Université Paris Nanterre

200 av. de la République
92001 Nanterre Cedex
www.parisnanterre.fr

École doctorale 139 :
Connaissance, langage, modélisation

Membre de l'université Paris Lumières

Hyun Jung KANG

Regards croisés sur les avis en ligne :

approches du TAL et de la linguistique de corpus

Thèse présentée et soutenue publiquement le 26/01/2021
en vue de l'obtention du doctorat de Sciences du langage
de l'Université Paris Nanterre

sous la direction de Mme Iris Eshkol-Taravella (Université Paris Nanterre)

Jury * **:**

Rapporteur·e :	Mme Farah Benamara	Maître de conférences – HDR, Université Paul Sabatier
Rapporteur·e :	M. Dominique Legallois	Professeur des universités, Université de Sorbonne-Nouvelle
Membre du jury :	Mme Caroline Brun	Chercheuse senior, Naver Labs Europe
Membre du jury :	M. Guillaume Desagulier	Maître de conférences – HDR, Université Paris 8
Membre du jury :	M. Jean-Luc Minel	Professeur émérite des universités, Université Paris Nanterre

* *Vous pouvez rajouter ou enlever des lignes au tableau, ou modifier les fonctions (remplacer Membre du jury par Rapporteur·e par exemple). N'oubliez pas de supprimer ces deux lignes de texte à la fin de votre rédaction.*

Université Paris Nanterre
École doctorale 139 – Connaissance, Langage, Modélisation
Traitement Automatique des Langues

THÈSE

présentée par

Hyun Jung KANG

soutenance présentée et soutenue publiquement le 28 janvier 2021
pour obtenir le grade de docteur en Science du Langage
Spécialité : Traitement Automatique des Langues

Regards croisés sur les avis en ligne : approches du TAL et de la linguistique de corpus

Thèse dirigée par :

Iris Eshkol-Taravella	Professeure des universités, Université Paris Nanterre

Rapporteurs :

Farah Benamara	Maître de conférences – HDR, Université Paul Sabatier
Dominique Legallois	Professeur des universités, Université de Sorbonne-Nouvelle

---

Membres du jury :

Farah Benamara	Maître de conférences – HDR, Université Paul Sabatier
Caroline Brun	Chercheuse senior, Naver Labs Europe
Guillaume Desagulier	Maître de conférences – HDR, Université Paris 8
Iris Eshkol-Taravella	Professeure des universités, Université Paris Nanterre
Dominique Legallois	Professeur des universités, Université de Sorbonne-Nouvelle
Jean-Luc Minel	Professeur émérite des universités, Université Paris Nanterre

Remerciements

Je tiens tout d'abord à remercier très sincèrement ma directrice de recherche, Iris Eshkol-Taravella, qui a accepté de m'encadrer en tant que doctorante. Elle a su m'encadrer tout en me laissant de l'autonomie dans mes choix et elle était toujours disponible pour me guider. Sans sa bienveillance, sa patience, son soutien et ses conseils précieux, je n'aurais pu mener à bien et à terme cette thèse.

Rien dans cette recherche n'aurait été possible sans Gaston Gross qui m'a recommandé ma directrice pour que je puisse faire cette thèse avec elle.

Je souhaite exprimer ma gratitude à Farah Benamara et Dominique Legallois pour avoir accepté de juger ce travail en qualité de rapporteurs. Merci également à Caroline Brun, Guillaume Desagulier et Jean-Luc Minel pour leur participation à ce jury et l'intérêt porté à mon travail.

Je tiens à remercier sincèrement Frédéric Isel et Florence Villoing en tant que directeurs du Laboratoire MoDyCo. Je les remercie de m'avoir accueillie au sein de ce laboratoire, où j'ai pu bénéficier d'un environnement scientifique et humain remarquable. Mes remerciements vont également aux doctorants de MoDyCo qui m'ont accompagnée tout au long de cette thèse (sans ordre particulier) : Laura, Gasparde, Chloé, Chunxiao, Jade, Aline, Alice, Élise, Ilaine et Seongmin. Je tiens à remercier tout particulièrement Laura pour sa gentillesse, sa disponibilité et ses qualités pédagogiques qui m'ont permis de progresser en français. Je suis aussi reconnaissante envers l'ensemble du personnel administratif du laboratoire pour leur aide et leur disponibilité : Myriam Djedi et Veronique Goret.

Je tiens également à exprimer toute ma gratitude à Gabriel Bergounioux de m'avoir accueillie au sein du Laboratoire Ligérien de Linguistique, où

j'ai passé ma première année de thèse. Je n'oublie pas les doctorants avec qui j'ai partagé de très bons moments, en particulier Hélène. Un grand merci donc à Hélène pour son accueil exceptionnel dès mon installation à Orléans, pour la gentillesse et la disponibilité dont elle a fait preuve à mon égard.

Je tiens à exprimer mes plus vifs remerciements à Soo Mi Lee, professeure à l'université d'Ewha, qui m'a initiée à la linguistique française depuis ma licence, et à Seong Heon Lee, professeur à l'Université Nationale de Séoul, pour leur conseil, leur soutien et pour l'attention qu'ils ont accordé à ma thèse.

J'adresse mes sincères remerciements de nouveau à Laura, Jade et Aline, ainsi que Clémentine, Pia, Tiffany, Jérôme et Olivier d'avoir consacré du temps à la relecture de ce mémoire et d'avoir fait des commentaires enrichissants.

Pour finir, je tiens à remercier de tout cœur mon entourage qui a tenu bon toutes ces années. Une belle pensée pour Inhye avec qui j'ai passé tous les samedis à la BnF avec ses merveilleux gâteaux et qui a été d'un grand soutien lors des périodes difficiles. Un énorme merci également à Jenny avec qui j'ai partagé des moments agréables pour sa gentillesse et sa bonne humeur. Toutes mes amitiés et mes sincères remerciements à mes amis en Corée qui ont été d'un grand soutien moral et m'ont apporté le réconfort nécessaire durant les périodes de doute.

Je souhaite plus que tout remercier ma famille, en particulier mes parents, qui n'ont cessé d'aider et d'encourager la petite dernière, durant cette thèse et bien avant. J'ai une pensée particulière pour ma maman, qui a toujours été à mes côtés à travers mes hauts et mes bas pendant toute la durée de ma thèse.

**

Je tiens également à remercier le gouvernement français et l'Ambassade de France en Corée qui m'ont permis de bénéficier de la bourse d'excellence Blaise Pascal pour la réalisation de cette recherche.

Table des matières

1 Introduction 1
- 1.1 Context 1
- 1.2 Notion d'évaluation 2
- 1.3 Présentation du corpus 5
- 1.4 Organisation de la thèse 6

PREMIÈRE PARTIE : ARRIÈRE-PLAN 9

2 État de l'art 11
- 2.1 Introduction 11
- 2.2 Évaluation en marketing expérientiel 11
- 2.3 Évaluation en linguistique 16
  - 2.3.1 Subjectivité et axiologie 17
  - 2.3.2 Niveau morphologique 19
  - 2.3.3 Niveau lexical 20
  - 2.3.4 Séquences lexico-grammaticales 25
  - 2.3.5 Niveau des séquences discursives 26
  - 2.3.6 Niveau énonciatif 28
  - 2.3.7 Théorie de l'appraisal 31
- 2.4 Évaluation en TAL 36
  - 2.4.1 Trois niveaux d'analyse 38
  - 2.4.2 Différentes techniques de classification 41
- 2.5 Positionnement 42

DEUXIÈME PARTIE : MODÉLISATION 43

3 Modélisation de l'évaluation 45
- 3.1 Introduction 45

3.2 Typologie 45
3.2.1 Opinion positive/négative/mixte 45
3.2.2 Suggestion 55
3.2.3 Intention 58
3.2.4 Description 60
3.3 Annotation manuelle 62
3.4 Conclusion 65

4 Analyse quantitative de l'évaluation 67
4.1 Introduction 67
4.2 Nettoyage et prétraitement 68
4.3 Analyse lexicométrique 71
4.3.1 Fréquence des lemmes 73
4.3.2 Analyse des mots-clés 74
4.3.3 TF-IDF des n-grammes 77
4.4 Distribution des parties du discours 83
4.5 Position : l'ouverture et la clôture 87
4.6 Conclusion 91

5 Détection automatique de l'évaluation 93
5.1 Introduction 93
5.2 Méthodologie générale 94
5.3 Préparation des données 95
5.3.1 Normalisation 95
5.3.2 Représentation vectorielle 95
5.4 Classification supervisée 108
5.4.1 Apprentissage de surface 108
5.4.2 Apprentissage profond 109
5.4.3 Validation croisée 111
5.5 Traitement des données déséquilibrées 112
5.5.1 Cost-Sensitive Learning (CSL) 112
5.5.2 Sur-échantillonnage 113
5.6 Résultats 115
5.6.1 Métrique d'évaluation 116
5.6.2 Apprentissage de surface 116
5.6.3 Apprentissage profond 121

5.6.4 Impact du sur-échantillonnage 123
5.7 Conclusion 128

TROISIÈME PARTIE : APPLICATION 131

6 Application de la typologie proposée 133
6.1 Introduction 133
6.2 Application du modèle au corpus portant sur l'hôtellerie 133
6.2.1 Corpus d'étude 134
6.2.2 Annotation manuelle 134
6.2.3 Analyse du corpus annoté 136
6.2.4 Annotation automatique 146
6.2.5 Pour conclure 149
6.3 Application du modèle sur le corpus en langue coréenne 149
6.3.1 Travaux en coréen 150
6.3.2 Corpus d'étude 153
6.3.3 Annotation manuelle 153
6.3.4 Analyse du corpus annoté 155
6.3.5 Pour conclure 157
6.4 Conclusion 157

QUATRIÈME PARTIE : CONCLUSION ET PERSPECTIVES 159

7 Conclusion et perspectives 161
7.1 Contributions 161
7.1.1 De l'élaboration du modèle à la détection automatique 162
7.1.2 Application de la typologie proposée 164
7.2 Perspectives 165
7.2.1 Détection des cibles et des évaluations 165
7.2.2 Résumé automatique 166
7.2.3 Visualisation 166
7.2.4 Développement d'un outil de la fouille d'opinion 167
7.2.5 Application à plus grande échelle 171

CINQUIÈME PARTIE : ANNEXES 173

A Études en TAL 175

A.1 Synthèse des tâches proposées aux campagnes d'évaluation 175

A.1.1 Défi Fouille de Textes (DEFT) 175

A.1.2 Semantic Evaluation (SemEval) 176

A.2 Synthèse des études dans la fouille d'opinions 178

B Exemples des outils de la fouille d'opinions 183

Bibliographie 185

Table des figures

1.1 Répartition des notes attribuées par les clients 6

2.1 Différentes approches de l'évaluation sensorielle 13
2.2 Boucle de la prescription 15
2.3 L'objectif de l'évaluation en marketing expérientiel 16
2.4 Vue globale de la théorie de l'appraisal 32

3.1 Typologie de l'évaluation 46
3.2 Distinction entre affect, émotions, sentiments et opinions . . 50
3.3 Capture d'écran du logiciel MAXQDA 63
3.4 Répartition des catégories d'évaluation 65
3.5 Synthèse de la typologie proposée 66

4.1 Nuage de mots de l'ensemble des données 72
4.2 Fréquences des lemmes 73
4.3 Mots-clés de chaque catégorie 76
4.4 Exemples des tri-grammes composés d'une même liste de mots 77
4.5 TF-IDF des bi-grammes 79
4.6 TF-IDF des tri-grammes 79
4.7 Clusters de réseaux de mots couramment utilisés ensemble 82
4.8 Diagramme en bâtons des deux variables (la catégories d'évaluation et la position) 84
4.9 Valeurs propres associées à chaque dimension 85
4.10 Analyse symétrique standard de l'analyse factorielle des correspondances 85
4.11 Diagramme en bâtons des deux variables (la catégories d'évaluation et la position) 89
4.12 Graphique d'association entre la catégorie d'évaluation et la position 90

5.1 Processus d'apprentissage supervisé 95
5.2 Représentation en sac de mots 97
5.3 Exemple de CBOW et *skip-gram* 99
5.4 Processus d'apprentissage de surface 109
5.5 Processus d'apprentissage profond 110
5.6 Sur-échantillonnage et sous-échantillonnage 114
5.7 Matrice de confusion normalisée (SVM) 121
5.8 Matrice de confusion normalisée (LSTM) 123
5.9 Synthèse de résultats (CSL, sur-échantillonnage) 129

6.1 Répartition des catégories d'évaluation (HotelFR) 136
6.2 Mots-clés de chaque catégorie (HotelFR) 140
6.3 Analyse symétrique standard de l'analyse factorielle des correspondances (HotelFR) 141
6.4 Graphique de lune 143
6.5 Graphique d'association entre la catégorie d'évaluation et la position (HotelFR) 145
6.6 Système d'écriture du coréen 152
6.7 Répartition des catégories d'évaluation (RestoKR) 155

7.1 Exemple de la détection de la cible et de l'opinion 166
7.2 Exemple de la détection de la cible et de la suggestion . . 166
7.3 Exemples de l'outil que nous proposons 169
7.4 Exemple de la détection de la cible et de l'opinion 170

B.1 Capture d'écran de l'outil Monkey Learn 183
B.2 Capture d'écran de l'outil Lexalytics 184
B.3 Capture d'écran de l'outil Social Searche 184

Liste des tableaux

3.1 Exemple d'annotation manuelle (RestoFR) 64

4.1 Comparaison du résultats selon les étiqueteurs (spaCy, StanfordNLP, Treetaggerwrapper) 69
4.2 Exemple des scores de TF-IDF 78
4.3 Tableau de contingence des deux variables (la catégories d'évaluation et la position) 89
4.4 Synthèse de l'analyse quantitative 92

5.1 Classement des traits par importance 107
5.2 Paramètres pour adapter CSL 113
5.3 Moyenne pondérée de la F-mesure (sac de mots/TF-IDF + apprentissage de surface) 117
5.4 Moyenne pondérée de la F-mesure (Word2vec + apprentissage de surface) 118
5.5 Moyenne pondérée de la F-mesure (apprentissage de surface) 119
5.6 Précision, rappel et F-mesure de chaque catégorie (SVM) . . 120
5.7 Moyenne pondérée de la F-mesure (apprentissage profond) . 122
5.8 Précision, rappel et F-mesure de chaque catégorie (LSTM) . 122
5.9 Moyenne pondérée de la F-mesure (apprentissage de surface) : CSL contre sur-échantillonnage 124
5.10 Moyenne pondérée de la F-mesure selon différentes méthodes d'échantillonnage (sac de mots/TF-IDF et apprentissage de surface) 125
5.11 Moyenne pondérée de la F-mesure (apprentissage profond) : CSL contre sur-échantillonnage 126
5.12 Précision, rappel et F-mesure de chaque catégorie (SVM) : CSL contre SMOTE 127

5.13 Précision, rappel et F-mesure de chaque catégorie (RF) : CSL contre SMOTE 127
5.14 Précision, rappel et F-mesure de chaque catégorie (LSTM) : CSL contre Random 127

6.1 Exemple d'annotation manuelle (HotelFR) 135
6.2 Moyenne pondérée de la F-mesure (HotelFR) 147
6.3 Précision, rappel et F-mesure de chaque catégorie (HotelFR-SVM) 148
6.4 Exemple d'annotation manuelle (RestoKR) 154

A.1 Tâches proposées au DEFT 175
A.2 Tâches proposées au SemEval 176
A.3 Synthèse des études dans la fouille d'opinions 178

Chapitre 1

Introduction

1.1 Context

Le Web 2.0, dit Web participatif, a fait exploser la quantité de données textuelles en multipliant les plateformes de communication. Dans ce contexte d'interconnectivité croissante et de développement de nouvelles technologies, il est très fréquent de se référer au bouche-à-oreille électroniques (*electronic word-of-mouth* ou eWOM) d'inconnus, facilement accessible sur des plateformes telles que les blogs, forums ou sites dédiés à la critique de divers services ou produits. En effet, l'eWOM est défini comme toute communication positive ou négative faite par des clients potentiels, actuels ou précédents, au sujet d'un produit ou d'une entreprise, qui est mise à disposition au grand public et aux organisations via l'Internet (Hennig-Thurau *et al.* 2004).

L'impact de l'eWOM sur la prise de décision des clients est important pour les services de restauration, qui ne peuvent pas être évalués avant la consommation. Choisir un restaurant particulier comporte un risque non négligeable, amenant finalement les clients à dépendre davantage de l'eWOM (Lewis et Chambers 2000). L'eWOM positif, généré par une expérience satisfaisante (dans cette thèse, au restaurant), a un impact significatif sur les restaurants, car il est susceptible de les aider à construire une image positive et de susciter des intentions d'achat chez les clients. En outre, une expérience gastronomique réussie encourage le comportement altruiste des clients vis-à-vis du restaurant et suscite chez ceux-ci une tendance psychologique à partager leur bonne expérience avec d'autres personnes en ligne (Jeong et Jang 2011). Les évaluations diffusées en ligne sont ainsi devenues très puissantes, tant pour les clients que pour les

fournisseurs.

Par ailleurs, une évaluation exprimée par un individu ne représente qu'un seul point de vue. Chaque individu peut avoir des intérêts et des idéologies hétérogènes, ce qui implique différentes perceptions et évaluations des expériences. En dépit de cette diversité, le traitement d'un ensemble d'évaluations émises par de nombreuses personnes non seulement permet au public concerné (le restaurant ou les clients, par exemple) de prendre des décisions, mais constitue également une évaluation générale partagée par la communauté. Par conséquent, bien que les évaluations soient subjectives, elles peuvent exprimer une certaine objectivité par leur accumulation, représentant ainsi des valeurs communes acceptées par la communauté.

Ainsi, c'est la notion d'évaluation, qui occupe une place importante dans l'ère numérique, que nous étudierons dans le cadre de cette thèse.

1.2 Notion d'évaluation

L'objectif de cette thèse est d'analyser la façon dont les clients expriment leur évaluation d'expériences vécues dans des restaurants en s'appuyant sur différents niveaux linguistiques : le niveau morphologique, le niveau lexical, le niveau énonciatif et pragmatique, etc.

Dans le cadre de cette thèse, nous considérons l'évaluation comme une notion étendue qui dépasse celle de l'opinion. Observons l'avis [1] : l'évaluation est constituée d'opinions qui portent sur plusieurs aspects du restaurant (le cadre, le service et la nourriture) dans la phrase *a* et la vitesse du service dans la phrase *b*, dont la polarité est positive (*a*) ou négative (*b*).

[1] **a.** *Cadre magnifique, serveurs très avenants, mets exquis.* **b.** *Seul bémol la lenteur du service.*

Or, l'évaluation ne se limite pas aux opinions et peut impliquer d'autres informations. Dans l'avis [2], après avoir exprimé ses opinions positives sur les divers aspects du restaurant (*a*), l'auteur donne une précision

sur son action future (*b*), c'est-à-dire sur son intention de revenir au restaurant.

[2] **a.** *Service souriant et agréable, mets et vins originaux et bons.* **b.** *Nous reviendrons !*

L'évaluation peut, par ailleurs, s'adresser simultanément à des publics différents (les clients potentiels et le restaurant évalué) : la phrase *a* de l'avis [3] représente une opinion positive à l'égard du restaurant, plus précisément de la cuisine et du service, tandis que la phrase *b* incite les clients potentiels à se rendre au restaurant donné.

[3] **a.** *Remarquable établissement où la cuisine et le service sont irréprochables.* **b.** *Un sans faute : rare par les temps qui courent alors n'hésitez pas, réservez et profitez de l'excellence !*

De même, dans les deux premières phrases de l'avis [4], l'auteur exprime des opinions positives (*a*). Il termine ensuite son évaluation par la phrase *b* évoquant des aspects à améliorer ; ces informations sont adressées non seulement aux clients futurs, mais aussi au restaurant.

[4] **a.** *C'est vrai que l'on se sent bien ici. La cuisine, le cadre, l'accueil, tout est parfait.* **b.** *Un service légèrement plus chaleureux, un peu plus de fanataisie dans la sélection des vins et ce serait encore mieux.*

Dès lors, l'évaluation englobe diverses informations et peut s'adresser à des publics différents. Ainsi, elle porte sur plusieurs aspects liés du restaurant, notamment la nourriture, le service, le prix et l'ambiance. Ce sont les clients qui partagent leurs expériences sur Internet ; il s'agit donc

d'évaluations laissées par des non-spécialistes. Ces évaluations s'adressent aux futurs clients, mais aussi aux restaurants visités et sont donc des sources d'information utiles pour les uns comme pour les autres. Les auteurs cherchent à aider les futurs clients à prendre de meilleures décisions, d'une part, et, d'autre part, à aider les restaurateurs à améliorer leur service.

En anglais, le terme *review* est utilisé pour désigner les évaluations en ligne déposées par les clients, qui signifie « *a critical evaluation*[1] (une évaluation critique) ; *an evaluation of a how sth is working*[2] (une évaluation du fonctionnement d'une chose) ». En français, le terme « avis » est généralement employé et correspond à « ce que l'on pense, ce que l'on exprime sur un sujet »[3]. Ainsi, bien que les deux termes représentent le même objet, leur signification n'est pas en conformité. À partir de ce constat, il semble nécessaire d'établir une distinction entre la notion d'avis et celle d'évaluation afin de lever la confusion.

