Il est également à noter que l'étape 3 de la méthodologie n'a pas pu être effectuée sur des datasets multiclasse. De plus, les étapes précédentes ont comparé, pour le multiclasse, des datasets différents avec des labélisations différentes et l'impossibilité d'isoler ces effets, nous empêche de tirer des conclusions pour le multiclasse.

Enfin, les résultats que nous avons obtenus ne sont pas les seuls déterminants du choix du type de transfert, la manière dont l'utilisateur veut utiliser la traduction compte aussi. Pour le transfert d'environnement (la traduction des observations à prédire vers l'anglais pour utiliser les modèles en anglais), il est nécessaire de traduire dès qu'une observation apparaît alors que pour le transfert de ressources (la traduction des datasets anglophones en français), une grande quantité de données doit être traduite en une ou plusieurs fois. Il faut donc d'un côté traduire en continu et de l'autre une ou plusieurs fois en grande quantité. La traduction a aussi un cout computationnel non négligeable et le choix du type de transfert devra se faire en fonction du rapport entre quantité de données d'entraînement nécessaire et quantité de données à prédire.

5. Conclusion partie II

L'objectif principal de ce chapitre était d'étudier la possibilité d'utiliser la traduction pour effectuer un *transfert d'environnement* ou un *transfert de ressources* entre l'anglais et le français et déterminer le transfert optimal. Pour réaliser ce travail, nous avons récupéré 3 datasets disponibles, mais nous avons dû en créer un selon les mêmes règles que celui lui correspondant. Nous avons aussi traduits les 720.000 observations cumulées de ces datasets vers l'anglais ou le français selon la langue d'origine. Nous avons normalisé les 1.440.000 observations obtenues après traduction selon les quatre variantes que nous avons définies dans la section normalisation. Pour les dernières étapes du prétraitement, nous avons considéré des n-grammes où $n = 1, 2$ et 3 lors de la vectorisation ainsi l'utilisation de bag of words et TF-IDF. Lors de la phase d'entraînement, nous avons utilisé deux classifieurs, à savoir une régression logistique et un Bayes naïf.

Un sous-objectif de ce chapitre était de déterminer les prétraitements optimaux à appliquer lors d'une analyse de sentiment. Nous avons déduit de nos tests que les n-grammes où $n = 2$ ou 3 sont envisageables, selon la rigueur voulue. TF-IDF semble à privilégier, mais avec une faible avance sur le bag of words et il ne permet pas toujours d'obtenir les meilleurs résultats. De plus, dans notre cas, la variante avec aucune normalisation nous a permis d'obtenir les meilleurs résultats.

Nous avons mis en place une méthodologie en trois étapes pour répondre à la problématique principale. L'étape 1 cherchait à savoir si la traduction détruisait l'information nécessaire pour procéder à une analyse de sentiment. La méthodologie était empruntée à la littérature et elle consistait à comparer l'accuracy des prédictions d'un modèle entraîné et prédisant des observations d'un même dataset original avec l'accuracy des prédictions d'un modèle entraîné et prédisant des observations du même dataset, mais lorsque celles-ci sont traduites.

La réponse est qu'en effet l'information relative au sentiment semble assez peu détruite, car il est possible de classifier presque aussi bien après traduction. En effet, si on se concentre sur deux classes, nous avons en moyenne une diminution de l'accuracy de 0,93% entre les prédictions d'observations originales et traduites pour l'étape 1, avec un écart maximal de 1,5%. Nos résultats abondent aussi dans le sens de la littérature, surtout ceux de Ghafoor et al (2021), concluant aussi aux alentours d'un pourcent pour l'allemand.

Malheureusement cette étape 1 nous renseigne sur la qualité de la traduction, mais pas sur la possibilité de « mêler » données traduites et originales et donc d’opérer un transfert entre les langues. Il y a deux façons d’utiliser des données traduites pour procéder au transfert. Une première possibilité (a) est un *transfert d’environnement* en traduisant les observations à prédire vers l’anglais pour utiliser les modèles développés en anglais pour les prédictions. C’est pourquoi nous avons testé la capacité d’un modèle entraîné sur des données originales à prédire la polarité d’observations traduites (étapes 2a et 3a). Une seconde possibilité (b) est un *transfert de ressources* en traduisant les datasets d’entraînement anglophones vers les autres langues pour leur permettre de développer des modèles. C’est pourquoi nous avons testé la capacité d’un modèle entraîné sur des données traduites à prédire la polarité d’observations originales (étapes 2b et 3b).

Pour l’étape 2, nous avons une différence de 11,4% pour un modèle entraîné sur des données originales et 9,8% pour un modèle entraîné sur des données traduites. Mais l’étape 2 instaurait un biais dans la méthodologie en recourant à un train set et un test set qui partagent le même dataset. L’étape 3 permet de se libérer de ce biais et nous offre une différence de 7,2% pour un modèle tiers entraîné sur des données originales et 4,8% pour un modèle tiers entraîné sur des données traduites.

Enfin, les résultats des étapes 2 et 3 révèlent des écarts bien plus importants qu’e ceux aperçus à l’étape 1 en plus d’introduire des matrices de confusion déséquilibrées. Il n’est donc pas aisé de prédire des données traduites ou originales avec un modèle qui n’a été sensibilisé qu’à l’autre type de données. Néanmoins, dans notre analyse, la prédiction de la polarité d’observations originales avec un modèle entraîné sur des données traduites semble avoir été celle subissant la plus petite baisse de performance, ce qui nous pousse à promouvoir le *transfert de ressources*, c’est-à-dire, traduire les datasets d’entraînement anglophones vers les autres langues pour développer des modèles dans ces autres langues.

Une autre contribution de ce mémoire est la mise à disposition de l’ensemble des datasets utilisés dans ce travail, dont le dataset de commentaires de film ayant dû être créé pour ce mémoire ainsi que toutes les traductions et prétraitements effectués, rendant disponible 1.440.000 observations étiquetées pouvant être utilisées dans des applications de machine learning.

Conclusion

Il est de plus en plus intéressant de déterminer le sentiment derrière du texte non structuré, comme des commentaires sur internet. Il est possible de le révéler grâce à une analyse de sentiments, mais pour laquelle la majorité des ressources existent pour l'anglais. L'objectif de ce mémoire était donc de déterminer s'il était possible d'utiliser la traduction pour opérer un transfert entre l'anglais et le français. Pour répondre à cette question, nous avons réalisé une analyse quantitative où nous avons testé les deux transferts possibles par une expérimentation sur données réelles.

Notre travail a montré que la traduction détruit peu l'information relative au sentiment, mais modifie suffisamment l'expression du langage pour qu'un modèle n'ayant été habitué qu'à un type de données (langue originale ou traduite) performe moins bien sur l'autre type de données. Néanmoins, si un type de transfert devait être recommandé, ce serait celui de la traduction des datasets d'entraînement anglophones vers le français. Néanmoins, le coût à payer pour utiliser des données traduites est à nuancer dans la réalité. Notre travail a montré que le problème est moins sur la perte d'information causée par la traduction, mais plutôt lorsqu'un modèle n'a pas été confronté au type de données qu'il prédit. Il est donc probable de corriger les performances inférieures d'un modèle entraîné avec des données traduites (par exemple de l'anglais vers le français) et prédisant des données originales (en français) en ajoutant à l'entraînement des données originales (en français). Nous proposons donc d'entraîner des modèles hybrides avec des données traduites et les données originales disponibles et tester leur capacité de prédictions sur des observations originales. Un ratio optimal entre données originales et traduites pourra être déterminé, ce qui permettra de connaître la quantité de données traduites pouvant être ajoutées aux données originales disponibles pour maintenir de bonnes prédictions.