L'évaluation est « une action d'évaluer, de déterminer la valeur de quelque chose »[4]. Il s'agit ainsi d'un procédé destiné à « estimer (une qualité, la valeur de quelque chose, le résultat d'une action) »[5]. En revanche, l'avis est le résultat d'une évaluation réalisée par les clients et se compose de texte et d'une note. **Le corpus d'avis est donc l'objet que nous étudions dans le cadre de cette thèse, et nous nous intéresserons aux particularités du procédé d'évaluation en ligne.**

Cette thèse vise à examiner sous différents angles des évaluations de restaurants publiées par des clients. L'étude réalisée permet de mettre en lumière les raisons qui poussent les clients à laisser leur avis, d'analyser les moyens linguistiques qu'ils utilisent pour exprimer et partager leurs expériences, ainsi que de relever les caractéristiques propres aux évalua-

[1] https://www.merriam-webster.com/dictionary/review [consulté le 4 novembre 2020].
[2] https://dictionary.cambridge.org/fr/dictionnaire/anglais-francais/review [consulté le 4 novembre 2020].
[3] https://dictionnaire.lerobert.com/definition/avis [consulté le 4 novembre 2020].
[4] https://www.larousse.fr/dictionnaires/francais/%c3%a9valuation/31794?q=%c3%a9valuation#31719 [consulté le 4 novembre 2020].
[5] https://cnrtl.fr/definition/evaluer [consulté le 4 novembre 2020].

tions rédigées en ligne. Pour atteindre cet objectif, nous nous sommes appuyée en premier lieu sur la méthode de linguistique de corpus. Nous avons ensuite tenté de réaliser une détection automatique des informations contenues dans les évaluations en faisant appel à des méthodes de Traitement Automatique des Langues (TAL). Nous avons testé ensuite la généralisabilité d'un modèle développé à partir d'un corpus relevant d'un autre domaine – celui de l'hôtellerie – et sur la base d'un corpus écrit dans une autre langue – le coréen. En résumé, cette thèse met en œuvre des méthodes de linguistique de corpus et du TAL pour analyser l'évaluation en ligne.

1.3 Présentation du corpus

Le corpus sur lequel s'appuie l'étude, que nous appellerons « RestoFR », est composé d'évaluations collectées sur un site Internet[6] dédié aux restaurants[7]. De manière générale, les évaluations peuvent être exprimées en texte libre ou à travers une notation (un chiffre, des étoiles, etc.).

Afin d'assurer l'homogénéité du corpus, nous avons sélectionné des restaurants situés à Paris et proposant une cuisine traditionnelle française. Les avis des restaurants ont également été choisies en fonction de leur date de parution : pour chaque restaurant, les avis antérieurs au mois de novembre 2016 ont été extraites, dans la limite de 200 avis. Nous avons donc sélectionné 87 restaurants, correspondant à 6 287 avis (3 109 avis de restaurants étoilés et 3 178 avis de restaurants non étoilés) et 17 268 phrases, avec une moyenne de dix mots par phrase.

Certains chercheurs affirment que la distribution des nombreux avis publiés sur le Web est biaisée : des études ont ainsi montré que les notes tendent vers le pôle positif (Ganu *et al.* 2009 ; Jurafsky *et al.* 2014 ; Potts 2010 ; Wachsmuth *et al.* 2014). Cette tendance biaisée à la polarité positive a également été observée dans la répartition des notes au sein de nos données, dans laquelle la majorité des avis est favorable (figure 1.1). Les notes de notre ensemble de données s'étendent de 1 à 10, 10 étant

[6] La Fourchette, https://www.lafourchette.com

[7] Cette tâche a été réalisée avec l'aide de Flora Badin, ingénieure d'études CNRS en ingénierie logicielle au laboratoire LLL.

la plus élevée ; leur répartition montre une moyenne de 9,2. En effet, cette tendance au positif n'est pas un phénomène particulier à Internet (Jurafsky *et al.* 2014). Il s'agit, en psychologie, de l'effet *pollyanna*, selon lequel la nature humaine a une forte propension à la positivité et à l'optimiste.

Cependant, l'avis peut ne pas correspondre, dans certains cas, aux notes attribuées. La note seule ne suffit donc pas à évaluer un restaurant : chacun perçoit ce dernier de manière univoque, selon de multiples aspects, auxquels est accordée une importance différente. Les contenus textuels ont ainsi été privilégiés puisqu'ils permettent de mieux cerner les avis que les scores numériques attribués par les clients.

En outre, afin de tester la généralisabilité des méthodes à d'autres domaines, nous étudierons également deux corpus supplémentaires : « HotelFR », un corpus français sur les hôtels et « RestoKR », un corpus coréen sur les restaurants. Ils seront détaillés dans le chapitre concerné (voir chapitre 6).

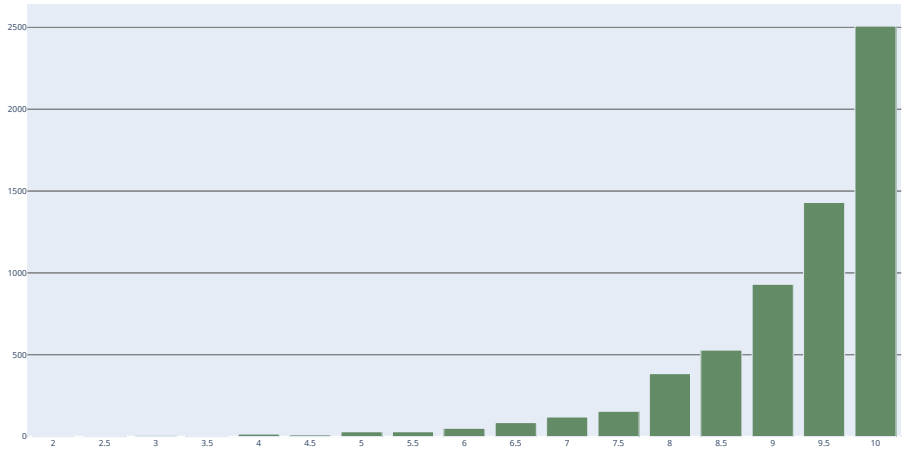


FIG. 1.1 : Répartition des notes attribuées par les clients

1.4 Organisation de la thèse

Cette thèse, fruit du travail effectué tout au long du programme de doctorat, est organisée en quatre parties. La première sera consacrée à

l'arrière-plan et présentera l'état de l'art dans les domaines dont relève la notion d'évaluation (chapitre 2). Nous examinerons cette dernière tout d'abord dans le domaine du marketing expérientiel, puis nous donnerons une vue d'ensemble des études réalisées en linguistique et en TAL, les deux domaines dans lesquels se situe notre travail. En linguistique, l'évaluation est définie à l'aide des notions de subjectivité et d'axiologie ; en TAL, c'est à travers la tâche de fouille d'opinions que ce domaine s'intéresse à l'évaluation.

La seconde partie concernera le développement du modèle qui a servi de cadre à l'étude de l'évaluation dans cette thèse. Cette partie sera composée de trois chapitres. Le chapitre 3 présentera le modèle d'évaluation élaboré sur la base d'une observation manuelle du corpus. L'analyse quantitative du corpus effectuée selon le modèle établi sera décrite dans le chapitre 4. Celle-ci a permis de mieux dégager les spécificités observées dans le corpus d'évaluations en ligne grâce aux méthodes de linguistique de corpus. Enfin, le dernier chapitre de cette partie (chapitre 5) présentera le module développé pour la détection automatique de l'évaluation. La démarche choisie s'appuie sur des méthodes couramment employées en TAL tel l'apprentissage supervisé (l'apprentissage de surface et l'apprentissage profond) et vise à obtenir de bonnes performances aussi bien sur les classes minoritaires que sur les classes majoritaires.

La troisième partie sera consacrée à l'évaluation du modèle établi sur d'autres données. Ainsi, le chapitre 6 démontrera la possibilité de généraliser notre modèle conceptuel à deux autres corpus : un corpus relevant d'un autre domaine (celui de l'hôtellerie) et un corpus écrit dans une autre langue (le coréen).

Enfin, la conclusion exposera les contributions et les perspectives qu'offrent cette thèse.

Première partie

ARRIÈRE-PLAN

Chapitre 2

État de l'art

2.1 Introduction

La vie quotidienne est aujourd'hui constamment préoccupée par le partage d'expériences, à cause de l'augmentation de la diffusion des avis sur Internet et sur les réseaux sociaux. Ces avis nous imposent l'évaluation faite par autrui d'une expérience, avant même que nous ayons pu la vivre par nous-même. L'évaluation des expériences s'appuie sur les réactions cognitives et affectives engendrées au cours de l'expérience (Giboreau et Body 2012) d'une part, et d'autre part, sur les principes axiologiques propres à l'acteur impliqué (Heinich 2017). Ces derniers renvoient plus précisément, à la convocation implicite ou explicite d'un ensemble de valeurs (Jackiewicz 2016), qui seront ensuite activés concrètement à travers des expressions verbales.

Dans ce chapitre, nous commencerons par une brève présentation de la notion de l'évaluation dans le domaine du marketing expérientiel, qui nous permettra de comprendre l'évaluation dans ce domaine. Dans un second temps, nous présenterons des travaux en linguistique portant sur l'évaluation, à travers divers niveaux linguistiques, suivis par des travaux en TAL relevant de la tâche de « la fouille d'opinions » dans laquelle la notion d'évaluation est également abordée.

2.2 Évaluation en marketing expérientiel

L'évaluation est un procédé utilisé dans une multitude de domaines (la sociologie, l'économie, la politique, psychologie, etc.). Aujourd'hui, dans le contexte du marketing, ce sont avant tout les composants intangibles,

irrationnels du produit qu'on cherche à vendre – comme l'expérience, les émotions et les sensations – qui détermineront l'évaluation (Giboreau et Body 2012). Ce nouveau courant est appelé « le marketing expérientiel » couvrant « tous les aspects de l'acte de consommation, de l'achat à l'utilisation du produit, de l'intuition au raisonnement, des émotions aux sensations » (Giboreau et Body 2012, p.3). Dans le cadre de cette étude, nous nous intéresserons à l'évaluation dans le domaine du marketing expérientiel dans lequel la voix du client a une place importante dans le but de comprendre comment les réactions cognitives et affectives, se déroulant lors de l'expérience, constituent l'évaluation.

L'ère postmoderne est caractérisée par une montée en puissance des valeurs individualistes et hédonistes (Body et Tallec 2015). L'individu ne se conforme plus à des valeurs considérées comme universelles, il cherche le plaisir et à agir selon ses envies. Les consommateurs ne donnent ainsi plus la priorité aux attributs fonctionnels et à la qualité d'un produit ou service, ces derniers étant considérés comme ce qui doit être offert au minimum. Ils sont désormais devenus plus guidés par l'affectif et recherchent des stimulations sensorielles et émotionnelles (Body et Tallec 2015).

Dans cette perspective, Pine et Gilmore (2011) proposent un nouveau type d'offre économique, il s'agit de l'expérience. Cette dernière recouvre l'ensemble des aspects de l'acte de consommation, de l'intention à l'action, de l'achat à l'utilisation, de l'intuition au raisonnement, des émotions aux sensations (Barbaray 2016). D'après Pine et Gilmore (2011), dans ce monde d'« économie d'expérience », les fournisseurs ne vendent pas que des produits ou des services, mais également des souvenirs. Le marketing expérientiel a donc pour objectif de dépasser les fonctionnalités du produit (ou service) et d'offrir des expériences positives et mémorables aux consommateurs. En intégrant la dimension sensorielle, l'évaluation se focalise ainsi davantage sur le processus de satisfaction que sur la satisfaction elle-même, permettant de créer des expériences qui enchantent le consommateur et stimulent ses sens.

L'évaluation de la dimension sensorielle des expériences se base sur une multitude de signaux perçus par l'individu. Dans les restaurants, par exemple, ce sont le goût et la présentation des plats, l'interaction avec les

personnels et la décoration intérieure qui contribuent ensemble à l'expérience de l'évaluateur. À partir de ces signaux, l'évaluateur peut ainsi se forger une impression, juger la situation de manière objective ou subjective, ce qui engendre, par conséquent, des sentiments positifs ou négatifs à l'égard de la situation. L'évaluation est menée soit auprès d'experts, soit auprès de consommateurs. Chacun a une mission différente : les experts sont impliqués pour des éléments liés aux premier niveaux de sensation, alors que les consommateurs interviennent pour des questions d'ordre cognitif supérieur, tels que l'interprétation des informations sensorielles au niveau socioculturel et personnel (Giboreau et Body 2012).

Par ailleurs, l'évaluation sensorielle est réalisée sur trois niveaux d'information. À chaque niveau, la mission concrète exigée auprès de chaque type d'évaluateurs (les experts et les consommateurs) est différente ; néanmoins, nous en exposerons uniquement les grandes lignes. Les trois niveaux (représentés par la figure 2.1) mettent en jeu, respectivement, des informations liées à l'entité donnée dans sa globalité (approches globales), des informations issues d'une analyse des caractéristiques (approches analytiques) et des informations procédant d'une interprétation subjective (approches interprétatives).

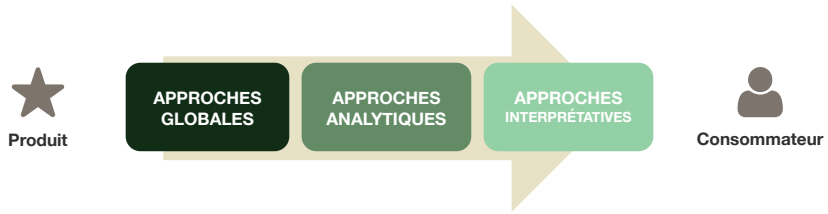


Fig. 2.1 : Différentes approches de l'évaluation sensorielle

Au premier niveau – celui de l'approche globale –, il ne s'agit pas de connaître l'avis personnel des individus ni d'en décomposer les diverses caractéristiques, mais d'appréhender et d'évaluer l'expérience dans son ensemble. Le second niveau est une approche analytique qui est liée à la fonctionnalité. À ce niveau, l'ensemble des attributs composant la dimension de fonctionnalité, comme la qualité des plats, le prix, l'environnement physique ou le service du personnel contribue à l'évaluation

de l'expérience, et constitue une réponse subjective transitoire. Le dernier niveau concerne les informations issues d'une interprétation subjective permettant à l'individu de catégoriser les signaux perçus selon deux tendances opposées (attractif ou répulsif), qui guident, par conséquent, son comportement vers une tendance soit d'approche, soit d'évitement (Bitner 1992). L'approche est associée au plaisir et à la satisfaction, résultant d'une sensation de confort, tandis que l'évitement concerne l'inconfort, qui peut modifier le comportement jusqu'à provoquer le rejet et la fuite (Mehrabian et Russell 1974). Pour renforcer la dimension de plaisir, les sources d'inconfort devraient être identifiées puis supprimées ; des alternatives pourraient, en outre, être proposées afin de convertir cet inconfort en source de plaisir et de satisfaction.

Depuis l'avènement d'Internet, l'évaluation de l'expérience par des consommateurs est devenue un important enjeu de marketing. Les consommateurs prennent plus de temps, avant l'achat, pour s'informer sur le produit et disposent d'informations de sources différentes (producteurs, consommateurs, experts, amateurs). Ceci leur permet d'élargir la palette de choix et de comparer : ils consultent des sites communautaires, des blogs ou même rejoignent des conversations sur des forums. Grâce à Internet, l'évaluation des consommateurs donne ainsi un nouvel essor au canal de consommation, et un pouvoir leur est donc conféré. L'évaluation marquant la fin d'un parcours pour un individu devient la première phase pour un autre ; il s'agit de « boucle de la prescription » (Body et Tallec 2015), comme le montre la figure 2.2.

Le parcours de consommation est composé de trois phases : la phase d'expérience, la phase de post-expérience et la phase de pré-expérience. La phase d'expérience correspond à la phase de consommation, où le consommateur rencontre le produit auquel sera lié l'expérience. Il juge la valeur générée par l'accord ou le désaccord entre les services attendus et perçus, qui déterminera sa satisfaction finale. La phase de post-expérience consiste à fidéliser plus encore les consommateurs. À titre d'exemple, un consommateur qui a une bonne expérience dans un restaurant aura une attitude favorable et sera enclin à y revenir. De plus, ce consommateur satisfait parlera de son expérience et recommandera le restaurant non seulement à son entourage, mais également aux autres. Avec le succès

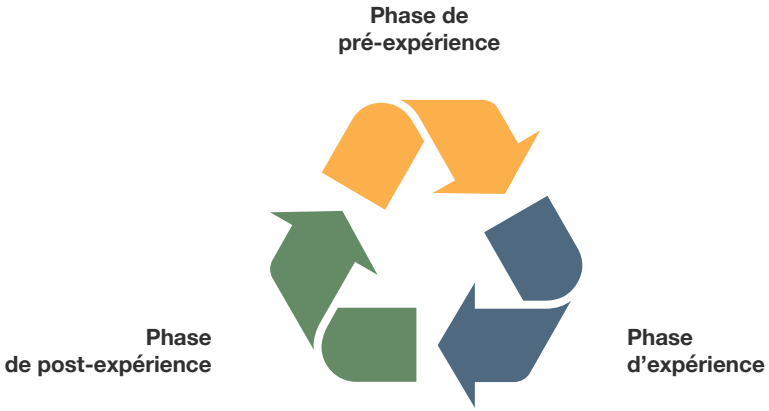


FIG. 2.2 : Boucle de la prescription

des réseaux sociaux et d'autres sites communautaires, qui disposent d'un pouvoir de viralité[8], une plus grande valeur est accordée au partage des évaluations. Ces dernières incluent les circonstances, le contexte général, les motivations ou les sentiments éprouvés. La satisfaction contribue ainsi à la fidélisation du consommateur et, grâce à eWOM généré, en incite d'autres à consommer. La phase de pré-expérience, enfin, est donc influencée par les avis rédigées en ligne, car un individu en quête d'information s'appuie souvent sur les avis et les recommandations pour prendre sa décision.

Par conséquent, Internet a créé une nouvelle relation entre fournisseurs et consommateurs, dans laquelle ces derniers disposent de plus de pouvoir qu'auparavant. Après avoir essayé un restaurant, le consommateur l'évalue sur Internet ; son avis devient ensuite l'une des principales sources d'information des autres consommateurs et fait partie de la conceptualisation de l'offre. En marketing expérientiel, afin de développer une connaissance approfondie du consommateur dans un contexte étendu, celui-ci est alors placé au cœur du parcours de consommation. Par ailleurs, l'évaluation s'efforce de rapprocher les deux publics (fournisseur et client), en confrontant

[8]Définie comme « diffusion rapide et imprévisible de contenus divers (photos, vidéos, etc.) sur Internet par effet viral, via les réseaux sociaux, particulier » d'après le dictionnaire Larousse (https://www.larousse.fr/dictionnaires/francais/viralit%c3%a9/10910800?q=viralit%c3%a9#11044522 [consulté le 23 septembre 2020]).

d'une part ce que le fournisseur veut contre ce qu'il a livré et, d'autre part, ce que le client attend contre ce qu'il a vécu (figure 2.3).

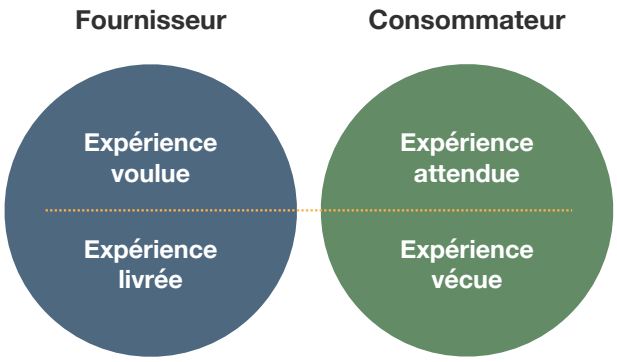


FIG. 2.3 : L'objectif de l'évaluation en marketing expérientiel[9]

Ainsi, l'évaluation en marketing expérientiel est un ensemble de procédures apportant des connaissances approfondies concernant l'expérience individuelle des consommateurs, qui permet de construire une vision complète et holistique sur le parcours de consommation.

2.3 Évaluation en linguistique

En linguistique, l'évaluation constitue un champ d'étude vaste et complexe, car il n'implique pas seulement les lexiques évaluatifs, mais également des phénomènes sensibles au cotexte et à l'information extralinguistique (Jackiewicz 2016). Nous aborderons en premier lieu l'évaluation à travers les notions de subjectivité et d'axiologie. Nous présenterons ensuite un ensemble d'études traitant l'évaluation selon les différents niveaux linguistiques. Enfin, nous nous arrêterons sur la théorie de l'*appraisal* proposée par Martin et White (2003), un cadre théorique conçu pour décrire l'évaluation d'une manière complète et systématique et nous montrerons également quelques limites de cette théorie observées dans le cadre de notre étude.

[9]Cette figure est une reprise du schéma présent dans (Body et Tallec 2015).

2.3.1 *Subjectivité et axiologie*

Les notions d'évaluation, d'opinion ou de jugement sont employées de manière équivoque dans différentes disciplines comme la psychologie, l'informatique et la linguistique. La confusion à l'égard de ces notions est due au double degré de la subjectivité (Vernier 2011).

La subjectivité, notion introduite par Benveniste (1966, p.259-260), est « la capacité du locuteur à se poser comme sujet dans et par le langage. [...] la subjectivité se détermine par le statut linguistique de la personne ». Selon Benveniste, *je*/*nous* et *tu*/*vous* sont des marqueurs renvoyant à la présence d'une personne dans l'acte de communication, alors que *il(s)* ou *elle(s)* font référence à un objet hors de la communication. Par ailleurs, la situation d'énonciation révèle la manière dont le locuteur se positionne dans son énoncé. Toutes les unités linguistiques faisant référence à la situation d'énonciation constituent ainsi des « embrayeurs » : les embrayeurs de personnes correspondent au locuteur (*je*, *nous*, *moi*, *mon*, etc.) ou au destinataire (*tu*, *vous*, *toi*, *ton*, etc.) et les embrayeurs spatio-temporels correspondent aux circonstances de l'énonciation – circonstances de lieu ou de temps (*ici*, *maintenant*, etc.). Les embrayeurs permettent de délimiter « l'instance spatiale et temporelle coextensive et contemporaine de la présente instance de discours contenant *je* » (Benveniste 1966, p.253) et constituent les traces d'une subjectivité dans le langage.

Cependant, il est légitime de questionner ce qui fait des avis un acte d'évaluation, au-delà d'un acte subjectif, car la subjectivité manifestée à travers les embrayeurs ne suffit pas à donner une vision de l'évaluation. La présence de *je* dans un journal intime ou une anecdote, par exemple, ne relève pas nécessairement de l'évaluation. D'après Kerbrat-Orecchioni (1980), les embrayeurs sont le premier degré de subjectivité explicite dans le discours. Kerbrat-Orecchioni note, en outre, qu'il existe un second degré de subjectivité, dit évaluatif et affectif. La réaction du sujet y est manifestée de manière implicite ou explicite grâce à des marqueurs de la subjectivité langagière, renvoyant à l'émotion (ou l'affect) et l'axiologie (ou l'évaluation).

Il semble alors indispensable d'interroger la nature axiologique[10] de l'évaluation pour décrire fidèlement cette dernière. L'évaluation est donc relative à l'expression d'un jugement, qui implique implicitement ou explicitement un système de normes et de valeurs dérivant des champs d'expérience humaine (Jackiewicz 2016). Nous pouvons le confirmer à partir de la définition proposée par les dictionnaires. D'après le TLFi[11], le verbe *évaluer* correspond, selon son premier sens, à « estimer, juger pour déterminer la valeur » et, selon le second, à « estimer, mesurer (une quantité, une grandeur, un ensemble mesurables) ». Pour le Larousse[12], il signifie « déterminer, fixer, apprécier la valeur, le prix de quelque chose ou d'un bien, etc. » ou « déterminer approximativement la durée, la quantité, le nombre, l'importance de quelque chose ». Le Robert[13] le définit comme le fait de « porter un jugement sur la valeur » ou de « déterminer (une quantité) par le calcul sans recourir à la mesure directe ».

Toutes ces définitions d'évaluation révèlent ainsi les différentes formes d'attribution de la valeur que distingue Heinich (2017) : la mesure, l'attachement et le jugement. La mesure renvoie à la simplification de l'évaluation par une mesure quantifiée, à titre d'exemple, les notes attribuées aux restaurants par les clients. Cependant, les évaluations que nous analysons dans cette étude sont plutôt liées à l'attachement et au jugement. L'attachement peut demeurer implicite, se manifestant par le biais de l'expression d'émotions ou de comportements. Le jugement, quant à lui, concerne « une problématique de la valeur d'un objet quelconque à une observation des opérations par lesquelles se manifeste concrètement la valeur attribuée par les acteurs à cet objet » (Heinich 2017, p.30). Cette dernière peut, par ailleurs, être observée uniquement à travers son application concrète dans les expressions verbales. Dès lors, la valeur

[10] L'axiologie « recouvre une zone sémantique, définie par un postulat empirique, qui renvoie à l'idée de préférence, de rupture de l'indifférence (étymologiquement : axios = « ce qui vaut »), id est à une logique binaire, à une polarité : positif/négatif, bon/mauvais, bien/mal » (Galatanu 2003, p.217)

[11] http://stella.atilf.fr/Dendien/scripts/tlfiv5/visusel.exe?11;s=4261488345;r=1;nat=;sol=0; [consulté le 8 juillet 2020].

[12] https://www.larousse.fr/dictionnaires/francais/%C3%A9valuer/31795 [consulté le 8 juillet 2020].

[13] https://dictionnaire.lerobert.com/definition/evaluer [consulté le 8 juillet 2020].

s'appuie sur « la description des situations où se produisent effectivement ces jugements, à l'intérieur de l'espace des possibles axiologiques ou, en d'autres termes, du système d'appréciation (pour soi) et d'expression (pour autrui) de la valeur » (Heinich 2017, p.31). Les valeurs peuvent ainsi être esthétiques (beau/laid), pragmatiques (utile/inutile, important/dérisoire, efficace/inefficace), cognitives ou intellectuelles (intéressant/inintéressant), hédonique-affectives (agréable/désagréable, plaisir/souffrance) et avoir une polarité positive ou négative (Galatanu 2003).

Ainsi, **l'évaluation renvoie au second degré de la subjectivité, à laquelle l'axiologie positive ou négative du locuteur est associée**. Elle fait intervenir un ensemble de normes et de valeurs qui s'inscrit dans le cadre des perceptions collectives, sensibles aux différences sociales et culturelles, et qui évolue au fil du temps.

L'évaluation est étudiée dans les travaux en linguistique selon différents niveaux que nous présenterons dans la section suivante.

2.3.2 *Niveau morphologique*

La morphologie évaluative a fait l'objet de nombreuses études (Bauer 1997 ; Dal 1997 ; Dressler et Barbaresi 2011 ; Fradin 2003 ; Grandi et Montermini 2003 ; Körtvélyessy et Stekauer 2011 ; Stump 1993). Elle désigne l'ensemble des procédés morphologiques (affixation, réduplication, apophonie, etc.) permettant de construire des lexèmes dont le sens inclut une évaluation (diminution, augmentation, péjoration, mélioration, etc.) à l'égard du lexème de base (Stosic et Amiot 2011). En français, elle se distingue par le biais des affixes évaluatifs : les préfixes (*super-*, *hyper-*, *sur-*, *hypo-*, *sous-*, etc.) et les suffixes (*-et*, *-ot*, *-asse*, *-on*, etc.). Ainsi, l'évaluation peut être exprimée par une combinaison de deux lexèmes : le lexème base qui offre une norme implicite et le lexème dérivé qui renvoie au décalage par rapport à cette norme (Stosic et Amiot 2011). En partant du lexème base *nager*, par exemple, le verbe *nageoter* implique une évaluation péjorative de l'action réalisée par le sujet concerné.

Pour décrire la morphologie évaluative, Mel'čuk (1994) relève deux dimensions liées à l'attitude du locuteur : la dimension mesurative et la dimension appréciative. La dimension mesurative est associée à une opposition petit/grand, tandis que la dimension appréciative repose sur

une opposition bon/mauvais. De même, Grandi (2002, 2015) distingue deux types d'évaluation morphologique en italien, chacun sur la base de deux oppositions : l'évaluation quantitative (petit/grand) et l'évaluation qualitative (bon/mauvais).

Inspirés à la fois de Mel'čuk (1994) et de Grandi (2002), Fradin (2003) présente une première tentative de recenser les préfixes évaluatifs en français. Il regroupe les préfixes évaluatifs autour de deux pôles : le pôle référentiel et le pôle locuteur. Le pôle référentiel fait intervenir une échelle réunissant l'ensemble des interprétations de Mel'čuk (1994) sur la dimension mesurative et celles de Grandi (2002) sur l'évaluation quantitative. Fradin (2003, p.60) précise qu'il s'agit « des significations ayant trait aux propriétés constitutives ou fonctionnelles de l'entité dénotée ». En revanche, le pôle locuteur correspond à la dimension appréciative et à l'évaluation qualitative ; il n'est pas lié à la propriété du référent, mais à la relation que le locuteur pose avec le référent.

En s'appuyant sur les deux pôles, plusieurs études ont été menées sur la morphologie évaluative en français, telles que l'étude des verbes évaluatifs (Amiot 2012 ; Amiot et Stosic 2015), l'étude comparative entre le français et d'autres langues (Efthymiou 2017 ; Stosic et Amiot 2011) et une sous-catégorisation sémantique des préfixes évaluatifs (Lefer et Grabar 2013, 2014).

2.3.3 *Niveau lexical*

Les études au niveau lexical sont nombreuses dans la littérature linguistique et portent plus particulièrement sur l'analyse et la constitution des lexiques évaluatifs.

2.3.3.1 Analyse des lexiques évaluatifs

De nombreuses d'approches traitant la dimension évaluative des textes se focalisent sur les lexiques particulièrement évaluatifs (ou subjectifs).

Chez Pupier (1998), on trouve une mention de la dualité de l'axiologie (favorable/défavorable) au niveau lexical. Selon lui, certains mots portent intrinsèquement une valeur positive (*bon*, *économe*) ou une valeur négative (*mauvais*, *avare*). Cependant, il note que les mots ne sont pas toujours

nettement positifs ou négatifs, car les préférences et les idéologies du locuteur, mais aussi le cotexte et le contexte peuvent influencer l'orientation axiologique. Ainsi, l'axiologie du même mot peut changer d'une acceptation à l'autre, en fonction de différents facteurs.

De son côté, Kerbrat-Orecchioni (1980) a étudié les lexiques axiologiques dans le cadre de sa théorie des « subjectivèmes ». L'auteure distingue les unités lexicales selon un axe d'opposition objectif/subjectif qui n'est pas dichotomique, mais graduel. À la différence des lexiques objectifs, les lexiques subjectifs véhiculent une trace de la subjectivité du locuteur, dans la mesure où ils lui permettent de manifester son attitude vis-à-vis d'une cible de jugement. Elle précise que « les unités lexicales sont en elles-mêmes (en langue) chargées d'une dose plus ou moins forte de subjectivité » (Kerbrat-Orecchioni 1980, loc.1902[14]).

Par ailleurs, les lexiques subjectifs (ou les subjectivèmes) se décomposent eux-mêmes en deux sous-catégories : les lexiques affectifs, relatifs à l'émotion du locuteur vis-à-vis d'un objet, et les lexiques évaluatifs, qui renvoient à l'évaluation d'un objet ciblé par le locuteur. Dans ce sens, Jackiewicz (2016, p.43) note que « tout subjectivème n'est pas à priori axiologique, mais cette classe a une forte propension à actualiser ce trait en discours ». En outre, comme Pupier (1998), Kerbrat-Orecchioni (1980) souligne le fait que l'idéologie que le locuteur attribue à un objet impose une orientation vers un lexique précis, par exemple politique, libéral, national, etc.

Les mots intrinsèquement axiologiques peuvent donner lieu à des classements selon leurs valeurs positives ou négatives, comme les adjectifs *bon*, *délicieux* ou *mauvais*. Cependant, l'orientation axiologique de certains mots est déterminée par des valeurs relatives qui sont sensibles aux différences culturelles, aux codifications sociales, à l'idéologie ou à la préférence de l'énonciateur et qui peuvent donc évoluer au fil du temps. Par exemple, la valeur axiologique de l'adjectif *rapide* semble changer selon la situation. Dans un restaurant étoilé, dans lequel le client prend le temps de savourer les plats et de profiter de l'ambiance, les consommateurs préféreront un service à vitesse modérée ; dans ce cas, la rapidité n'est pas nécessairement associée à une évaluation positive. En revanche, un restaurant fréquenté

[14] [version Kindle iOS].

par des employés pressés sera évalué positivement si leur service y est rapide. L'adjectif *élevé* offre un exemple de la polarité changeant selon le cotexte et le contexte. Lorsqu'une personne parle de la qualité, l'adjectif a une polarité positive, comme dans [1]. À l'inverse, lorsqu'il s'agit du prix, l'adjectif renvoie à une évaluation négative [2]. Ainsi, ces facteurs complexifient la détermination de l'orientation axiologique.

[1] *Le niveau de gastronomie reste toujours aussi élevé.*

[2] *Prix à mon sens très élevés.*

Les lexiques évaluatifs sont étudiés principalement autour de l'adjectif. Cheminée (2006), par exemple, se penche sur les adjectifs utilisés par les pianistes pour évaluer le son de leur instrument (*clair*, *brillant*, *vert*). Ces adjectifs sont spécifiques, c'est-à-dire qu'ils ont un sens différent de celui proposé dans un dictionnaire de français standard. Les adjectifs gardent néanmoins leur sens et possèdent simultanément une connotation positive ou négative. Jarukan (2014) mène une analyse sémantique sur les adjectifs axiologiques dans le discours touristique, en s'interrogeant sur la difficulté à classifier et à identifier la valeur axiologique ainsi que la polarité des adjectifs axiologiques. Dans le cadre sociolinguistique, Bensoussan (2009) analyse les adjectifs utilisés, au cours d'une enquête, par des personnes invitées à évaluer la qualité de leur logement. Dans ses travaux sur l'évaluation au sein des productions scientifiques, Tutin (2007, 2010a,b, 2014) a pour objectif de dégager les spécificités du lexique scientifique. Cependant, si les adjectifs sont directement associés au phénomène axiologique dans les travaux mentionnés (Bensoussan 2009 ; Cheminée 2006 ; Jarukan 2014), selon Tutin (2010a,b), dans les productions scientifiques, les adjectifs purement évaluatifs sont moins « productifs », car les auteurs préfèrent recourir à des adjectifs moins subjectifs renvoyant au temps, à l'importance ou à la nouveauté.

En linguistique anglophone, Biber et Finegan (1988, 1989) proposent la notion de *stance*, qui renvoie « l'expression lexicale et grammaticale d'attitudes, de sentiments, de jugements ou d'engagements vis-à-vis du contenu

propositionnel d'un message (*the lexical and grammatical expression of attitudes, feelings, judgements, or commitment concerning the propositional content of a message*) » (Biber et Finegan 1989, p.93). Plus précisément, selon les auteurs, « les participants à une situation de communication peuvent exprimer différents degrés d'engagement et d'attitudes » (Grisot 2019, p.23), qui sont liés à l'évidentialité et à l'affect. L'évidentialité renvoie à « toute expression linguistique des attitudes à l'égard de la connaissance (*any linguistic expression of attitudes toward knowledge*) » (Chafe 1986, p.271). Elle est divisée en deux catégories sémantiques : la certitude et le doute. L'affect est quant à lui associé à « l'expression d'un large éventail d'attitudes personnelles, dont les émotions, les sentiments et les humeurs (*the expression of a broad range of personal attitudes, including emotions, feelings, moods*) » (Biber et Finegan 1989, p.94). Il possède des significations positives (*content*, *joyeux*) ou négatives (*bouleversé*, *malheureux*). Biber et Finegan (1988) s'intéressent d'abord aux adverbes (*personally*, *frankly*, *unquestionably*, *of course*, *apparently*) marquant la *stance*. Ils élargissent ensuite la portée des marqueurs de *stance* aux adjectifs, verbes, verbes modaux et marques d'atténuation (*hedges*) (Biber et Finegan 1989).

De nombreuses études se focalisent ainsi sur les unités lexicales fonctionnant comme des marqueurs d'évaluation. Néanmoins, comme le remarquent Polanyi et Zaenen (2006), la perspective lexicale ne suffit pas à rendre compte de la nature de l'évaluation là où le contexte permet de récupérer des informations sur l'axiologie des adjectifs, telles que la négation, les intensifieurs, les connecteurs, la modalité, le genre, le discours rapporté, etc. De même, Pupier (1998) met en évidence la nécessité de la prise en compte des préférences et idéologies du locuteur, d'une part, et du cotexte et du contexte, d'autre part. Legallois et Lenepveu (2014, p.15) affirment également que « l'évaluation n'est pas seulement un acte subjectif ou intersubjectif exprimé par le lexique et par les propriétés formelles des énoncés » et adoptent une approche qui prend en compte l'organisation du texte.

2.3.3.2 Constitution des ressources

En parallèle des analyses fines de termes évaluatifs, d'autres travaux visent à constituer des ressources. Il s'agit non seulement de termes évaluatifs, mais aussi de ceux sur les émotions qui peuvent également intervenir dans l'évaluation (voir chapitre 3). Toutefois, ces recherches sont davantage axées sur l'anglais que sur le français.

En anglais, les ressources couramment utilisées sont General Inquirer[15] (Stone *et al.* 1966), SentiWordNet[16] (Baccianella *et al.* 2010), MPQA Subjectivity Lexicon[17] (Wilson *et al.* 2005), NRC Word-Emotion Association Lexicon[18] (Mohammad et Turney 2010, 2013) et Semantic Orientation Calculator (SO-CAL)[19] (Taboada *et al.* 2011). Plus récemment, des lexiques ont été développés spécifiquement pour Twitter, comme Hashtag Sentiment Lexicon[20] ou Sentiment140 Lexicon[21] (Kiritchenko *et al.* 2014 ; Mohammad et Turney 2010 ; Zhu *et al.* 2014). La plupart des ressources ne s'intéressent pas au domaine dont sont issues les données ; néanmoins, le projet SocialSent[22] (Hamilton *et al.* 2016) a constitué des ressources lexicales spécifiques à deux domaines, l'histoire et la communauté de reddit.com[23].

En revanche, il existe moins de ressources lexicales pour le français. Abdaoui *et al.* (2017) ont repris le NRC Word-Emotion Association Lexicon évoqué ci-dessus et l'ont traduit en français, créant le French Expanded Emotion Lexicon (FEEL). Leur travail s'est composé de deux étapes : les lexiques ont été d'abord traduits automatiquement grâce à un traducteur en ligne, puis un traducteur humain professionnel est intervenu pour valider l'association du terme et de son émotion correspondante. Dans le cadre du projet ANR Blogoscopie, un lexique de l'évaluation a été

[15] http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm [consulté le 4 juin 2020].
[16] http://sentiwordnet.isti.cnr.it/ [consulté le 4 juin 2020].
[17] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/ [consulté le 4 juin 2020].
[18] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm [consulté le 4 juin 2020].
[19] https://github.com/sfu-discourse-lab/SO-CAL [consulté le 4 juin 2020].
[20] http://saifmohammad.com/WebPages/AccessResource.htm [consulté le 4 juin 2020].
[21] http://saifmohammad.com/WebPages/AccessResource.htm [consulté le 4 juin 2020].
[22] https://nlp.stanford.edu/projects/socialsent/ [consulté le 4 juin 2020].
[23] https://www.reddit.com/ [consulté le 4 juin 2020].

constitué, comprenant 982 termes évaluatifs (Vernier 2011). La ressource constituée par Mathieu (2005, 2006) comporte un lexique sémantique de 600 verbes français décrivant les sentiments, les émotions et les états psychologiques.

2.3.4 *Séquences lexico-grammaticales*

Le lexique d'évaluation a été décrit également en termes de grammaire locale fondée sur l'analyse du contexte d'une unité lexicale étudiée. Ainsi, Legallois et Ferrari (2006) proposent une grammaire locale de l'évaluation, reconnaissant à cette dernière une portée importante au sein de diverses branches de la linguistique. Nous entendons par une grammaire locale un mécanisme formel représentatif, qui décrit « le cotexte d'une unité lexical donnée en tant qu'ensemble de configurations de mots (cf. Gross 1995 ; Maurel 1993) » (Buvet 2010, p.42). Les auteurs ont d'abord fait l'inventaire des séquences lexico-grammaticales portant sur l'évaluation et constituées de simples collocations (*conseiller vivement*) ou de configurations plus larges (*ne jamais (aus)si bien* + verbe de représentation/explication). Ces séquences ont ensuite été étiquetées selon les rôles thématiques propres à l'évaluation ou à la fonction de connecteur à portée énonciative. À titre d'exemple, une phrase *je suis déçu par ce livre* est étiquetées comme [Évaluateur] [décevoir] [Évalué], un syntagme *la fameuse trilogie* est étiqueté comme [adj. Évaluatème] [Évalué] et une locution adverbiale *quand même* comme connecteur argumentatif. Cette étude a plus tard été implémentée informatiquement dans le cadre du Défi Fouille de Textes (DEFT) 2007 (Vernier *et al.* 2007).

En linguistique anglophone, Hunston et Francis (2000) introduisent ces séquences en tant que « grammaires de patrons » (*pattern grammar*) définies comme l'ensemble des mots et structures couramment associés au mot donné et contribuant à sa signification. Un patron peut donc être identifié (i) si une combinaison de mots se produit relativement fréquemment, (ii) si cette combinaison dépend d'un choix de mots particuliers, ou (iii) si une signification précise est associée à cette combinaison. Il s'agit donc d'identifier et d'étudier, dans le corpus, « les patrons distributionnels qu'intègrent certaines classes de mots sémantiquement homogénéisés par le pattern » (Legallois 2016, p.11).

Hunston (2010) a par la suite réalisé une analyse approfondie des grammaires de patrons, qui distinguent quelques types de patrons d'évaluation en anglais. À titre d'exemple, les patrons dits « performatifs », accomplit une fonction d'évaluation. Ces patrons sont principalement formés de *it* ou *there* : ***it*** *is obvious* ***that*** *(il est évident que) ou* ***there*** *is* ***something*** *special* ***about*** *(il y a quelque chose de spécial).* D'autres types de patrons peuvent également procéder une évaluation, mais seulement si le porteur de l'évaluation est le locuteur lui-même, sinon l'évaluation est rapportée par autrui, comme *People find it strange that people on a high income end up getting the same child benefit as everyone else (les gens considèrent qu'il est étrange que les personnes ayant des revenus élevés finissent par recevoir le même montant d'allocations familiales que les autres*). L'auteure affirme ainsi que la distinction entre les évaluations performatives et rapportées sont essentielles pour la mise en œuvre de la grammaire locale.

Ainsi, les travaux conduits au niveau des séquences lexico-grammatiques montrent que, lorsque des mots partageant une certaine similarité sémantique sont regroupés, on peut leur associer des rôles fonctionnels d'évaluation.