En permettant de fournir des datasets et donc des modèles en d'autres langues, ce *transfert de ressources* pourrait ouvrir de nouvelles possibilités pour les applications de l'analyse de sentiment en d'autres langues, comme connaître l'opinion générale sur les réseaux sociaux, repérer les commentaires mécontents, voir haineux, vérifier la cohérence entre les notes et les commentaires ou encore mettre en avant du contenu sur les réseaux sociaux selon leur polarité, et idéalement le contenu considéré comme positif, pour éviter les phénomènes de bulle de filtre autour du contenu négatif. Mais au-delà, c'est une opportunité pour toutes les applications du NLP reposant sur du machine learning supervisé comme les chat bots, résumé automatique ou détecteur de faux commentaires. Ce transfert propose une solution à un problème fondamental du machine learning supervisé, celui de la disponibilité de données d'entraînement labellisées.

Recul critique et recommandations

Tout d'abord, nos résultats ne sont valides que pour le français, mais la littérature laisse présager des résultats similaires pour une langue comme l'allemand et potentiellement pour d'autres langues européennes proches (originellement ou géographiquement). En effet, dans une méthodologie similaire à l'étape 1 de ce travail, Ghafoor et al. (2021) ont obtenu des résultats proches des nôtres pour l'allemand, nous laissant croire qu'appliquer les mêmes méthodologies que celles présentées aux étapes 2 et 3 de ce travail pour l'allemand, produirait des résultats similaires à ceux que nous avons obtenus pour le français. Cependant, pour des langues plus éloignées de l'anglais, comme l'hindi et l'urdu, pour lesquelles les mêmes auteurs avaient obtenu des résultats moins satisfaisants, notre méthodologie apporterait probablement aussi des résultats plus décevants. En outre, pour toutes les langues, le potentiel d'utilisation des données traduites est aussi conditionné à la qualité et à l'amélioration de la traduction automatique.

Ensuite, nous restons prudents vis-à-vis des résultats, car nous n'avons utilisé que quatre datasets différents dont le contenu était uniquement des commentaires de films ou de produits vendus sur Amazon. Il est possible, par exemple, que le sentiment soit plus difficile à déterminer après traduction sur les réseaux sociaux. Une plus grande présence de l'ironie ou une utilisation plus forte de l'argot pourrait mettre à mal la traduction automatique.

Il serait aussi nécessaire de procéder à des méthodologies similaires en ayant recours à d'autres types de classifieurs ou d'autres approches en machine learning, d'autres méthodes de vectorisation comme le plongement lexical ou d'autres systèmes de traduction.

Enfin, comme mentionné lors de la discussion, nous ne pouvons pas conclure pour le multiclasse. Il devra néanmoins être évalué, car il est souvent utile. Par exemple, avoir recours à trois classes permet d'avoir une classe neutre et de permettre au classifieur de ne pas catégoriser un commentaire comme positif ou négatif si son taux de certitude est faible, évitant les faux positifs et négatifs. Cela sera probablement utile dans le cas des réseaux sociaux où de nombreux messages peuvent être considérés comme neutres.

Bibliographie

- Abdaoui, A., Tapi Nzali, M. D., Azé, J., Bringay, S., Lavergne, C., Mollevi, C. & Poncelet, P. (2015). ADVANSE: Sentiment, Opinion and Emotion Analysis in French Tweets. *DEFT: Défi Fouille de Texte*, Caen, France. hal-01222629
- Aditya, P., Abhilash B. & Abhijit M. (2020). IMDb Movie Reviews Dataset. *IEEE Dataport*. <https://dx.doi.org/10.21227/zm1y-b270>
- AL-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y. & Gupta, B. (2017). Deep Recurrent Neural Network vs. Support Vector Machine for Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews. *Journal of Computational Science*, 27, 386-393. <https://doi.org/10.1016/j.jocs.2017.11.006>
- Al-Shabi, A., Adel, A., Omar, N., & Al-Moslmi, T. (2017). Cross-Lingual Sentiment Classification from English to Arabic using Machine Translation. *International Journal of Advanced Computer Science and Applications*, 8(12). <https://doi.org/10.14569/IJACSA.2017.081257>
- Analytics & Insights. (2021). *10 applications du nlp*. Consulté le 7 juillet, 2021 sur <https://analyticsinsights.io/10-applications-du-nlp/>
- Ardi, M. (2020). *An Implementation of Semi-Supervised Learning*. Consulté le 8 aout, 2021 de Becoming, sur <https://becominghuman.ai/an-implementation-of-semi-supervised-learning-e0054ab4fa02>
- Balahur, A., & Turchi, M. (2012). Multilingual Sentiment Analysis using Machine Translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, 52–60, Jeju, Korea. Association for Computational Linguistics.
- Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech Language*, 28(1), 56-75. <https://doi.org/10.1016/j.csl.2013.03.004>
- Blard, T. (2020). *French sentiment analysis with BERT*. Consulté le 20 aout de Huggingface Site Web : <https://github.com/TheophileBlard/french-sentiment-analysis-with-bert>
- Brooke, J., Tofiloski, M., & Taboada, M. (2009). Cross-Linguistic Sentiment Analysis: From English to Spanish. In *Proceedings of the International Conference RANLP-2009*, 50–54, Borovets, Bulgaria. Association for Computational Linguistics.