2.3.5 *Niveau des séquences discursives*

L'évaluation peut également dépasser les frontières du lexique et de morphosyntaxe et se manifester à l'échelle du texte entier. Hunston et Thompson (2000) posent l'évaluation comme un terme général couvrant les expressions de l'attitude, la *stance*, le point de vue ou les sentiments du locuteur quant aux entités dont il parle. Les auteurs considèrent que l'évaluation s'exprime autant par le texte que par les lexies et la grammaire et elle peut avoir trois fonctions :

- Elle exprime l'opinion du locuteur et reflète donc son système de valeurs et celui de la communauté concernée ;
- Elle établit une relation entre le locuteur et l'interlocuteur, en s'appuyant sur des attitudes, des valeurs et des réactions partagées. Le locuteur présente, manipule ou synthétise des informations, tandis

que l'interlocuteur, étant subordonné dans cette relation, ne peut pas s'y opposer, mais seulement se contenter de l'accepter ;

– Elle organise le discours. Le locuteur ne se restreint pas à exprimer son opinion et à interagir avec l'interlocuteur ; il présente également ses arguments avec une organisation choisie.

D'après les auteurs, les deux premières fonctions sont largement abordées dans les travaux relatifs à l'évaluation, alors que la troisième fonction a suscité moins d'intérêt.

En ce sens, Legallois et Lenepveu (2014, p.15) affirment que « l'évaluation [...] est un mode d'interaction entre les unités du texte (Legallois 2006) de dimensions locales et globales ». Certaines études menées sur les séquences discursives se fondent sur l'approche ascendante, qui observe les unités minimales de niveau inférieur pour ensuite reconstruire l'architecture du discours aux niveaux supérieurs ; des travaux autour de la Rhetorical Structure Theory (RST) (Mann et Thompson 1988), la Segmented Discourse Representation Theory (SDRT) (Asher et Lascarides 2003) ou la Discourse Lexicalised Tree-Adjoining Grammar (DLTAG) (Creswell *et al.* 2003 ; Webber 2004) supposent que les structures discursives couvrent différents niveaux, de la phrase aux paragraphes, jusqu'à l'ensemble du texte donné. Dans cette perspective, le discours est constitué d'unités élémentaires (UDE) qui servent de base à la construction d'une structure discursive.

L'étude de l'évaluation selon cet axe a soulevé des problématiques telles que le rôle des relations discursives dans l'analyse de la subjectivité et l'impact de la structure du discours sur la détermination de l'opinion globale véhiculée par un document (Benamara *et al.* 2017b). Dans ce cadre, Benamara *et al.* (2016) ont constitué le tout premier corpus (des avis sur des films en français et en anglais, des lettres envoyées à l'éditeur en français) annoté manuellement selon le modèle SDRT. Il s'agit de l'annotation à plusieurs niveaux concernant la structure complète du discours, l'orientation des unités de discours élémentaires et des expressions d'opinion, et leurs caractéristiques associées. Ce travail a permis d'étudier la corrélation entre le discours et la catégorie sémantique des expressions d'opinion, le rôle de l'évaluation dans l'identification des relations discursives ainsi que l'impact des relations discursives sur l'analyse de la

subjectivité et de la polarité. Leur étude montre ainsi l'importance de la structure discursive dans l'analyse de l'évaluation.

2.3.6 *Niveau énonciatif*

Un contexte est construit par un ensemble d'éléments linguistiques, stylistiques, paralinguistiques, situationnels, culturels et psychologiques (Prihodko *et al.* 2019). Le niveau énonciatif permet de comprendre les circonstances dans lesquelles l'évaluation s'est produite, offrant les informations qui y sont associées. Nous évoquerons en particulier les travaux sur la modalité, principalement autour de la théorie de l'énonciation, et les actes de langage, qui se sont surtout développés en pragmatique anglo-saxonne.

2.3.6.1 Modalité

Diverses architectures modales ont été élaborées en linguistique du français, dont certaines relèvent de l'évaluation.

Dans sa théorie des opérations énonciatives, Culioli (1980, 1985) propose une typologie des modalités réduite à quatre grands types, dont l'un est la modalité appréciative (ou modalité de type III) marquant un jugement qualitatif de l'énonciateur.

Charaudeau (1992) introduit trois types d'« actes locutifs » déterminés par l'implication du locuteur dans son propos : les actes allocutif, élocutif et délocutif. L'acte est dit allocutif lorsque l'interlocuteur est présent dans l'acte d'énonciation et « le locuteur lui impose le contenu de son Propos » (1992, p.574) ; dans l'acte élocutif, l'interlocuteur est absent et le locuteur « révèle sa propre position quant à ce qu'il dit » (1992, p.575) ; enfin, lors d'un acte délocutif, le locuteur et l'interlocuteur sont absents et « le locuteur laisse s'imposer le Propos en tant que tel, comme s'il n'en était nullement responsable ». Chaque acte locutif est spécifié par les « modalités énonciatives », par exemple, l'opinion et l'appréciation correspond à l'acte élocutif et la suggestion à l'acte allocutif (voir chapitre 3).

Vernier (2011) regroupe les modalités proposées par Charaudeau (l'opinion, l'accord/désaccord, l'appréciation et le jugement) et distingue deux types d'évaluation : l'évaluation logique et l'évaluation axiologique. La

première regroupe l'opinion et l'accord/désaccord en s'appuyant sur l'axe vrai/faux ou l'axe réel/irréel. Ce type d'évaluation n'active pas d'axiologie positive ou négative. En revanche, l'évaluation axiologique regroupe l'appréciation et le jugement, permettant au locuteur d'adopter une attitude favorable (positive) ou défavorable (négative) par rapport à l'entité évaluée.

De son côté, Gosselin (2010, 2017, 2018) fait une distinction entre les modalités appréciatives et axiologiques. La modalité appréciative relève de l'expression du désirable et de l'indésirable et donc à l'évaluation des objets et des procédés en prenant en compte les désirs et aversions. La modalité axiologique, elle, renvoie au jugement de valeur de nature morale, idéologique ou légale. Elle indique donc une évaluation du caractère louable ou blâmable des comportements, des actions ou des situations contrôlées par des agents.

Ainsi, les études sur les différentes architectures de valeurs modales ne distinguent pas nécessairement la subjectivité de l'énonciateur. Bien que des divisions de modalité différentes soient adoptées d'un travail à l'autre, la modalité évaluative est intégrée dans les divers systèmes.

2.3.6.2 Actes de langage

L'évaluation est conçue pour produire un effet sur l'interlocuteur (Ananko 2017), en articulant non seulement l'aspect sémantique, mais aussi l'aspect pragmatique du contexte communicatif. En outre, elle n'est pas nécessairement exprimée de manière explicite : elle exige des inférences de la part des interlocuteurs afin qu'ils accèdent au sens non littéral de l'énoncé. Les actes de langage ont ainsi le potentiel de véhiculer une évaluation qui se fonde sur le contexte.

La théorie classique des actes de langage repose sur une distinction entre les énoncés affirmatifs décrivant le monde et ceux accomplissant une action. Les premiers sont des énoncés constatifs, qui dépendent des conditions de vérité (vrai/faux), tandis que les seconds sont des énoncés performatifs, s'appuyant sur des conditions de félicité (heureux/malheureux). Dans cette dernière perspective, d'après Austin (1975), le langage permet de modifier une situation, notamment en agissant sur autrui. Plutôt que d'opposer énoncés constatifs et énoncés performatifs, Austin distingue trois actes différents qui peuvent être accomplis par le biais du langage : les actes

locutoires, qui consistent à dire quelque chose ; les actes illocutoires, qui désignent l'acte accompli sur l'interlocuteur par le locuteur ; et les actes perlocutoires, qui se rapportent aux conséquences indirectes visées par les deux actes précédents. Selon cette théorie, le locuteur non seulement exprime son attitude, mais vise également à provoquer une action chez l'interlocuteur. En ce sens, Arutyunova (1988, p.44)[24] affirme que lorsqu'un locuteur exprime une évaluation, il est enclin à faire une recommandation, rendant celle-ci analogue aux impératifs. L'impératif permet d'amener l'auditeur à effectuer une action ou à s'en abstenir, de montrer que le locuteur souhaite réaliser son propos et inciter l'auditeur à le mettre en œuvre (Condoravdi et Lauer 2012).

Le travail d'Austin a été repris et développé par Searle (1972), qui conserve l'idée selon laquelle la production d'un énoncé revient à accomplir un certain acte qui vise à modifier la situation des interlocuteurs. Selon lui, le contenu d'un énoncé résulte de sa force illocutoire, conjuguée à son contenu propositionnel. Des énoncés différents peuvent donc avoir le même contenu propositionnel tout en correspondant à des actes de langage différents. Par ailleurs, Searle introduit également la notion d'acte de langage indirect. Par opposition aux actes de langage directs, qui ne doivent leur existence qu'à la signification de la phrase énoncée, sans qu'un processus de dérivation illocutoire n'intervienne, les actes de langage indirects doivent être reconstruits par l'auditeur. Ils font appel à un arrière-plan d'informations linguistiques et non linguistiques mutuellement partagées et à la capacité d'inférence rationnelle de l'auditeur.

Tout comme Searle, Grice (1975) admet la possibilité de communiquer au-delà de ce qui est dit. Selon son modèle d'inférence, la situation de communication dépend du fait que l'interlocuteur identifie l'intention de communiquer du locuteur. L'interlocuteur d'un énoncé doit donc déduire la meilleure interprétation possible en partant du principe que le locuteur est rationnel, comme indique l'ouvrage de Moeschler et Auchlin (2009, loc.4470[25]) :

> « Dire qu'un locuteur L a voulu signifier quelque chose par X, c'est dire que L a eu l'intention, en énonçant X, de produire

[24] Nous avons consulté l'extrait traduit en anglais dans Ananko (2017).
[25] [version Kindle iOS].

> un effet sur l'auditeur A grâce à la reconnaissance par A de cette intention. »

Autrement dit, la compréhension du sens transmis par un énoncé n'est pas liée aux règles conventionnelles, mais plutôt à la capacité de déduction de l'interlocuteur. La communication ne fonctionne donc pas comme un processus d'encodage et de décodage, mais selon un principe de coopération entre les locuteurs, précisé par une série de maximes de la conversation (maximes de quantité, de qualité, de relation, de manière). Par conséquent, l'un des défis de la compréhension de l'évaluation – notamment celle qui est exprimée de manière implicite –, est que l'interlocuteur doit correctement repérer l'intention communicative du locuteur.

L'évaluation occupe ainsi une place importante dans la théorie des actes de langage ; elle permet d'éclaircir des problèmes spécifiques à la communication en tenant compte de ses propriétés fonctionnelles et sémantiques, ainsi que des phénomènes de réflexion de différents aspects de la réalité existante dans les activités de raisonnement et de langage de l'individu (Ananko 2017 ; Bezuhla et Romanchenko 2013 ; Prihodko 2018).

2.3.7 *Théorie de l'appraisal*

La théorie de l'*appraisal* offre une vision globale de l'évaluation permettant d'envisager des recherches appliquées (en TAL, par exemple). Elle a été développée par Martin et White (2003) dans le cadre de la linguistique systémique de Halliday (1970) et donne une description plus riche et plus précise de l'évaluation. La notion d'évaluation dans la théorie de l'*appraisal* dépasse les notions d'opinion et de sentiment car elle explore les différents niveaux d'expression de l'évaluation, allant de l'explicite à l'implicite.

Le modèle est construit autour de trois catégories (figure 2.4) : l'ATTITUDE, l'ENGAGEMENT et la GRADUATION[26].

[26]Nous utilisons les majuscules pour faciliter la distinction entre les notions.

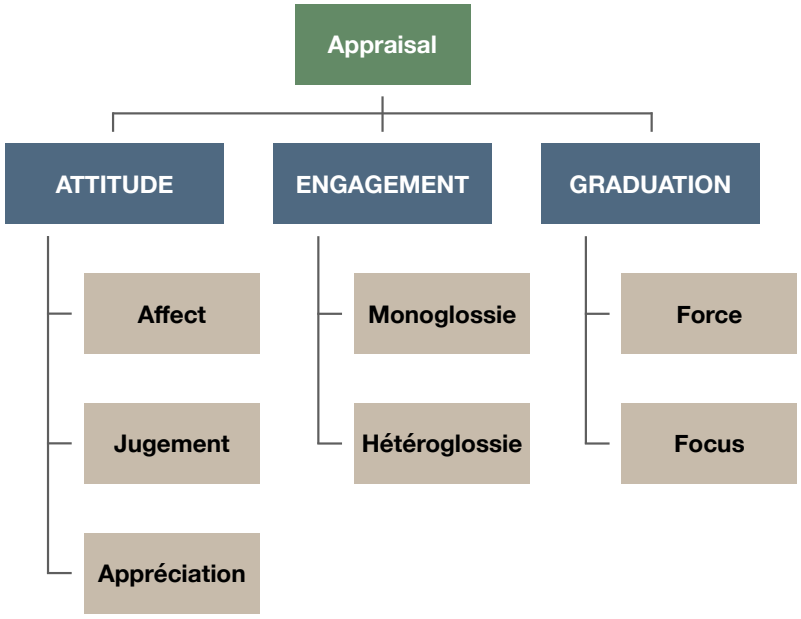


FIG. 2.4 : Vue globale de la théorie de l'appraisal

L'**ATTITUDE** est liée à l'interprétation d'entités variées (produits, services, comportement, etc.) par le locuteur et chargée de valeur positive ou négative[27]. Elle implique les sentiments, y compris les réactions émotionnelles, les jugements de comportement et l'évaluation des objets. L'étude de l'évaluation d'un restaurant – à laquelle nous nous intéressons particulièrement ici – concerne principalement cette interprétation des valeurs de chaque locuteur. L'ATTITUDE semble donc être la catégorie la plus pertinente pour ce travail.

Il existe plusieurs façons de véhiculer ou de mettre en avant une ATTITUDE ; certaines sont plus évidentes à l'identifier que d'autres. La plus simple consiste à utiliser des mots indiquant explicitement la position du locuteur. La réalité est, cependant, plus complexe. Tout d'abord, l'ATTITUDE est observée non seulement au moyen de lexies seuls, mais aussi à travers des propositions et des séquences de propositions du texte dans lequel elle opère. De plus, elle peut se manifester de manière explicite

[27] https://www.grammatics.com/appraisal/appraisalguide/unframed/stage1-attitude-affect.htm [consulté le 18 juin 2020].

ou implicite. Une ATTITUDE explicite s'exprime à travers des mots ou combinaisons des mots clairement évaluatifs. Une ATTITUDE implicite, en revanche, est moins évidente à identifier, car le locuteur s'appuie sur l'interlocuteur, qui évalue les événements comme justes ou faux, étranges ou normaux, attirants ou désagréables, réconfortants ou bouleversants.

L'ATTITUDE s'articule autour de trois zones sémantiques : l'émotion, l'éthique et l'esthétique, qui correspondent respectivement à l'affect, au jugement et à l'appréciation.

L'**affect**, impliquant la dimension de l'émotion, désigne les réactions émotionnelles du sujet parlant. Il peut se diviser en trois groupes :

- Le groupe du bonheur/malheur : des émotions qui représentent le bonheur ou la tristesse (la misère, l'antipathie, la joie, l'affection) ;
- Le groupe de la sécurité/insécurité : des émotions qui reflètent la confiance et l'inquiétude du locuteur à propos de son environnement (l'inquiétude, la surprise, la confiance, l'assurance) ;
- Le groupe de la satisfaction/insatisfaction : des émotions qui renvoient à la réussite ou à la frustration par rapport aux activités dans lesquelles le locuteur est engagé (l'ennui, le mécontentement, l'intérêt, le plaisir).

Le **jugement** concerne l'évaluation des personnes et de leurs comportements selon des normes sociales ou religieuses et peut être divisé en deux catégories :

- Le jugement de type « estime sociale » se réfère à l'évaluation des comportements à l'aune des normes sociales codifiées par des lois, codes et règlements qui dictent comment se comporter ;
- Le jugement de type « sanction sociale » correspond, quant à lui, à l'évaluation des comportements au regard des attentes sociales ou des systèmes de valeurs, moins précis et présents implicitement dans la culture.

L'**appréciation** est liée aux évaluations des objets sur la base de valeurs couvrant trois aspects :

- La réaction (les objets attirent-ils l'attention ? sont-ils agréables ?) ;
- La composition (sont-ils cohérents ? sont-ils complexes ?) ;
- La valeur (à quel point sont-ils innovants et authentiques ?).

L'**ENGAGEMENT** renvoie aux « signifiés par lesquels les locuteurs reconnaissent ou ignorent la diversité des points de vue que leurs messages mettent en jeu et à travers lesquels ils négocient un espace interpersonnel pour leur(s) propre(s) position(s) » (Jackiewicz 2016, p.59). Il s'agit de la façon dont un locuteur exprime son évaluation dans le discours qui est liée au dialogisme (monoglossique et hétéroglossique) de Bakhtine (1978, 1979). Lorsque la présence de l'opinion du locuteur ne concerne que lui, « il actualise le discours en utilisant sa voix : c'est le cas de la monoglossie » (Jarukan 2014, p.30). En revanche, si le locuteur « active son opinion dans le discours en utilisant la voix des autres par exemple, il s'agit d'hétéroglossie ».

La **GRADUATION** consiste, pour le locuteur, à faire varier l'intensité de l'évaluation par l'emploi d'éléments spécifiques. Elle se compose de deux sous-catégories : la force et le focus. La force exprime le degré d'intensité et de quantité par rapport à l'évaluation. En revanche, le focus correspond au degré de prototypicalité, qui peut prendre les mêmes valeurs que la force (Gardin 2009). La GRADUATION est notamment observée par le biais de certains modifieurs, tels que les intensifieurs et les atténuateurs, qui altèrent la valeur initiale d'un mot en agissant directement sur l'orientation de la polarité.

La théorie de l'*appraisal* permet ainsi d'acquérir une compréhension de l'évaluation et de ses réalisations linguistiques. Parmi les catégories (l'ATTITUDE, l'ENGAGEMENT et la GRADUATION), la première est étroitement liée à l'axiologie. L'évaluation d'un restaurant – sur laquelle porte notre étude – implique les trois sous-catégories de l'ATTITUDE (l'affect, le jugement, l'appréciation). Elle renvoie d'abord à l'affect, puisque le locuteur a la volonté de décrire son ressenti lors de l'expérience. Par ailleurs, le locuteur essaie d'inviter l'interlocuteur à partager sa réponse évaluative ; si l'interlocuteur l'accepte, l'évaluation est considérée comme pertinente et justifiée. Quant au jugement, certaines évaluations portent sur les personnels (gérant(e), serveur(se), chef) et leurs comportements, en

se fondant sur des attentes sociales ou des systèmes de valeurs imposés aux restaurants. Enfin, l'évaluation est également associée à l'appréciation, car le locuteur évalue les différents aspects du restaurant (le plat, le service, l'ambiance, etc.) d'après la valeur qu'il leur accorde.

Cependant, notre étude se heurte à une difficulté pour établir une distinction nette entre les différentes sous-catégories de l'ATTITUDE (l'affect, le jugement, l'appréciation) au sein de notre corpus d'avis sur des restaurants. Étant donné que l'affect implique les réactions émotionnelles et que l'un des composants de l'appréciation est la réaction à la cible évaluée, il n'est pas évident de différencier les deux sous-catégories sur le plan d'impact émotionnel.

Par ailleurs, Martin et White (2003) différencient l'appréciation et le jugement : l'appréciation est liée à l'évaluation d'un objet, alors que le jugement concerne plutôt des personnes et leurs comportements. Pourtant, la distinction entre ces deux sous-catégories, qui n'est fondée que sur la cible, a peu d'intérêt dans le cadre de la présente étude car nous nous intéressons pas si la cible est un objet (le plat et l'ambiance) ou une personne (les personnels et leurs comportements). En outre, une cible n'est pas toujours exprimée de manière explicite : la phrase [3], par exemple, ne permet pas d'identifier précisément la cible évaluée. De même, la phrase [4] renvoie à une évaluation globale du restaurant et, puisque sa cible n'est pas précisée explicitement (au-delà du restaurant), il peut s'agir des objets (les plats, la décoration), du personnel (le gérant, les personnels, le chef) ou des deux.

[3] *Juste parfait !*

[4] *Un excellent restaurant.*

En ce sens, Bednarek (2009a,b) critique la théorie de l'*appraisal* et propose un aménagement de la catégorie d'ATTITUDE. Selon l'auteure, les sous-catégories de cette théorie devraient être réparties, à des fins opérationnelles, au sein de métacatégories plus simples à distinguer et à manipuler. L'aménagement de la catégorie d'ATTITUDE qu'elle propose regroupe le jugement et l'appréciation dans une métacatégorie « opinion »

articulée autour de sept dimensions : prévisibilité, émotivité, importance, authenticité, pouvoir, fiabilité et causalité. En revanche, l'affect est englobé par la métacatégorie « émotion », qui possède quatre dimensions : bonheur/malheur, in/sécurité, in/satisfaction et surprise.

Ainsi, même si la théorie de l'*appraisal* offre un cadre théorique complet et systématique pour l'évaluation, la délimitation des frontières entre les sous-catégories de l'ATTITUDE qu'elle propose n'est pas entièrement satisfaisante.

Dans le chapitre 3, portant sur la modélisation de l'évaluation, nous nous intéresserons à l'analyse du fonctionnement de l'évaluation des clients qui évaluent. Nous chercherons plus précisément à montrer que les évaluations en ligne peuvent avoir plusieurs fonctions et contenir différentes informations. Le client met en ligne son évaluation pour donner son opinion, pour faire des suggestions, pour exprimer ses intentions ou pour décrire son expérience. Avant de procéder à l'élaboration du modèle conceptuel décrivant les évaluations observées dans le corpus, nous décrirons la manière dont la notion d'évaluation est abordée en TAL et la façon dont l'évaluation est automatiquement détectée dans les données.

2.4 Évaluation en TAL

Les avis en ligne aboutissent aujourd'hui à une production abondante de données véhiculant l'évaluation du consommateur vis-à-vis de son expérience. En TAL, la tâche de la fouille d'opinions s'intéresse à cette évaluation et est au centre de plusieurs campagnes d'évaluation tels que le Défi Fouille de Texte (DEFT) et Semantic Evaluation (SemEval) (voir l'annexe A.1).

Cependant, peu de discussions ont portées sur la différence entre deux tâches : la fouille d'opinions et l'analyse de sentiments. La première a initialement été employée par Dave *et al.* (2003) et la seconde par Nasukawa et Yi (2003). Toutefois, les recherches sur l'opinion (Das et Chen 2001 ; Morinaga *et al.* 2002 ; Pang *et al.* 2002 ; Tong et Koller 2001 ; Turney 2002 ; Wiebe 2000) ont débuté avant même l'apparition de dénominations spécifiques. En TAL, l'évaluation est désignée par le terme

d'« opinion », qui recouvre également les notions d'appréciation, d'attitude et de sentiment. En conséquence, l'évaluation est une notion mobilisée dans des tâches différentes, mais apparentées. Hemmatian et Sohrabi (2019, p.1498) présentent une série de désignations et les tâches associées à la fouille d'opinions :

- L'analyse des sentiments (*sentiment analysis*) : l'objectif est l'identification des sentiments et de l'opinion, considéré comme un domaine de la fouille de texte (*text mining*) ;

- L'extraction d'opinion (*opinion extraction*) : il s'agit du processus d'extraction des opinions à partir des documents disponibles sur Internet. L'objectif principal est de découvrir « la façon de penser des utilisateurs (*the users' ways of thinking*) » ;

- La fouille de sentiments (*sentiment mining*) : cette tâche a deux objectifs principaux, l'identification de phrases objectives ou subjectives, d'une part, et l'extraction des opinions et leur classification selon trois catégories (positives, négatives et neutres), d'autre part ;

- L'analyse de la subjectivité (*subjection analysis*) : elle permet d'identifier, de classer et de rassembler des phrases subjectives[28] ;

- L'analyse des affects ou émotions (*affect or emotion analysis*) : cette analyse permet de spécifier les éléments exprimant des émotions parmi les nombreux mots positifs ou négatifs d'un texte (Grefenstette *et al.* 2004).

- La fouille des avis (*review mining*) : un sous-thème de l'analyse des sentiments des textes (*text sentiment analysis*), elle a pour objectif d'extraire l'objet évalué à partir du sentiment exprimé par l'auteur et de créer un résumé des sentiments. D'autres recherches ont été axées sur les avis rédigés sur des produits par les utilisateurs (Zhuang *et al.* 2006).

[28] Ici, « les phrases subjectives » sont celles qui expriment des points de vue et des opinions subjectifs contrairement aux phrases objectives (Liu 2015).

Ainsi, l'objectif de la fouille d'opinions est de développer un système qui permet de reconnaître et d'extraire les expressions évaluatives ainsi que leur polarité (Liu 2015).

Nous nous focaliserons plus particulièrement, dans cette section sur les méthodes employées dans le domaine du TAL pour détecter l'opinion. Dans un premier temps, nous présenterons les trois niveaux d'analyse utilisés dans la fouille d'opinions pour détecter les opinions, les sentiments, etc. (document, phrase et aspect), puis nous décrirons des techniques couramment utilisées par cette tâche (l'apprentissage non supervisé, l'apprentissage supervisé de surface et l'apprentissage profond).

2.4.1 *Trois niveaux d'analyse*

Les travaux existants dans le domaine de la fouille d'opinions se distinguent selon trois niveaux d'analyse : les niveaux du document, de la phrase et de l'aspect.

2.4.1.1 Niveau du document

Le niveau d'analyse portant sur un document s'appuie sur le principe que chaque document (ou texte) est considéré comme un tout. Dans cette optique, cette méthode consiste à déterminer si un document est globalement positif ou négatif. Il s'agit d'une méthode restrictive qui s'applique à un seul objet. Cependant, en pratique, ce principe ne se vérifie pas toujours. Un document peut comprendre des opinions portant sur plusieurs objets (restaurants, films, etc.). En outre, même si un document ne contient des opinions qu'à l'égard d'un seul objet, les opinions peuvent porter sur plusieurs aspects, certains positifs et d'autres négatifs. Par exemple, si un avis sur un appareil électronique accompagné d'un score de 5/5 est globalement positif, on peut y trouver une phrase d'opinion négative mentionnant que la durée de vie de la batterie est courte. Une telle opinion est plus intéressante pour ceux qui cherchent une batterie de longue durée. Ainsi, ce niveau d'analyse est plus ou moins limité, du fait qu'il ne permet pas de discerner les différents aspects de l'objet évalué.

2.4.1.2 Niveau de la phrase

À ce niveau, le premier objectif consiste à opérer des distinctions entre phrases objectives et subjectives. Les premières se caractérisent par l'observation, ne faisant pas intervenir d'interprétation personnelle. Les secondes révèlent, en TAL, un sentiment, un point de vue, un jugement ou une croyance personnelle (Liu 2015). Lorsqu'une phrase est subjective, nous pouvons y associer une polarité positive, négative ou neutre, cette dernière renvoyant généralement à l'absence d'opinion.

Tout comme l'analyse au niveau du document, l'analyse au niveau de la phrase est quelques peu restrictive lorsque des détails supplémentaires sont recherchés (par exemple, quels aspects de l'objet évalué ont été appréciés ou détestés par des utilisateurs). Le défaut de ce type d'analyse est qu'il n'attribue qu'une seule polarité à la phrase, alors que dans certains cas une phrase peut comporter deux polarités distinctes et opposées comme dans l'exemple [5]. L'auteur donne une opinion négative sur le service (*quelques petits couacs*) et une opinion positive sur la cuisine (*surprenants*, *très bons*). Une personne accordant donc de l'importance au service ne se rendra probablement pas dans le restaurant donné.

[5] *Malgré quelques petits couacs de service, les plats proposés étaient surprenants et très bons.*

En outre, une autre limite de ce niveau d'analyse est qu'une opinion positive ou négative peut être inférée à partir d'une phrase objective. Autrement dit, les phrases objectives peuvent être traitées comme relevant de la subjectivité dans la mesure où elles impliquent un point de vue positif ou négatif sur l'entité concernée. Bien qu'une déclaration de l'exemple [6] soit objective à propos du service, elle implique une opinion négative, car elle décrit une situation non désirable.

[6] *L'hôtesse n'a pas été capable de trouver notre réservation.*

2.4.1.3 Niveau d'aspect

Ce niveau d'analyse a été introduit par Hu et Liu (2004) sous le nom de *feature level*, qui a ensuite été remplacé par le niveau de l'*aspect-based sentiment analysis* (niveau ABSA) par Liu (2012). Même si le document évalue un objet comme positif, cela n'implique pas que la perception de l'auteur soit positive sur tous les aspects de cet objet. De même, un document d'avis négatif ne signifie pas que l'auteur a un positionnement négatif à propos de tout. En ce sens, Liu (2012) a proposé d'extraire l'opinion comme un tuple composé d'éléments représentés par les lettres E, A, S, H et T :

- ***Entity*** (E) : l'entité (ou l'objet) représentant la cible d'opinion (par exemple, le restaurant, le film, etc.) ;
- ***Aspect*** (A) : l'aspect de l'entité sur lequel l'opinion est exprimée (la cuisine, le service, l'ambiance, etc.) ;
- ***Sentiment*** (S) : l'orientation sémantique de l'opinion (positive, négative, neutre) ;
- ***Holder*** (H) : le porteur de l'opinion ;
- ***Time*** (T) : le temps d'émission de l'opinion.

Cette approche permet ainsi de creuser l'analyse des différents aspects d'une entité ou des diverses opinions portées sur celle-ci. Dans l'exemple présenté précédemment [5], les opinions portent sur deux aspects (le service, le plat) d'une seule entité (le restaurant). Considérons un autre exemple [7]. Plusieurs opinions sont exprimées ici (*très souriant*, *avenant*, *pas très professionnel*) sur un seul aspect (le service). Ainsi, ce niveau d'analyse permet une description fine des opinions, pour des applications pratiques.

[7] *Le service, très souriant et avenant, pas très professionnel.*

2.4.2 *Différentes techniques de classification*

La fouille d'opinions est souvent conçue comme un problème de classification des textes. Les techniques de classification couramment employées peuvent être scindées en trois groupes : l'apprentissage non supervisé, l'apprentissage supervisé de surface et l'apprentissage profond. Dans cette section, nous présenterons ces techniques d'une manière synthétiques. Elles seront détaillées par la suite dans le chapitre 5.

Les travaux issus de l'apprentissage non supervisé effectuent la classification automatique en exploitant des patrons syntaxiques d'une part, et, d'autre part, des ressources lexicales, constituées d'une liste de mots portant une orientation sémantique (*semantic orientation* ou SO) positive ou négative. La première étude utilisant l'apprentissage non supervisé dans le cadre de la classification des avis est celle de Turney (2002), qui a recours à des patrons syntaxiques et à des étiquettes morphosyntaxiques. Une autre approche non supervisée est fondée sur les ressources lexicales déjà présentées dans la section 2.3.3.2. Le principe de cette méthode est qu'elle réalise une classification automatique sur la base d'un dictionnaire de lexiques subjectifs, avec les orientations et le degré de subjectivité/polarité qui leur sont associés. Cette approche a été adoptée pour la tâche de classification au niveau de la phrase par Kim et Hovy (2004) et au niveau des aspects par Hu et Liu (2004).

La classification automatique des textes est un objectif phare de la fouille d'opinions et de l'analyse de sentiments. Elle consiste à associer automatiquement une classe définie à chaque donnée d'entrée. L'apprentissage de surface et l'apprentissage profond supervisés sont aujourd'hui les techniques dominantes pour atteindre cet objectif. Les deux approches nécessitent des étiquettes préalablement associées aux données, par exemples « spam » ou « non-spam » pour identifier la nature d'un email, « positive » ou « négative » pour classifier les documents en polarité. Les classifieurs courants de l'apprentissage de surface visant à classifier les textes sont les modèles linéaires (Logistic Regression et Support Vector Machine), le modèle génératif (Naïve Bayes) et les modèles ensemblistes (Random Forest et XGBoost) (Aggarwal et Zhai 2012 ; Bhatia *et al.* 2020 ; Khan *et al.* 2010). En ce qui concerne l'apprentissage profond, les réseaux de neurones convolutifs (*convolutional neural network* ou CNN) et les ré-

seaux récurrents à mémoire court et long terme (*long short-term memory* ou LSTM) sont aussi couramment exploités dans la fouille d'opinions.

Pour clore cette section, nous présentons, dans l'annexe A.3, une synthèse de quelques études dans la fouille d'opinions en fonction du niveau d'analyse et des techniques utilisées.

2.5 Positionnement

Suite à la présentation d'études autour de la notion d'évaluation dans différents domaines, nous sommes parvenus à positionner notre étude vis-à-vis de l'état de l'art.

Notre travail se situe dans la lignée des recherches en TAL portant sur la fouille d'opinions au niveau de la phrase en exploitant deux méthodes de l'apprentissage supervisé (apprentissage de surface et apprentissage profond). Cependant, à la différence de nombreux travaux dans ce domaine, nous dépasserons largement la notion d'opinion positive ou négative. Dans cette optique, nous nous interrogerons sur la manière dont une évaluation émerge dans le langage en tant que jugement axiologique. De nombreux travaux en linguistique présentés dans ce chapitre abordent l'évaluation en fonction des différents niveaux linguistiques, mais il existe peu d'études regroupant ces divers niveaux. Notre travail s'inscrit dans cette dernière démarche. Il vise d'étudier, d'une part, les différentes formes et fonctions de l'évaluation présentes dans le corpus des avis en ligne, et d'autre part, la façon dont l'évaluation crée une axiologie positive ou négative.

Nous adopterons, d'abord, une approche qualitative pour décrire la nature de l'évaluation par le biais de l'observation du corpus avant de procéder à une approche de linguistique corpus à partir des moyens offerts par l'annotation manuelle. Cette double approche apportera un éventail des différentes caractéristiques liées à l'évaluation. Ensuite la détection automatique sera réalisée à partir du modèle appris sur les annotations manuelles, suivie par une évaluation de l'application de notre modèle proposé.

Deuxième partie

MODÉLISATION

Chapitre 3

Modélisation de l'évaluation

3.1 Introduction

Afin d'étudier l'évaluation des expériences vécues dans les restaurants, plus particulièrement la nature de l'évaluation, la façon dont les clients évaluent les restaurants, ainsi que les motivations des clients à exprimer une évaluation en ligne, nous en proposons une modélisation. Dans ce chapitre, nous présenterons la typologie élaborée et décrivant les différents types d'évaluation. Cette typologie est établie à partir de l'observation et l'étude du corpus et permet de montrer que les évaluations peuvent avoir plusieurs fonctions et contenir différentes informations : le client laisse son évaluation dans le but de donner ses opinions, de faire des suggestions, d'exprimer ses intentions ou de décrire son expérience. Cette typologie s'appuie donc sur quatre catégories (figure 3.1) : l'opinion (positive/négative/mixte), la suggestion, l'intention et la description. Afin d'étudier et de traiter automatiquement les évaluations, nous avons procédé à leur annotation manuelle selon la typologie élaborée.

3.2 Typologie

3.2.1 *Opinion positive/négative/mixte*

Affects, émotions, sentiments et opinions sont des notions utilisées de manière interchangeable au sein de différentes disciplines, telles que la psychologie, l'informatique et la linguistique (Vernier 2011). Cependant, elles ne semblent pas être dissociables les unes des autres en ce qui concerne l'évaluation (Jackiewicz 2016).



FIG. 3.1 : Typologie de l'évaluation

En psychologie, l'« affect », généralement considéré comme un état psychologique, est lié au ressenti non réfléchi et primitif (Russell 2003), dont la cible est omise. Ce ressenti est traité par un processus cognitif et se manifeste ensuite sous forme d'« émotion ». Selon Munezero *et al.* (2014), les affects sont définis comme ce qui précède les émotions, c'est-à-dire qu'ils interviennent avant la prise de conscience de leur ressenti et, par conséquent, ne se manifestent pas dans le langage[29]. L'émotion, en revanche, étant un type d'état affectif, est intense et de courte durée (Batson *et al.* 1992 ; Chatar-Moumni 2013). L'émotion peut se déclencher par des événements perceptibles : lorsqu'une personne regarde un film d'horreur par exemple, elle s'implique émotionnellement et éprouve un ressenti de peur. Le processus cognitif est responsable de ce ressenti, qui est ensuite projeté à travers des émotions comme l'angoisse, l'horreur et la panique. Ainsi, une émotion est en général très spécifique, suscitée par des événements perceptibles, ce qui signifie qu'une émotion a une cible précise (Liu 2017).

Quant au « sentiment », des définitions plus ou moins analogues ont été proposées par plusieurs chercheurs. Cattell (1940) le définit comme

[29]Toutefois, le terme affect est couramment utilisé en linguistique pour désigner l'émotion (Liu 2017).

une tendance neuropsychique acquise et relativement permanente à réagir émotionnellement, cognitivement et conativement envers un certain objet (ou une certaine situation) d'une manière stable, tout en étant conscient de cet objet. Pour Gordon (1991), les sentiments correspondent à des représentations sociales de sensations, de gestes expressifs et de significations culturelles organisées autour d'une relation à un objet social. Cette relation concerne généralement une autre personne ou un groupe, tel qu'une famille. De la même manière, Murray et Morgan (2014) définit les sentiments comme une tendance plus ou moins constante au sein d'une personnalité, qui consiste à répondre par un affect positif ou négatif à une entité donnée. À partir de ces définitions, un sentiment peut être résumé comme suit : il est créé lorsqu'un certain objet est constamment perçu par une personne, et que cette dernière se fait progressivement des idées sur celui-ci (Broad 1954). Cette idée est, en outre, associée à des tendances émotionnelles qui se déclenchent à chaque fois que la personne perçoit l'objet.

Les émotions et les sentiments renvoient tous deux à des expériences qui sont le fruit d'interactions entre des éléments biologiques, cognitifs et sociaux (Stets 2006). Cependant, les émotions se différencient des sentiments par la durée de leur expérience : elles correspondent à de courts instants où se produisent des changements dans le cerveau, le système nerveux autonome et le comportement. Les sentiments, en revanche, se construisent et se maintiennent sur une plus longue période (Ben-Ze'ev et Ben-Ze'ev 2001). À titre d'exemple, une personne peut avoir un sentiment d'amour sans se trouver dans un état d'amour. Une autre différence entre l'émotion et le sentiment tient au fait que la première n'est pas toujours exprimée vis-à-vis d'un objet, alors que la seconde est, en général, créée et vécue sur un objet. Bien que les termes d'émotions et de sentiments soient des notions différentes (Munezero *et al.* 2014), la distinction entre les deux, c'est-à-dire entre les expériences de courte durée (émotion) et de longue durée (sentiment) n'est pas évidente et n'est pas pertinente dans le cadre d'analyse réalisée sur le corpus d'avis sur les restaurants.

La notion d'« opinion » est articulée selon deux dimensions, épistémique et axiologique (Zhang 2012). Dans la première dimension, l'opinion représente une qualification relative à l'ensemble des ressources intellectuelles,

comme le raisonnement et la connaissance. Dans la dimension axiologique, l'opinion renvoie à un jugement de valeur basé sur les normes et les valeurs acceptées par la communauté. Il s'agit donc de juger « de ce qui vaut, de ce qui est bon ou mauvais, digne ou non d'être poursuivi, souhaitable ou à éviter, etc. » (Zhang 2012, p.39). Charaudeau (1992) distingue ces deux dimensions (épistémique et axiologique) en les nommant respectivement l'« OPINION » et l'« APPRÉCIATION »[30]. En ce qui concerne l'OPINION, le locuteur évalue la vérité de son propos et révèle son attitude de croyance plus ou moins certaine, relevant d'un ordre intellectif. Charaudeau (1992, p.601-602) définit l'OPINION de la manière suivante :

> « Le locuteur est *présupposé* un fait (ou une information) à propos duquel le locuteur *explicite* la place que celui-ci (ou celle-ci) occupe dans son *univers de croyance*. Le locuteur *évalue* ainsi la vérité de son propos et *révèle* du même coup ce qu'est son point de vue. Ce point de vue est d'**ordre *intellectif***, c'est-à-dire que, tout en gardant le propos à distance (point de vue externe), **le locuteur exprime une *attitude de croyance* plus ou moins certaine qui relève de la raison** (« *je pense* », « *je crois* », « *je doute* »). C'est ce qui distingue cette modalité de l'Appréciation. L'interlocuteur *n'est pas impliqué*. Il est le *témoin* de l'Opinion du locuteur. »

En revanche, l'APPRÉCIATION est relative à une évaluation structurée dans les champs de valeurs, notamment éthiques, esthétiques, hédoniques et pragmatiques. Charaudeau (1992, p.604-606) en propose la définition suivante :

> « Il est *présupposé* un fait à propos duquel le locuteur *dit* quel est son sentiment. Le locuteur évalue donc, non pas la vérité du propos, mais *sa valeur*, en révélant *ses propres sentiments*. Cette évaluation est donc d'**ordre *affectif***, c'est-à-dire que le locuteur *s'approprie* le Propos (point de vue

[30] Nous utilisons la majuscule pour distinguer les notions proposées par Charaudeau des nôtres.

> interne) en le *qualifiant* d'après **un jugement qui ne repose pas sur la raison mais sur l'*affect***. Ce jugement est nécessairement *polarisé*, c'est-à-dire qu'il varie sur **un axe dont les pôles sont : *jugement favorable (« je trouve positif que… »)/jugement défavorable (« je trouve négatif que… »)***. L'interlocuteur *n'est pas impliqué*. Il est le *témoin* de l'Appréciation du locuteur. »

L'APPRÉCIATION se charge alors d'une orientation axiologique positive ou négative. Par ailleurs, une APPRÉCIATION positive peut engendrer une satisfaction, un soulagement et une euphorie, alors qu'une APPRÉCIATION négative peut entraîner une insatisfaction, une déception, une résignation ou un désespoir (Jackiewicz 2016). Le locuteur peut ainsi évaluer la valeur d'une entité donnée en se fondant sur l'affect, autrement dit, en révélant ses propres sentiments.

Les réflexions menées dans le cadre de cette étude nous ont conduite à nous restreindre à **la dimension axiologique de l'opinion**, c'est-à-dire à l'APPRÉCIATION que propose Charaudeau (1992). Selon Zhang (2012), il n'existe pas de ligne de démarcation précise entre les deux dimensions, car un énoncé peut les impliquer toutes deux. Nous accordons néanmoins la priorité à la dimension axiologique, car, dans le corpus d'avis sur les restaurants, les clients mobilisent les valeurs en exprimant ce qui est, selon eux, bon ou mauvais, agréable ou désagréable, satisfaisant ou insatisfaisant, souhaitable ou à éviter. Les opinions reposent ainsi sur une interprétation personnelle des informations (objectives ou/et subjectives) qui peut ou non engager le sentiment (ou l'émotion) (Munezero *et al.* 2014).

Ayant précisé la définition de l'opinion que nous adoptons, nous pouvons distinguer celle-ci des autres notions mentionnées précédemment (les émotions, les sentiments)[31] – toutes trois étroitement liées à la subjectivité – par les caractéristiques suivantes :

- Les opinions peuvent être exprimées concernant une situation donnée, de manière descriptive et évaluative. Elles peuvent ou non être chargées émotionnellement (Sokolova et Lapalme 2011) ;

[31] L'affect étant défini comme une réponse primitive qui ne se manifeste pas dans le langage, nous ne le mentionnons pas ici.