- Can, E.F., Ezen-Can, A., & Can, F. (2018). Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data. *ArXiv preprint*.
<https://doi.org/10.48550/arXiv.1806.04511>
- Chaouche, Y. (2021). *Analysez vos données textuelles*. Consulté le 15 juillet de Openclassroom sur <https://openclassrooms.com/fr/courses/4470541-analysez-vos-donnees-textuelles>
- Conseil de l'Europe. (2021). *Histoire de l'intelligence artificielle*. Consulté le 9 juillet, 2021 sur <https://www.coe.int/fr/web/artificial-intelligence/history-of-ai>
- Dashtipour, K., Poria, S., Hussain, A., Cambria, A., Hawalah, A.Y.A., Gelbukh, A. & Zhou, Q. (2016). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. *Cognitive Computation*, 8, 757–771.
<https://doi.org/10.1007/s12559-016-9415-7>
- de Séguin, S. (2019). *Introduction au NLP – Partie 1/6 – Test de Turing, définition et premiers jalons*. Consulté le 9 juillet, 2021 de Clevy, sur <https://blog.clevy.io/conversationnel/introduction-au-nlp-1ere-partie/>
- Devika, M.D., Sunitha C. & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science*, 87, 44-49.
<https://doi.org/10.1016/j.procs.2016.05.124>
- Dhaoui, C., Webster, C. & Tan, L. (2017). Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*, 34(6), 480-488.
<https://doi.org/10.1108/JCM-03-2017-2141>
- Exper.ai. (2020). *Natural Language Understanding: What is it and How is it Different from NLP?* Consulté le 10 juillet, 2021 sur <https://www.expert.ai/blog/natural-language-understanding-different-nlp/>
- Fenner, R. (2018). Alibaba's AI Outguns Humans in Reading Test. *Bloomberg*. Consulté de <https://www.bloomberg.com/news/articles/2018-01-15/alibaba-s-ai-outgunned-humans-in-key-stanford-reading-test>
- Francq, P. (2020) *Web mining - Introduction générale*. Syllabus, Université Catholique de Louvain FUCaM Mons.
- Furqan, R., Imran, A., †, I., Arif, M., Saleem, U., Sang C. G. & Yar, K. (2019). Tweets Classification on the Base of Sentiments for US Airline Companies. *Entropy*, 21(11), 1078. <https://doi.org/21.10.3390/e21111078>