– Les opinions sont fondées sur des informations objectives et/ou sur des informations relativement subjectives sur un sujet. Les sentiments, en revanche, sont suscités par des émotions (Munezero *et al.* 2014) ;

– Les opinions sont des interprétations personnelles des informations créées dans les pensées et qui ne sont pas nécessairement restreintes par les attentes sociales. En revanche, les sentiments, dérivés de l'émotion, sont généralement construits par la société (Cattell 1940) ;

– Les opinions sont principalement reliées aux entités évaluées alors que les émotions ne le sont pas. De plus, le porteur d'une opinion est souvent omis, tandis que celui d'une émotion est généralement précisé de manière explicite (Liu 2017).

La figure 3.2 illustre la relation entre les notions d'affect, d'émotion, de sentiment et d'opinion, et constitue une reprise légèrement modifiée[32] de celle présentée par Munezero *et al.* (2014).

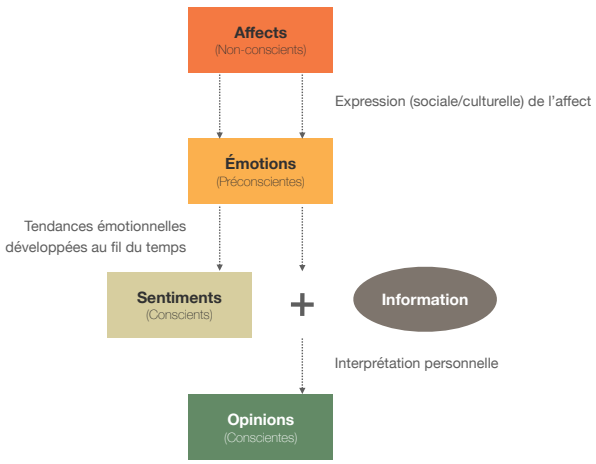


FIG. 3.2 : Distinction entre affect, émotions, sentiments et opinions

[32]Les auteurs abordent également la notion de *feeling*, alors que, dans ce travail, nous ne l'avons pas prise en compte.

Les opinions représentent donc, dans cette étude, les idées que le consommateur se fait du restaurant et la façon dont il qualifie ce dernier. La polarité des opinions est déterminée selon l'ordre émotionnel, comme le précisait Charaudeau (1992) dans sa définition de l'appréciation : si l'interprétation axiologique de la valeur accordée à l'entité donnée se réfère à une qualification favorable, il s'agit d'une opinion positive, alors que si l'interprétation relève d'une appréciation défavorable, l'opinion est négative. On parle de polarité mixte lorsque l'opinion comporte les deux polarités (positive et négative).

Une analyse du corpus a permis de relever différents indices concernant la polarité. L'une des façons de déterminer cette dernière est d'examiner l'emploi des lexiques évaluatifs. Ceux-ci conduisent en effet à une polarité positive ou négative. Les phrases suivantes constituent des exemples d'opinions positives [1]-[4] et négatives [5]-[8] :

[1] ***Bon*** *repas, plats* ***copieux****.*

[2] ***Super*** *restaurant, avec de* ***supers*** *produits.*

[3] *Accueil très* ***courtois*** *et* ***professionnel****.*

[4] *Ambiance* ***sympathique*** *et* ***détendue****.*

[5] *Les portions sont plutôt* ***maigres****.*

[6] *La cuisine* ***médiocre*** *et le temps d'attente très, très* ***long****.*

[7] *Les vins au verre sont* ***chers****.*

[8] *Le restaurant est très* ***bruyant*** *!*

Bon, *copieux*, *super(s)*, *courtois*, *professionnel*, *sympathique* et *détendue* sont des marqueurs de l'opinion positive, alors que *maigres*, *médiocre*, *long*, *cher* et *bruyant* sont plutôt des marqueurs de l'opinion négative.

Par ailleurs, les opinions incluent l'engagement émotionnel créé pendant l'expérience au restaurant, enrichissant le souvenir ; plus l'expérience est marquante, plus le souvenir restera longtemps dans la mémoire (Giboreau et Body 2012). La polarité des opinions est donc également déterminée grâce aux lexiques des émotions ou des sentiments, ces derniers étant

révélés au cours de l'expérience, comme dans les exemples suivants :

[9] *Nous avons été très* ***contente*** *de l'accueil et du service.*

[10] *Nous sommes rentrés* ***ravis*** *de notre soirée.*

[11] *Un* ***plaisir*** *pour les yeux et les papilles !*

[12] *Expérience* ***malheureuse****.*

[13] *Nous avons été* ***déçues*** *par la qualité des plats.*

[14] *Grosses* ***déceptions !***

Les exemples [9]-[11] représentent des opinions positives sur le restaurant, qui ont été exprimées à travers des émotions positives (*contente*, *ravi* et *plaisir*), alors que les extraits [12]-[14] présentent des opinions négatives, au moyen de mots exprimant des émotions négatives (*malheureuse*, *déçues* et *déceptions*).

Les clients manifestent également leur satisfaction ou leur mécontentement à travers une représentation objective de la situation concernée, jugée favorable ou défavorable selon les stéréotypes présents dans la société. À titre d'exemple, les clients font état d'une longue durée d'attente [15]-[17] ou d'une situation désagréable vécue [18], [19], qui sont considérées, par conséquent, comme une opinion négative d'un restaurant. De plus, le mécontentement peut également se manifester lorsqu'il existe un décalage entre les attentes du client et ce qu'il a vécu, comme dans les exemples [20]-[22].

[15] *Réservation confirmée à 21H 30, 45 minutes d'attente.*

[16] *Moins de 10 minutes entre le plat et le dessert !*

[17] *Nous avions réservé pour 9, il y avait 8 couverts.*

[18] *Sommelière qui présente le vin pendant que son collègue nous présentait les plats donc on n'entendait rien.*

[19] *Le restaurant nous a informé la veille qu'il n'avait pas de place pour nous, le restaurant étant complet alors qu'on nous avait confirmé la réservation.*

[20] *Leur super offre, pourtant affichée, n'était pas disponible.*

[21] *Je n'ai pas retrouvé la virtuosité attendue.*

[22] *Je ne lui attribuerai pas d'étoile comme je l'entends souvent.*

Cependant, la polarité de certains mots, tels que *courte*, *simple*, *petit*, *classique* ou *surprenant* est ambiguë. Une carte *courte* ou *simple*, par exemple, peut être rassurante ou restrictive, selon la préférence des lecteurs : le recours à peu d'ingrédients suggère que ceux-ci sont probablement de meilleure qualité, mais, dans le même temps, cette carte limite la variété des choix.

La détermination de la polarité devient plus complexe lorsque certains modifieurs – ou *valence shifters* (Polanyi et Zaenen 2006) – sont employés, tels que les intensifieurs, les atténuateurs, la négation, ou les connecteurs, qui modifient la valeur initiale d'un mot en agissant directement sur l'orientation de la polarité. Dans ce sens, Zhang (2012), dans sa thèse, se focalise sur le système de graduation et de polarité des modifieurs dans un corpus journalistique, dans le but de proposer une adaptation du modèle théorique au traitement automatique des langues.

Observons les exemples [23]-[29]. Les intensifieurs (*très*, *beaucoup* et *trop*) [23], [28], [29] et les atténuateurs (*un peu*) renforcent [23] ou affaiblissent [24] les caractéristiques évaluées. Autrement dit, les adjectifs (*belle* et *élevé*) sont complétés par l'intensifieur et l'atténuateur qui déterminent la granularité de l'opinion. Lorsque la négation (*ne…pas*, *aucun* et *peu*) est appliquée aux mots ayant une valeur positive, le mot positif est placé dans la zone négative, et inversement [25]-[27]. Si les intensifieurs, les

atténuateurs et la négation modifient la polarité au niveau lexical, les connecteurs concessifs (par exemple, *mais* et *cependant*) influencent la polarité d'une phrase ou d'un texte [28], [29] :

[23] *Une* ***très*** *belle découverte.*

[24] *Le prix des vins est* ***un peu*** *élevé.*

[25] *Le service* ***n'****est* ***pas*** *à la hauteur !*

[26] *Mais il* ***n'****y avait* ***aucun*** *plaisir dans cette soirée.*

[27] *Accueil* ***peu*** *chaleureux.*

[28] ***Pas très*** *inventif* ***mais très*** *bien exécuté.* **[négatif ⇒ positif]**

[29] *J'adore cet endroit qui est* ***très*** *convivial et sympa,* ***cependant*** *le service est* ***vraiment très****, voire* ***trop*** *long !* **[positif ⇒ négatif]**

Dans les écrits numériques, d'autres éléments peuvent influer sur la polarité et l'intensité, notamment les émoticônes [30], les ponctuations multiples en cascade [31] ou les mots en majuscules [32]. D'après Herring (1996) et Georgakopoulou (2006), ces caractéristiques, associées à la fois à l'écrit et à l'oral, sont des formes courantes dans la communication médiée par l'ordinateur (CMO ; Marcoccia 2016).

[30] *Un cadre charmant, équipe pro, plats et vins au top* ***:)***

[31] *Pas une fausse note* ***!!!***

[32] *L'accueil sympathique et un rapport qualité-prix* ***INCROYABLE****.*

L'opinion est ainsi définie, dans cette étude, selon sa dimension axiologique : il s'agit des évaluations de la valeur

d'une entité, par la révélation de sentiments. La polarité est déterminée sur un axe dont les pôles sont bons ou mauvais, agréable ou désagréable, satisfaisant ou insatisfaisant, souhaitable ou à éviter. Bien qu'elle constitue une réaction physiologique, l'émotion peut également intervenir sur cet axe de polarité, renseignant sur ce qui est bon ou mauvais d'après le locuteur. Les opinions polarisées sont fréquemment observées dans le corpus à travers les lexiques évaluatifs et émotionnels, dans lesquels les modifieurs renforcent ou affaiblissent la polarité.

3.2.2 *Suggestion*

La notion de suggestion est constatée depuis 2006, dans les travaux consacrés aux opinions dans le cadre de l'analyse de sentiments (Negi et Buitelaar 2017). Dans l'étude de Chesley *et al.* (2006), les suggestions sont identifiées au sein de textes chargés de sentiments. Elles sont marquées par l'emploi de certains verbes, utilisés comme des caractéristiques pour classifier des textes exprimant des sentiments. La détection de suggestions a ainsi émergé progressivement en TAL, et l'intérêt pour celle-ci s'est encore accru récemment, ce dont témoigne l'une des tâches proposées lors de l'atelier SemEval-2019 (Negi *et al.* 2019) : extraire des suggestions dans des forums et des évaluations sur l'hôtellerie.

Dans la littérature, la notion de suggestion apparaît sous différentes formes. Le travail fondateur de la détection automatique de suggestions est, selon nous, celui de Goldberg *et al.* (2009), qui utilise le terme « souhait » (*wish*). De leur côté, Ramanand *et al.* (2010) proposent de distinguer deux types de souhaits : le souhait d'améliorer un produit et le souhait de l'acheter. Brun et Hagège (2013) qualifient ce premier type de souhait de « suggestion d'amélioration (*suggestions for improvements*) », alors que Wicaksono et Myaeng (2012, 2013) utilisent, quant à eux, le terme de « conseil » (*advice*) dans le contexte de conseils de voyages publiés sur des forums de discussion. Negi et Buitelaar (2015) s'intéressent par ailleurs aux suggestions de client à client (*customer-to-customer* ou CTC), où un client cherche à aider les autres, ainsi qu'aux suggestions d'amélioration du produit ou du service adressées aux fournisseurs.

Dans ces lignes, Charaudeau (1992, p.589) propose une définition de la suggestion qui englobent plus ou moins tous ces aspects :

> « Le locuteur *pose*, dans son énoncé, **une action à réaliser (ou à ne pas réaliser)**. Il *sait* (ou *suppose*) que l'interlocuteur est dans une situation défavorable. **Il *propose* à l'interlocuteur d'exécuter l'action décrite comme moyen d'améliorer sa situation**. Il *fait comme* s'il était en lieu et place de l'interlocuteur (« *moi à ta place* »). **Il *se donne* un statut de savoir (plus ou moins conforté par l'expérience)**. L'interlocuteur *est supposé* se trouver dans une situation défavorable. Il *est le bénéficiaire* d'une proposition de faire pour améliorer sa situation. **Il *est libre* d'utiliser ou non cette proposition**. »

Dans le cadre de cette étude, nous pouvons dire que l'interlocuteur est dans une position défavorable car il dispose de moins d'informations sur le restaurant que le locuteur.

Ainsi, **la suggestion vise, d'une part, une amélioration relative des produits (ou services) et, d'autre part, s'adresse aux lecteurs dans le but de leur donner des conseils**. Dans cette optique, **les suggestions sont destinées à la fois aux restaurateurs (afin de leur faire prendre conscience des problèmes) et aux autres clients potentiels (futurs clients)**. Cependant, les travaux antérieurs sur la suggestion ne prennent en compte que l'un des deux destinataires.

Une diversité de marqueurs linguistiques mis en jeu a été observée dans le corpus traité :

- les verbes de parole désignant l'ordre ou le conseil, comme *recommander* et *conseiller* [33], [34] ;
- le mode conditionnel exprimant un souhait, une possibilité ou une hypothèse [35], [36] ;
- le mode impératif, qui permet aux clients de s'adresser directement aux lecteurs et de les inciter à prendre des mesures [37]-[39] ;
- les pronoms personnels ou possessifs et les adjectifs possessifs de la deuxième personne, qui permettent de construire un lien entre l'auteur et les lecteurs [38]-[42] ;

– la conjonction de subordination *si* exprimant une hypothèse, qui est souvent employée avec le pronom personnel *vous*. En limitant la portée de l'évaluation à un certain type d'individu ou à des circonstances spécifiques, la suggestion peut aider les lecteurs à déterminer dans quelle mesure l'évaluation s'applique à eux [34], [39]-[42] ;

– le terme *attention*, qui permet de signaler une information particulière [42].

[33] *Nous le* ***recommandons*** *fortement.*

[34] *Bref, une grande déception et je* ***déconseille*** *vivement* ***si*** *l'on veut s'organiser une fête de la gourmandise.*

[35] *Cadre sympa mais* ***mériterait*** *d'être un plus tamisé.*

[36] *Très bonne soirée dans un très beau cadre, un fond musical* ***serait*** *un plus.*

[37] *Surtout* ***prenez*** *cette entrée c'est succulent.*

[38] *Les gros mangeurs* ***passez votre*** *chemin.*

[39] ***N'hésitez*** *pas* ***si vous aimez*** *ce type de restaurant.*

[40] ***Si vous voulez vous régaler*** *avec un super plateau de fruits de mer c'est le bon endroit.*

[41] ***Si*** *les plateaux déjà composés sur la carte ne conviennent pas à* ***votre*** *goût, on peut choisir à la carte en respectant un minimum de prix correspondant au plateau le moins cher proposé.*

[42] ***Attention*** *à ne pas y aller* ***si vous avez*** *très faim.*

Tous ces éléments incitent les lecteurs à effectuer une action (commander ou non un plat particulier, visiter au restaurant, etc.) ou à améliorer l'un

des aspects du restaurant (accueil, ambiance, propreté, etc.).

Ceci confirme que les suggestions sont ainsi destinées aussi bien aux clients potentiels qu'aux restaurateurs. Elles apportent des détails qui peuvent être utiles aux clients potentiels dans leur prise de décision. Les restaurateurs, quant à eux, peuvent penser avoir offert une expérience positive aux clients alors que ces derniers n'ont pas nécessairement le même ressenti. Les suggestions peuvent donc aider les restaurateurs non seulement à percevoir l'expérience telle qu'elle a été vécue par les clients, mais aussi à résoudre les problèmes identifiés. Les restaurateurs pourront en outre proposer des alternatives, qui deviendront éventuellement des sources de satisfaction et de plaisir. Par conséquent, la suggestion permet de rapprocher les perspectives des deux acteurs (restaurateurs et clients), entre ce que le restaurant souhaitait et a livré et ce que le client attendait et a vécu.

3.2.3 *Intention*

La notion d'intention a été peu étudiée dans le cadre de la fouille d'opinions, alors qu'elle a été considérablement abordée dans le domaine de l'intelligence artificielle. Le modèle logiciel croyance-désir-intention (*Belief-Desire-Intention* ou BDI) modélise le comportement d'un agent intelligent, inspiré de la théorie de Bratman (1987). Cette théorie explique le raisonnement d'un agent via trois catégories d'état : la croyance, qui représente les informations que l'agent possède sur le monde ; le désir, qui indique l'état du monde que l'agent aimerait voir réalisé ; et l'intention, qui désigne l'action que l'agent entreprend dans le but de satisfaire son désir.

Dans la lignée du modèle BDI, Benamara *et al.* (2017a) utilisent le terme « intention » (*intent*) comme un concept plus large couvrant les désirs, les préférences et les intentions (*intention*), qui sont des attitudes mentales contribuant au comportement rationnel d'un agent. Selon les auteurs, les désirs peuvent être ordonnés en fonction de la préférence, qui marque ce que l'agent considère plus favorablement. Un ordre prioritaire des désirs est donc déterminé. La préférence d'un agent, quant à elle, est étroitement liée à l'action, qui permet de prédire la réaction de l'agent, si celui-ci est rationnel. Par exemple, une personne peut se rendre dans un restaurant pour goûter un plat, sans tenir compte de son cadre. En revanche, si cette

personne privilégie l'environnement, le décor ou l'ambiance, elle préfèrera manger ailleurs, dans un restaurant au cadre élégant. En ce qui concerne l'intention, un agent considère différents désirs, qui peuvent être en conflit et incohérents, avant de décider d'accomplir une action (Bratman 1990). Parmi ces désirs, seuls ceux retenus par l'agent de manière rationnelle et cohérente constituent une intention. En réunissant ces trois notions, pour Benamara *et al.* (2017a), les « intentions » correspondent à des attitudes mentales qui contribuent au comportement rationnel d'un agent et qui peuvent donc inciter ce dernier à agir.

Liu (2015) propose deux définitions de l'intention : i) une action qu'une personne (ou un groupe de personnes) a l'intention de suivre ; ii) le but sous-entendu d'une action (ou d'un ensemble d'actions spécifiques). Il s'intéresse au premier emploi de l'intention, car il considère que le second est implicite et donc difficile à identifier. Dans la lignée de ce travail, Chen *et al.* (2013) ont effectué une classification des intentions explicites. Carlos et Yalamanchi (2012) proposent une catégorisation des différentes intentions (acheter, se renseigner, se plaindre, critiquer, complimenter, informer, quitter, comparer, souhaiter et vendre) pour une application commerciale. De la même manière, Ding *et al.* (2015) identifient les intentions de l'utilisateur dans le cadre de la consommation.

Dans notre travail, **l'intention désigne le souhait d'un client de retourner ou non dans un restaurant**. Cette intention peut être considérée comme positive ou négative. L'intention positive implique souvent le souhait de revenir au restaurant, ce qui correspond à une évaluation positive ; en revanche, l'intention négative manifeste le désir de ne pas répéter l'expérience et constitue donc une évaluation négative. L'intention représente ainsi un engagement volontaire du client, initié par lui-même.

L'intention est marquée explicitement, dans les évaluations observées, à travers les verbes au futur ou avec le préfixe *re-*, qui indique généralement la réitération d'une action (*revenir*, *retourner*, *refaire* et *renouveler*). Notons également que dans le corpus, l'intention s'exprime en majorité dans la dernière phrase. Elle résume ainsi l'ensemble de l'expérience gastronomique, en présentant une impression ou une évaluation globale du restaurant, comme dans les exemples suivants :

[43] [...] *Un repas plein de surprises et d'une qualité incomparable ! Nous* ***reviendrons****.*

[44] [...] *Au final addition assez salée. Nous* ***n'y retournerons pas****.*

[45] [...] *Très agréable soirée, repas excellent, bien qu'un peu cher. Mais nous y* ***retournerons****...*

La détection des intentions peut ainsi compléter l'analyse de sentiments et d'opinions (Benamara *et al.* 2017b ; Liu 2015) car elle permet aux restaurants de se pencher sur les actions futures qu'un client souhaite réaliser ou, plus précisément, de prévoir si un client a l'intention de cesser de fréquenter l'établissement après une expérience négative.

3.2.4 *Description*

Si les trois catégories précédentes (l'opinion, la suggestion et l'intention) représentent l'évaluation exprimée par un client, **la description renvoie aux informations factuelles associées à l'expérience vécue, ayant peu de lien avec l'évaluation**. Ce type d'information est peu reconnu dans la fouille d'opinions et selon Ganu *et al.* (2009) par exemple, ces informations sont même considérées comme étant peu utiles. Cependant, la description permet aux lecteurs de découvrir l'arrière-plan de l'expérience : la raison pour laquelle les clients se rendent dans ce restaurant, les personnes qui les accompagnent, les plats qu'ils ont choisis, etc. Notons que, souvent, les clients commencent leur évaluation par une description permettant ainsi de plonger les lecteurs dans leur expérience :

[46] ***Nous étions six pour un dîner anniversaire en famille****. Tout a été parfait et pourtant nous sommes difficiles.* [...]

[47] ***Couple strasbourgeois de passage à Paris, nous avons opté pour ce restaurant après une soirée théâtre, nous nous y sommes rendus à 22 h un mardi****. Accueillis par le chef, il nous a détaillé les 3 plats au menu avec toute sa sympathie.* [...]

[48] ***Nous avons découvert ce restaurant par hasard en cherchant.*** *Nous n'avons pas été déçus.* [...]

[49] ***Nous avions réservé lundi soir, une table de 4.*** *On s'est régalés !!* [...]

[50] ***Nous avons réservé pour 3 personnes.*** *Nous avons été agréablement surprises par le cadre du restaurant.* [...]

[51] *Ainsi, le kir royal est au gingembre frais, le canard à la pastèque caramélisée, le saumon laqué à la réglisse, le fondant au chocolat à la liqueur Amarula.*

[52] *Nous avons optez pour le mille-feuille à la vanille et une meringue à la figue.*

[53] *J'ai pris un croustillant de saumon avec frites et mon ami une cassolette de veau avec du riz.*

[54] *Nous avons dégusté en entrée une raviole au bœuf pour les uns, tartare de saumon pour les autres.*

La description est liée, d'une part, aux informations factuelles ayant un impact avant la visite du restaurant, telles que le motif de cette dernière [46]-[48] ou la précision sur une réservation effectuée [49], [50]. Dans ce cas, la description a tendance à être positionnée au début d'un texte d'évaluation. D'autre part, elle est relative à des informations factuelles concernant l'expérience vécue lors de la visite, comme le choix du menu ou la composition du plat [51]-[54] et la référence au(x) compagnon(s) [53], [54]. La description fait donc intervenir les tendances, les modes de vie et le quotidien des clients et exerce des influences sur la perception, l'attitude et le comportement des lecteurs. Malgré sa faible fréquence (voir figure 3.4), ce type d'information est étroitement lié au contexte de l'expérience vécue, permettant de mieux cerner chacun des clients.

3.3 Annotation manuelle

L'annotation manuelle du corpus est indispensable puisqu'elle permet d'enrichir les possibilités d'exploration et de constituer un corpus de référence qui servira pour l'apprentissage automatique (Poudat et Landragin 2017). Le corpus annoté est ainsi une seconde version du corpus dans lequel l'interprétation et la perception peuvent différer selon les connaissances du domaine, l'orientation théorique et les variables sociologiques des annotateurs (Eshkol-Taravella 2015).

Afin d'étudier et de traiter automatiquement les évaluations, nous avons procédé à leur annotation manuelle selon la typologie élaborée et présentée dans la section précédente. Dans cet objectif, le corpus a d'abord été segmenté en phrases en fonction des signes typographiques (.!?), puis annoté selon la typologie présentée précédemment. La tâche d'annotation a consisté à attribuer à chaque phrase une étiquette parmi des catégories prédéfinies : POS_OPINION (opinion positive), NEG_OPINION (opinion négative), MIX_OPINION (opinion mixte), SUGGESTION, INTENTION et DESCRIPTION. Le nombre d'avis à annoter s'élevait à 6 287 avis, pour un total de 17 268 phrases.

Parmi les différents outils disponibles pour l'exploration de corpus et son annotation manuelle (GATE[33], BRAT[34], WebAnno[35], etc.), nous avons sélectionné le logiciel MAXQDA[36]. Ce logiciel est conçu pour l'analyse de données qualitatives telles que les textes, les enregistrements audio et vidéo, les sondages ou les photographies. Un texte donné est codé à l'aide de différentes couleurs ; en se fondant sur ces segments codés, l'extraction et la compilation sont ensuite réalisées afin d'analyser l'ensemble des données. Le logiciel peut également créer une liste de fréquence et un nuage de mots, qui permettent d'explorer les données textuelles dans une perspective globale.

Chaque phrase a donc été codée selon la catégorie concernée à l'aide de ce logiciel (figure 3.3). En cas d'ambiguïté, lorsque plusieurs catégories étaient pertinentes, les catégories les moins représentées dans le corpus,

[33] http://gate.ac.uk/ [consulté le 13 septembre 2020].
[34] http://brat.nlplab.org/index.html [consulté le 13 septembre 2020].
[35] https://webanno.github.io/webanno/ [consulté le 13 septembre 2020].
[36] https://www.maxqda.com/ [consulté le 13 septembre].

comme l'intention ou la suggestion, ont été privilégiées. La phrase [55], qui peut être assignée à deux catégories (opinion positive et intention), a ainsi été annotée comme intention. De plus, si une phrase impliquait à la fois l'intention et la suggestion, nous la classions dans la première catégorie. Nous justifions ce choix par des études en psycholinguistique (Carreiras *et al.* 1995 ; Gernsbacher 1990 ; Kim *et al.* 2004) qui affirment que le contenu présenté au début de la phrase semble avoir un effet plus marquant dans la détermination du jugement final. La phrase [56] par exemple, est, annotée ainsi comme suggestion.

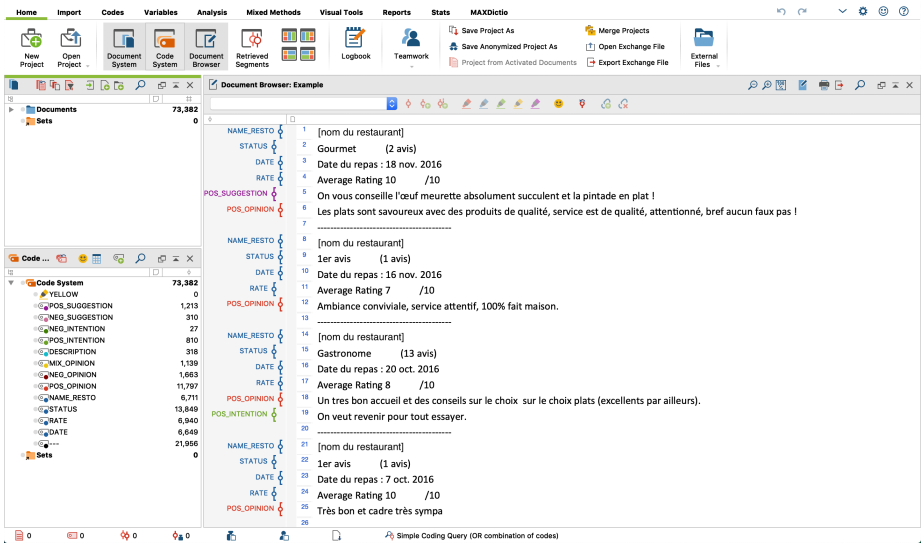


FIG. 3.3 : Capture d'écran du logiciel MAXQDA

[55] *Un bon moment, à refaire vite !*

[56] *Adresse à recommander, nous y retournerons.*

Avant de procéder à l'annotation manuelle finale, nous avons validé la typologie proposée en calculant l'accord inter-annotateur (AIA), afin de garantir un bon niveau d'annotation. Pour cela, un test d'annotation a été réalisé sur un extrait de corpus (100 phrases choisies de manière aléatoire, mais en respectant la distribution de chaque catégorie). L'an-

notation manuelle a été effectuée par trois personnes (des doctorants en linguistique[37]). Afin de faciliter la tâche aux annotateurs qui ne disposent pas du logiciel MAXQDA, l'annotation a été effectuée dans un fichier Excel, dans lequel chacun a attribué une catégorie à chaque phrase donnée, comme l'illustre le tableau 3.1.

TAB. 3.1 : Exemple d'annotation manuelle (RestoFR)

Phrases	**Catégories**
Nous nous sommes tous régalés.	**POS_OPINION**
Les desserts m'ont déçu.	**NEG_OPINION**
L'addition est corsée mais tout est top.	**MIX_OPINION**
À essayer pour toutes occasions !	**SUGGESTION**
Un léger fond musical serait un plus.	**SUGGESTION**
On reviendra !	**INTENTION**
Nous n'y retournerons pas.	**INTENTION**
Nous y étions pour les 30 ans de ma conjointe.	**DESCRIPTION**

L'annotation manuelle étant un processus d'interprétation individuelle, sa validité est évaluée par le biais du calcul de l'AIA. Celui-ci s'applique aux résultats d'accord et de désaccord entre les annotateurs. L'AIA entre les trois annotateurs a été calculé à l'aide de la mesure kappa de Fleiss (Fleiss *et al.* 1981), qui convient pour comparer l'annotation entre trois annotateurs ou plus. L'accord porte sur l'attribution d'une même étiquette par chaque annotateur. Pour interpréter la valeur de kappa (κ), Landis et Koch (1977) ont proposé une échelle des valeurs de kappa donnant un bon point de départ pour l'évaluation. Le score obtenu par nos annotateurs (0,90) est considéré comme « presque parfait » selon leur échelle et démontre que nos catégories et nos règles d'annotations sont bien définies et fiables. La typologie élaborée a donc été jugée pertinente, et le reste du corpus (17 168 phrases) a été annoté par un annotateur, aboutissant au traitement d'un total de 17 268 phrases.

La répartition des catégories est présentée par la figure 3.4. Les opinions

[37]Hélène Flamein, Jennifer Ganaye et l'auteure.

positives constituent la majorité des catégories (68,05 %), alors que les descriptions ou les intentions n'en représentent que 1,85 % et 4,86 % respectivement. Ainsi, la distribution des catégories est fortement déséquilibrée, ce qui a été pris en compte lors de la classification automatique (voir chapitre 5).

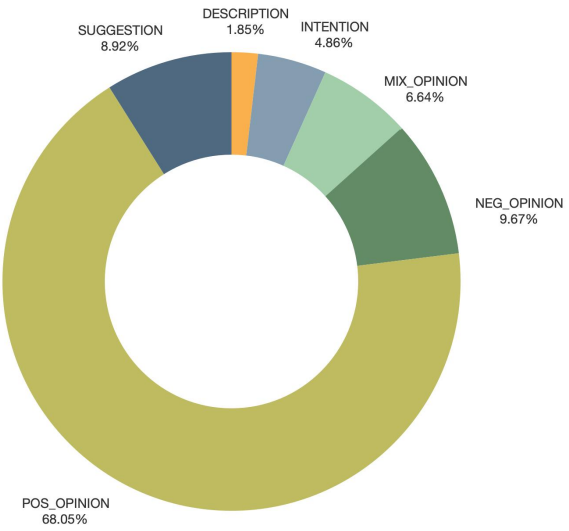


FIG. 3.4 : Répartition des catégories d'évaluation

3.4 Conclusion

Dans ce chapitre, nous avons mis en évidence le fait que les évaluations en ligne peuvent avoir plusieurs fonctions et contenir différentes informations (figure 3.5) : le client laisse un avis pour donner son opinion (positive/négative/mixte), faire des suggestions d'amélioration (suggestion), exprimer ses intentions (intention) ou pour décrire son expérience (description). Toutes ces informations reflètent les différents objectifs que poursuit l'internaute en laissant son évaluation. L'hétérogénéité de ces objectifs et fonctions s'explique aussi par la nature double de destinataires visés par les évaluations en ligne : une entité évaluée, un lieu visité par exemple (les restaurateurs dans notre corpus), d'une part, et les clients potentiels, d'autre part.

Grâce à l'observation manuelle du corpus, nous avons également relevé des indices linguistiques propres à chaque catégorie. Ces indices servent à définir et décrire chaque catégorie ainsi que la détecter automatiquement.

La typologie élaborée permet de replacer l'expérience individuelle d'un internaute dans un contexte plus large : avant, pendant et après la visite du restaurant. Malgré la diversité des origines et des expériences de chacun des clients, la typologie permet de développer l'analyse des comportements des clients et ainsi d'anticiper leurs besoins en général.

Le processus d'annotation manuelle a été réalisé selon les bonnes pratiques et l'accord entre les trois annotateurs, calculé à l'aide de la mesure kappa de Fleiss (Fleiss *et al.* 1981) obtient (0,90), ce qui est considéré comme l'accord « presque parfait » selon l'échelle de Landis et Koch (1977). Le corpus annoté a permis d'effectuer différentes analyses quantitatives à l'aide de méthodes et outils statistiques présentées dans le chapitre 4. Ce corpus a servi également de corpus de référence pour le traitement automatique réalisé, qui sera abordé dans le chapitre 5.



FIG. 3.5 : Synthèse de la typologie proposée

Chapitre 4

Analyse quantitative de l'évaluation

4.1 Introduction

Une typologie d'évaluation a été présentée dans le chapitre précédent. Elle contient six catégories : POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION et DESCRIPTION. Dans ce chapitre consacré à l'analyse quantitative du corpus annoté, nous étudierons la manière dont les clients expriment leurs évaluations en utilisant une méthode de linguistique corpus. Dans un premier temps, nous décrirons le prétraitement du corpus effectué pour procéder à l'analyse quantitative. Le corpus sera d'abord nettoyé et prétraité dans le but de faciliter son exploration de manière automatique en supprimant ainsi des anomalies ou des valeurs incorrectes qui entraînent un traitement des données erronées (Flamein 2018). Cette étape que l'on peut considérer comme essentielle permettra de garantir de bons résultats. À partir de ce corpus nettoyé, nous mettrons en place une chaîne d'analyse quantitative afin de mieux dégager les spécificités observées dans l'évaluation exprimée par les clients en ligne. Cette chaîne comprend plusieurs étapes : l'analyse lexicométrique, la distribution des parties du discours et la position des catégories d'évaluation au sein d'un avis.

L'analyse lexicométrique sera exposée, d'abord, pour observer la fréquence des lemmes. Nous procéderons aussi à une analyse des mots-clés et des Term Frequency-Inverse Document Frequency (TF-IDF). Nous nous intéresserons, ensuite, à la distribution des parties de discours afin d'étudier les aspects morphologiques et syntaxiques des catégories d'évaluation

annotées. Enfin, nous procéderons à l'examen de la position des catégories au sein d'un avis, plus précisément l'ouverture et la clôture d'un avis. Toutes ces analyses permettront ainsi de caractériser le dispositif linguistique de chaque catégorie sous un angle nouveau. Pour ce faire, nous utiliserons des outils d'analyse et de visualisation des données, comme le langage Python, le langage R, le logiciel AntConc[38] (Anthony 2019) et le logiciel Tableau[39].

4.2 Nettoyage et prétraitement

En vue de réaliser une analyse quantitative et automatique des données, un prétraitement s'avère nécessaire. Son objectif est d'homogénéiser les données afin de réduire la variation des mots à indexer lors du traitement automatique. Ce processus est particulièrement crucial pour une exploration des données provenant de communications médiées par l'ordinateur (Poudat *et al.* 2020), car ces dernières se distinguent de celles de l'écrit standard par un ensemble de phénomènes hétérogènes : l'utilisation de majuscules pour donner de l'emphase (*EXCELLENT*), d'émoticônes (« :) », « :(») et d'abréviations (*resto* au lieu de *restaurant*), les fautes d'orthographe ou de frappe (*acceuil* au lieu de *accueil*), etc.

Le premier prétraitement réalisé est le nettoyage du corpus qui comprend les démarches suivantes :

- Passage des mots en minuscules ;
- Remplacement des chiffres par « NUM », les symboles € par *euros* et % par *pourcent* ;
- Remplacement des émoticônes par « emoPOS » ou « emoNEG » selon la polarité[40] ;
- Normalisation des abréviations, ce qui transforme *resto* en *restaurant* et *déco* en *décoration* ;

[38] https://www.laurenceanthony.net/software/antconc/ [consulté le 26 septembre 2020].

[39] https://www.tableau.com/fr-fr [consulté le 26 septembre 2020].

[40] Les émoticônes se distinguent habituellement par la polarité : ceux qui ont une polarité positive comme « :) », « :-) » et ceux qui sont plutôt de polarité négative « :(». Une liste comprenant les émoticônes positifs et négatifs a été créée préalablement.

– Correction orthographique des mots connus de manière erronée dont la fréquence est supérieure à deux dans le corpus[41].

Ensuite, le corpus est lemmatisé et étiqueté. Il s'agit d'assigner à chaque mot du corpus une étiquette comportant des informations de nature morphologique (la forme de base, le genre, le nombre, etc.). Ces prétraitements essentiels offrent une vue complémentaire pour l'analyse des textes bruts (Lebart *et al.* 2019).

Différentes bibliothèques sont conçues sur Python, comme spaCy[42], Treetaggerwrapper[43] ou Stanford CoreNLP[44] pour réaliser ces traitements. Le tableau 4.1 illustre le résultat de l'étiquetage proposé par les trois outils.

TAB. 4.1 : Comparaison du résultats selon les étiqueteurs (spaCy, StanfordNLP, Treetaggerwrapper)

Lemme@Étiquetages morphosyntaxiques	
Phrase originale	**La cuisine est composée de ingrédients frais avec des plats.**
spaCy	le@DET cuisine@NOUN être@AUX composer@VERB de@ADP ingrédient@NOUN **frais@NOUN** avec@ADP un@DET plat@NOUN très@ADV souvent@ADV renouveler@ADJ .@SENT.@SENT
StanfordNLP	le@DET cuisine@NOUN **être@AUX** composer@VERB de@ADP ingrédient@NOUN frais@ADJ avec@ADP un@DET plat@NOUN très@ADV souvent@ADV renouveler@VERB .@PUNCT.@SENT
Treetaggerwrapper	le@DET:ART cuisine@NOM **être@VER:pres** composer@VER:pper de@PRP ingrédient@NOM frais@ADJ avec@PRP du@PRP:det plat@NOM très@ADV souvent@ADV renouveler@VER:pper .@SENT

[41] Considérés inconnus par le lemmatiseur.
[42] spaCy, https://spacy.io/models/fr/ [consulté le 28 juillet 2020].
[43] Treetaggerwrapper, https://treetaggerwrapper.readthedocs.io/en/latest/ [consulté le 28 juillet 2020].
[44] Stanford CoreNLP, https://stanfordnlp.github.io/CoreNLP/download.html [consulté le 28 juillet 2020].

Bien que spaCy soit connu pour donner de bonnes performances en général[45], dans le cadre de cette étude, l'étiqueteur a tendance à être moins performant, ce qui est le cas par exemple pour l'adjectif *frais* (frais@NOUN), alors que StanfordNLP et Treetaggerwrapper rendent des résultats comparables. La différence porte sur une catégorie morpho-syntaxique AUX (auxiliaire) absente chez Treetaggerwrapper (être@AUX). Pour cette raison, StanfordNLP est choisi pour ce travail. L'identification des auxiliaires permet de les éliminer dans les traitements et de ne se concentrer que sur les verbes plus significatifs. Treetaggerwrapper présente, néanmoins, l'avantage de pouvoir préciser le temps et le mode des verbes (êtreVER:pres). Cet étiqueteur sera ainsi utilisé partiellement plus tard dans le chapitre sur la détection automatique afin d'extraire quelques caractéristiques à introduire dans l'algorithme.

D'une manière générale, les catégories morphosyntaxiques peuvent être classées dans deux classes : une classe ouverte, en évolution constante (verbes, noms, adjectifs et adverbes), et une classe fermée, plus stabilisée (déterminants, prépositions, pronoms et conjonctions). Cette dernière constitue en général les mots-outils (Lebart *et al.* 2019).

Il est courant que les mots-outils soient considérés comme peu intéressants, et donc éliminés systématiquement lors du prétraitement du corpus. Or, la liste des mots-outils doit être constituée avec prudence selon la tâche et la nature du corpus. Comme nous l'avons mentionné précédemment, les embrayeurs de personnes – les pronoms personnels, les pronoms possessifs et les adjectifs possessifs de la première personne (*je*, *nous*, *on*, *moi*, *mon*, *notre*, *mien*, etc.) et de la deuxième personne (*vous*, *votre*, etc.) – renvoient au locuteur ou au(x) destinataire(s), relevant la subjectivité. La conjonction *mais*, quant à elle, joue un rôle important dans la détermination de l'opinion mixte[46], opposant les opinions positives et négatives au sein d'une phrase. De plus, les adverbes qui expriment l'intensité comme *très* permettent de renforcer l'expression de l'opinion. Les chiffres, pour leur part, sont souvent employés dans la description

[45] Les performances sont indiquées sur le site spaCy. Ce dernier reconnaît, cependant, que leurs étiqueteurs peuvent produire des résultats incohérents sur de nombreux genres, tels que les textes des médias sociaux, parce qu'ils sont entraînés sur Wikipédia.

[46] Une analyse de notre corpus montre que 57,5 % des opinions mixtes contiennent la conjonction *mais*.

pour indiquer les informations factuelles (le prix, le nombre de personnes, etc.). Tous ces marqueurs ont une valeur informative dans cette étude, alors qu'ils sont considérés comme des mots-outils à supprimer, selon les listes des mots-outils français proposées, par exemple, par NLTK[47], Snowball[48] ou l'équipe CLEF[49]. Dans le cadre de l'analyse quantitative, nous retenons donc les lemmes des classes ouverts (les noms, les verbes et les adjectifs), les embrayeurs de personnes, la conjonction *mais* et les chiffres (*NUM*).

4.3 Analyse lexicométrique

La fréquence des lemmes (ou des mots) fait depuis longtemps partie de la méthodologie standard d'exploitation des corpus (Baron *et al.* 2009). La fréquence de lemmes peut être visualisée de manières différentes : les nuages de mots, les graphiques, etc. Les nuages de mots permettent d'obtenir un aperçu de l'ensemble du corpus, comme le suggère la figure 4.1 dans laquelle les lemmes fréquents sont illustrés à l'aide du langage R. La taille de la police représente la fréquence des mots : plus ils sont fréquents, plus leur taille est grande et plus ils se situent au centre de la figure, alors que les mots moins fréquents se trouvent à la périphérie. La figure 4.1 illustre la fréquence élevée de l'adverbe *très*. Nous observons aussi que les nuages de mots représentent assez bien le thème global du corpus, c'est-à-dire les cibles d'évaluation (*service*, *plat*, *qualité*, *accueil*, etc.) et les adjectifs évaluatifs (*bon*, *excellent*, *agréable*, etc.) qui décrivent les restaurants (*restaurant*).

[47] https://www.nltk.org/ [consulté le 28 juillet 2020].
[48] http://snowball.tartarus.org/algorithms/french/stop.txt [consulté le 28 juillet 2020].
[49] http://members.unine.ch/jacques.savoy/clef/frenchST.txt [consulté le 28 juillet 2020].