- Ghafoor, A., Imran, A.S., Daudpota, S.M., Kastrati, Z., Batra, R. & Wani M.A. (2021). The Impact of Translating Resource-Rich Datasets to Low-Resource Languages Through Multi-Lingual Text Processing. *IEEE Access*, 9, 124478-124490. <https://doi.org/10.1109/ACCESS.2021.3110285>
- Gupta, N. & Agrawal, R. (2020). Chapter 1 - Application and techniques of opinion mining. S. Bhattacharyya, V. Snášel, D. Gupta & A. Khanna (Eds.), *Hybrid Computational Intelligence : Challenges and Applications* (pp. 1-23)). Cambridge : Academic Press. <https://doi.org/10.1016/B978-0-12-818699-2.00001-9>.
- Inbenta. (2021). *Traitement Automatique du Langage Naturel*. Consulté le 7 juillet, 2021 sur <https://www.inbenta.com/fr/technologie/ia-nlp/technologies-du-langage-naturel/>
- Joarson. (2020). Le test de Turing est-il toujours valide sur l'IA après 70 ans ? *Lebigdata.fr*. Consulté le 7 juillet, 2021 de <https://www.lebigdata.fr/test-de-turing-validite>
- Keung, P., Lu, Y., Szarvas, G. & Smith, N. (2020). The Multilingual Amazon Reviews Corpus. *Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.48550/arXiv.2010.02573>
- Kolchyna, O., Souza, T.T., Treleaven, P., & Aste, T. (2016). Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. Mitra, G. and Yu, X. (Eds.), *Handbook of Sentiment Analysis in Finance Kindle Edition*. New York : Albury Books. <https://doi.org/10.48550/arXiv.1507.00955>
- Koppe, M. (2019). L'IA traque le harcèlement en ligne. *CNRS*. Consulté le 12 juillet, 2021 de <https://lejournel.cnrs.fr/articles/lia-traque-le-harcelement-en-ligne>
- L'OBS. (2021). *Dans ces 8 domaines, l'intelligence artificielle va changer nos vies*. Consulté le 7 juillet, 2021 sur <https://www.nouvelobs.com/societe/20180228.OBS2855/dans-ces-8-domaines-l-intelligence-artificielle-va-changer-nos-vies.html>
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093-1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Mukhtar, N. Khan, M. & Chiragh, N. (2018). Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. *Telematics and Informatics*, 35(8), 2173-2183. <https://doi.org/10.1016/j.tele.2018.08.003>

- Nguyen Hoang, L. (2021). *Traffiquant d'humains et marchand de haine*. Consulté le 28 janvier, 2022 de Science4All, sur <https://www.youtube.com/watch?v=sAjm3-IaRtI>
- Nikil, T. & Aloysius, A. (2019). A Comparative study of Lexicon based and Machine learning based classifications in Sentiment analysis. *International Journal of Data Mining Techniques and Applications*, 8, 43-47. <https://doi.org/10.20894/IJDMTA.102.008.001.005>
- Olivier, P. (2021). *Introduction au NLP*. Consulté le 9 juillet, 2021 de Ekino, sur <https://www.ekino.com/articles/introduction-nlp-partie-i>
- Poibeau, T. (2014). Le traitement automatique des langues pour les sciences sociales: Quelques éléments de réflexion à partir d'expériences récentes. *Réseaux*, 188, 25-51. <https://doi.org/10.3917/res.188.0025>
- Satinet, C., & Fouss, F. (2022). A Supervised Machine Learning Classification Framework for Clothing Products' Sustainability. *Sustainability*, 14(3), 1334. <https://doi.org/10.3390/su14031334>
- Shalunts, G., Backfried, G., & Commeignes, N. (2016). The Impact of Machine Translation on Sentiment Analysis. *The Fifth International Conference on Data Analytics*.
- Sullivan, K.O. (2018). *Comparing the Effectiveness of Support Vector Machines and Convolutional Neural Networks for Determining User Intent in Conversational Agents*. Masters dissertations, DIT, 2018. <https://doi.org/10.21427/D78B9J>
- Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, 2(1). <https://doi.org/10.1146/annurev-linguistics-011415-040518>
- Towards AI. (2021). *Sentiment Analysis (Opinion Mining) with Python — NLP Tutorial*. Consulté le 12 juillet, 2021 sur <https://pub.towardsai.net/sentiment-analysis-opinion-mining-with-python-nlp-tutorial-d1f173ca4e3c#8efc>
- Vaibhav, J. (2020). *Laplace smoothing in Naïve Bayes algorithm*. Consulté le 28 janvier, 2022 de Towards data science, sur <https://towardsdatascience.com/laplace-smoothing-in-na%C3%AFve-bayes-algorithm-9c237a8bdece>

Annexes

Cfr. Document Massagé 27481700 2022 Annexe.pdf