Fig. 4.1 : Nuage de mots de l'ensemble des données

D'autres visualisations et analyses de la fréquence ont été réalisées dans le cadre de ce travail. Nous présenterons, d'abord, une fréquence de lemme sous forme d'une liste, ce qui permettra de mettre en évidence les lemmes les plus récurrents dans chaque catégorie. Cependant, afin de mieux caractériser chaque catégorie d'évaluation en tenant compte de la fréquence relative des catégories, nous effectuerons également une analyse de mots-clés (*keyword analysis*) (Baker 2010), obtenus par le biais du degré de saillance. Une autre méthode utile pour étudier le lexique est d'appliquer la méthode de pondération TF-IDF aux n-grammes (des séquences de mots consécutifs), elle sera présentée par la suite. La méthode TF-IDF offre la possibilité de dépasser la limite d'une simple fréquence, tout comme l'analyse des mots-clés.

4.3.1 *Fréquence des lemmes*

Une liste de fréquence de mots est, en général, un tableau à deux colonnes : dans une colonne, figurent tous les mots apparaissant dans le corpus, et dans l'autre, leur fréquence (Gries 2009). Cette méthode peut donc apporter des informations intéressantes concernant les mots qui apparaissent (ou n'apparaissent pas) dans un texte. Le corpus d'étude contient 183 137 lemmes dont 6 408 lemmes distincts. Pour chaque catégorie d'évaluation, la liste de fréquence est illustrée dans la figure 4.2 : il s'agit des dix lemmes les plus fréquents dans l'ordre décroissant.

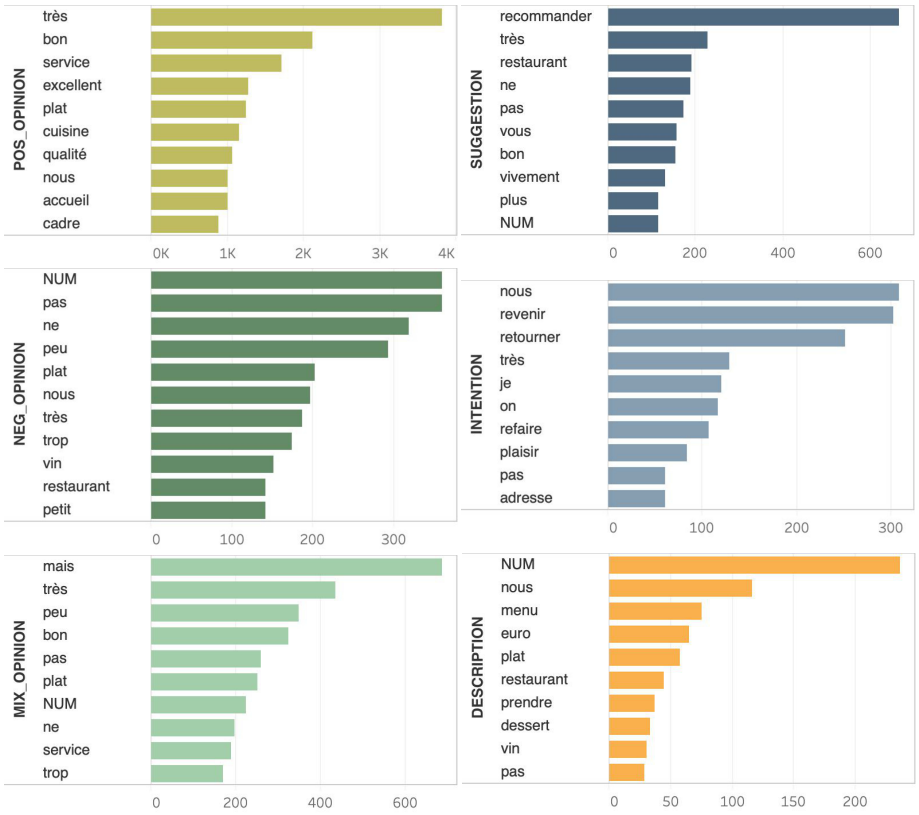


FIG. 4.2 : Fréquences des lemmes

De nombreux noms (*service*, *plat*, *cuisine*, *qualité*, *accueil*, *cadre*) constituent la liste de l'opinion positive ; ils sont utilisés pour décrire un restaurant et ses différents aspects, autrement dit ils représentent les

cibles sur lesquelles porte l'évaluation. La négation (*pas*, *ne*, *peu*) et l'adverbe *trop* sont présents dans l'opinion négative, et la conjonction *mais* dans l'opinion mixte. En revanche, les verbes sont observés dans les phrases exprimant la suggestion (*recommander*), l'intention (*revenir*, *retourner*, *refaire*) et la description (*prendre*), car toutes ces catégories sont étroitement liées à l'action attendue ou réalisée. Pour étudier le lien entre les catégories morphosyntaxiques et les catégories d'évaluation, nous avons réalisé l'analyse factorielle des correspondances qui sera présentée dans la section 4.4.

Cependant, la fréquence des lemmes selon les catégories ne suffit pas à dégager clairement les particularités de chaque catégorie, comme l'affirme Müller (1979, p.243) :

> « Dire qu'un fait est plus fréquent dans le texte A que dans le texte B ne saurait s'appliquer à la fréquence absolue, mais à une fréquence relative qui ne peut être établie que si l'on connaît avec précision suffisante le rapport de longueur des deux textes ».

En outre, une fréquence trop élevée d'un lemme (c'est-à-dire le lemme présent dans presque toutes les catégories) permet d'identifier les spécificités de la catégorie dans laquelle il apparaît, mais ne permet pas de différencier les catégories. C'est le cas de certains lemmes comme *service*, *plat* et *très*. À cet égard, une analyse des mots-clés (*keyword*), qui sera détaillée par la suite, permet de mettre en évidence la saillance lexicale, tout en tenant compte de la fréquence relative entre les catégories.

4.3.2 *Analyse des mots-clés*

L'analyse des mots-clés est un moyen pratique pour mettre en évidence l'importance lexicale (Baker 2010). Elle consiste à comparer la fréquence relative des mots dans un corpus d'étude par rapport à celles observées dans un corpus de référence. Il ne s'agit pas d'identifier le sujet (« *aboutness* ») du corpus d'étude (Scott 1998), mais de révéler les caractéristiques saillantes et fonctionnelles liées au corpus. Pour cette raison, l'analyse des mots-clés est souvent employée dans de nombreuses études

linguistiques basées sur les corpus (Aull 2017 ; Gabrielatos 2018 ; McEnery 2016 ; Poole 2016 ; Rayson 2008 ; Scott et Tribble 2006 ; Seale *et al.* 2006).

Les mots-clés ne sont pas non plus les mots les plus fréquents dans l'ensemble du corpus, mais plutôt ceux qui sont particulièrement fréquents dans le corpus d'études. Cette approche est donc fondée sur une comparaison de deux listes de mots, l'une issu du corpus d'études et l'autre d'un corpus de référence. La différence entre le corpus d'études et celui de référence est mesurée grâce au « degré de saillance (*keyness*) », qui est calculé au moyen de mesures statistiques, tels que la log-vraisemblance (Dunning 1993), %DIFF (Gabrielatos et Marchi 2012), le ratio logarithmique (Hardie 2014), le *d* de Cohen (Brezina 2014) et SMP (*Simple Maths Parameter* ; Kilgarriff 2009). Au cours de cette analyse, afin d'extraire les mots-clés d'une catégorie (une suggestion, par exemple), le sous-corpus qui correspond à cette catégorie est pris comme corpus d'étude, et le reste du corpus (dans ce cas-là, composé de l'opinion positive, de l'opinion négative, de l'opinion mixte, de l'intention et de la description) comme corpus de référence.

Les outils de linguistique de corpus, comme WordSmith (Scott 2020), TermoStat (Drouin 2003) ou Lancsbox (Brezina *et al.* 2018) peuvent générer une liste de mots-clés d'un corpus. Dans le cadre de notre étude, nous avons utilisés le logiciel AntConc (Anthony 2019) avec les paramètres par défaut. Nous avons suivi les recommandations données dans la documentation de l'outil[50]. La log-vraisemblance a été donc appliquée pour le calcul du degré de saillance. Les résultats de l'extraction des mots-clés sont montrés dans la figure 4.3. Il s'agit des dix premiers lemmes les plus saillants pour chaque catégorie. Notons que certains lemmes fréquents (voir figure 4.2) sont également identifiés comme des mots-clés.

Les opinions positives se caractérisent par les lexiques évaluatifs (*bon*, *parfait*, *agréable*, *beau*) et les cibles sur lesquelles elles portent (*accueil*, *service*, *cuisine*, *cadre*). Les opinions négatives sont associées aux lexiques qui impliquent une polarité négative (*bémol*, *dommage*, *cher*, *déception*) et la négation (*peu*, *pas*, *ne*). Les opinions mixtes sont liées aux lexiques concernant l'opinion positive et négative, mais notons que le terme le plus

[50] https://www.laurenceanthony.net/software/antconc/releases/AntConc343/help.pdf [consulté le 28 juillet 2020].

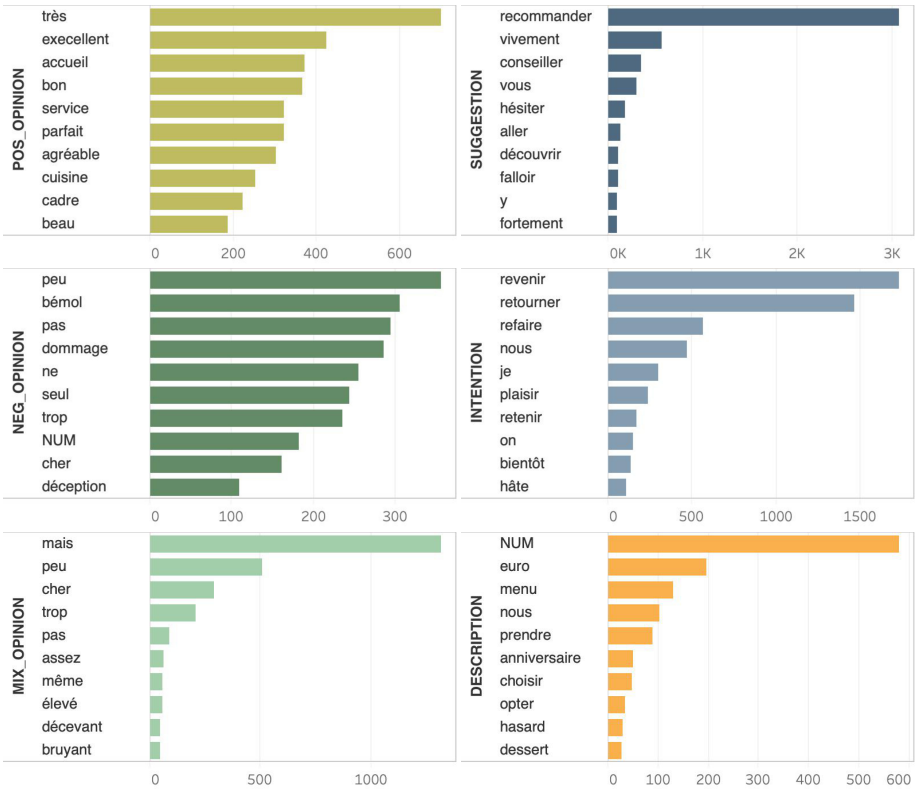


FIG. 4.3 : Mots-clés de chaque catégorie

caractéristique de cette catégorie est la conjonction mais. Les suggestions contiennent les verbes d'ordre et de conseil tels que *recommander* et *conseiller*. Elles se manifestent également à travers le pronom *vous* qui construit un lien entre l'auteur et les lecteurs. Les intentions comprennent les verbes itératifs (*revenir*, *retourner*, *refaire*) et les pronoms (*nous*, *je*, *on*). Les descriptions incluent les mots concernant des menus choisis (*menu*, *dessert*, *opter*), leur prix (*NUM*, *euro*) et le motif de la visite (*anniversaire*, *hasard*).

Nous remarquons également une présence majoritaire de noms et d'adjectifs dans les opinions et de verbes dans les suggestions, intentions et descriptions, ce qui sera abordé plus en détail lors d'une observation de l'analyse factorielle des correspondances (section 4.4). Une catégorie morphosyntaxique semble donc guider l'apparition de chacune des catégories d'évaluation.

4.3.3 *TF-IDF des n-grammes*

Si dans la partie précédente, nous avons examiné les lemmes pris isolément sans leur contexte, dans cette partie, nous nous concentrons sur les suites de lemmes (les n-grammes) selon la catégorie d'évaluation.

Un n-gramme est une séquence de n mots consécutifs. Les tri-grammes de [1] sont : *personnel et ambiance*, *et ambiance très*, *ambiance très agréable*. Contrairement à une simple liste de fréquence des mots, les n-grammes prennent en compte les particularités syntaxiques et sémantiques. À titre d'exemple, [2] et [3] disposent d'une même liste de mots : *assez*, *bons*, *copieux*, *les*, *mais*, *pas*, *plats* et *sont*. Or, les significations générées par certains tri-grammes ne seront pas identiques comme le montre la figure 4.4. Plus précisément, certains tri-grammes obtenus à partir de la première phrase ne figurent pas dans ceux de la deuxième phrase.

[1] *Personnel et ambiance très agréable.*

[2] *Les plats sont bons mais pas assez copieux.*

[3] *Les plats sont copieux mais pas assez bons.*

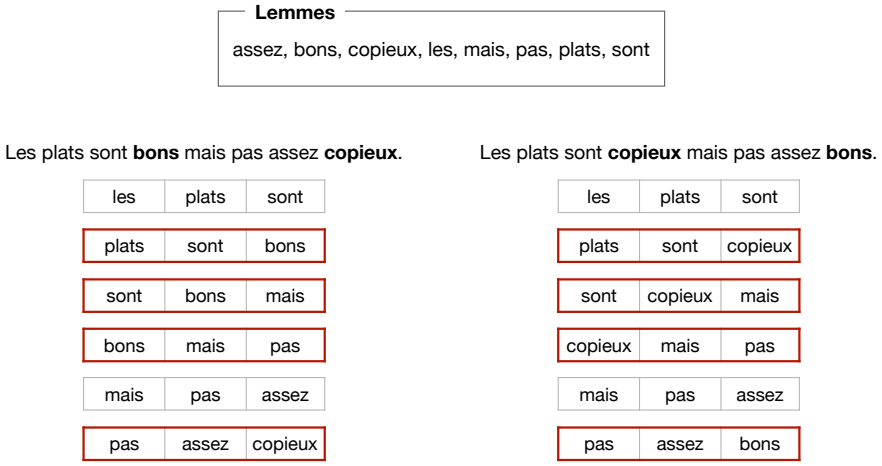


FIG. 4.4 : Exemples des tri-grammes composés d'une même liste de mots

En raison de la courte longueur des phrases (en moyenne de 2,75 mots), nous n'avons étudié que les n-grammes ayant une valeur *n* de deux et trois, donc les bi-grammes et les tri-grammes. Tout comme nous l'avons constaté à travers la fréquence de lemmes, une simple fréquence n'est pas suffisante pour dégager les spécificités des expressions de chaque catégorie d'évaluation, car elle ne prend pas en compte la fréquence relative des catégories. Pour cette raison, nous réalisons une analyse TF-IDF sur la fréquence des n-grammes.

La méthode TF-IDF, introduite par Salton et Buckley (1988), permet de connaître les mots significatifs. Le TF indique la fréquence d'un terme : plus le mot est présent, plus il est saillant dans le document en question. En revanche, l'IDF est l'inverse de la proportion des documents où le terme apparaît : il s'agit de considérer si un mot apparaît souvent dans de trop nombreux documents. Si tel est le cas, cela implique que le mot est peu significatif et que son importance doit donc être sous-pondérée, comme les mots-outils par exemple. Dans l'ensemble TF-IDF, un poids élevé est ainsi attribué aux mots qui se trouvent fréquemment dans un document particulier, mais non dans de nombreux documents du corpus. Autrement dit, si un mot apparaît fréquemment dans un document, mais très rarement ailleurs, il est probablement très représentatif du contenu du document donné. Le tableau 4.2 montre le score de TF-IDF concernant quelques mots figurant dans notre corpus : notons que les mots-outils (*un*, *de*, *le*) obtiennent les scores les plus faibles, qui sont égaux à zéro.

TAB. 4.2 : Exemple des scores de TF-IDF

Lemme	**TF-IDF**
retourner	0,0251125782795246
recommander	0,00729505297835829
hâte	0,0043966384204499
…	…
un	0
de	0
le	0

Une analyse TF-IDF permet ainsi d'identifier les mots qui sont discriminants d'un document (l'ensemble des phrases correspondant à une catégorie donnée dans notre cas) par rapport à d'autres documents (toutes les phrases des autres catégories). Pour ce faire, le langage R et le package `tidytext` (Silge et Robinson 2016) ont été exploités. Les figures 4.5 et 4.6 illustrent respectivement l'analyse TF-IDF des bi-grammes et des tri-grammes.

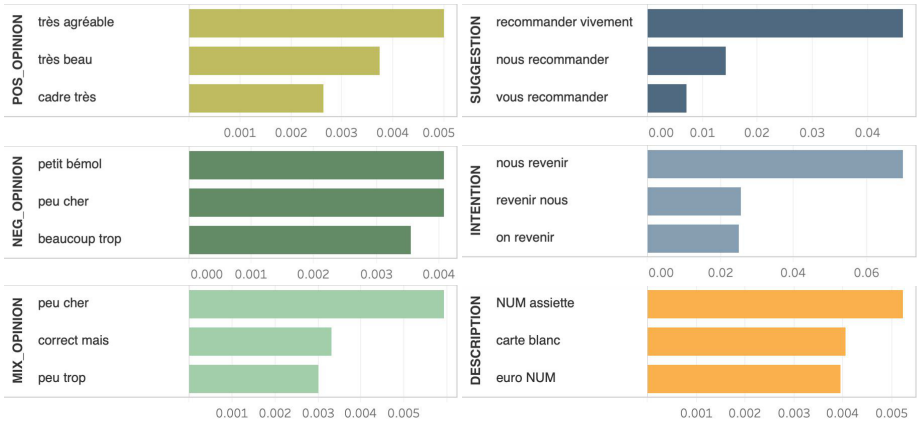


FIG. 4.5 : TF-IDF des bi-grammes

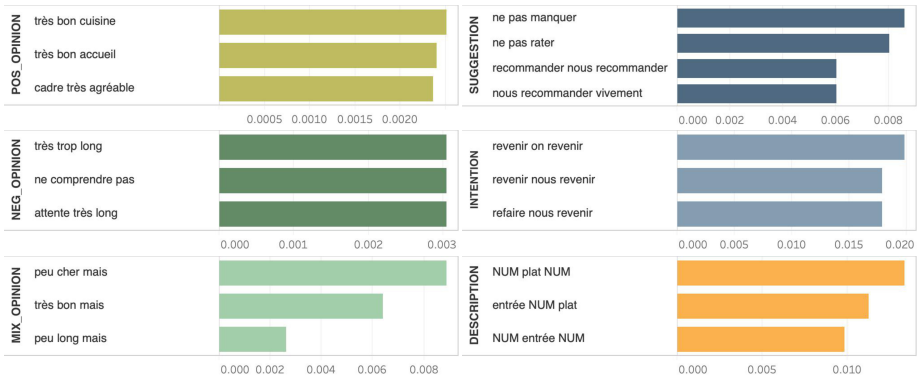


FIG. 4.6 : TF-IDF des tri-grammes

Nous pouvons constater qu'en général, les lemmes qui constituent les bi-grammes et les tri-grammes se retrouvent également présents dans les mots-clés extraits précédemment. En outre, sur ces deux figures :

- l'opinion positive contient les lemmes ayant une polarité positive (*agréable*, *beau*, *bon*) et, notamment, l'adverbe *très* qui renforce l'opinion. De surcroît, en comparaison avec d'autres catégories, l'opinion positive comporte relativement plus de lemmes qui indiquent les cibles de l'évaluation concernant le restaurant (*cadre*, *cuisine*, *accueil*) dans les tri-grammes ;

- l'opinion négative, quant à elle, comprend les indices qui peuvent conduire l'opinion à une polarité négative (*bémol*, *cher*, *trop*) ;

- l'opinion mixte est marquée non seulement par la conjonction *mais*, mais aussi par l'adverbe (*peu*), qui semble être dérivé d'*un peu*. Ce dernier, expression d'une faible mesure par opposition à *beaucoup*, a pour rôle d'atténuer l'opinion négative avant ou après l'introduction d'une opinion positive dans une phrase [4], [5] ;

- en ce qui concerne la suggestion, les bi-grammes contiennent particulièrement les verbes de parole désignant l'ordre ou le conseil (*recommander*) et les tri-grammes contiennent la négation (*ne...pas*) pour inciter le destinataire à mener une action [6] ;

- l'intention est très présente avec les verbes ayant le préfix *re-* pour indiquer l'itération d'une action (*revenir*) ;

- nous remarquons également la présence du pronom *nous* assez fréquente dans la suggestion et l'intention, qui se réfère à la réalisation d'une action (au moment de la communication ou dans le futur) par le locuteur ;

- finalement, la description est liée aux expressions comportant des chiffres (*NUM*) permettant de décrire les informations factuelles associées à l'expérience vécue.

[4] *Les prix sont un peu élevés mais cela les vaut vraiment !!*

[5] *Très bon et inventif mais un peu trop cher.*

[6] *À ne pas manquer !*

Nous avons observé les n-grammes significatifs à l'aide de l'analyse TF-IDF. Cependant, cette observation ne nous renseigne pas suffisamment sur les multiples relations qui existent entre les lemmes. Pour mieux les comprendre, nous avons réalisé une visualisation graphique des relations entre les deux lemmes dans l'ensemble du corpus grâce aux bi-grammes. En exploitant le langage R et le package `tidytext`, le graphique obtenu permet de présenter la relation entre deux mots au moyen d'un graphique assemblé avec des nœuds connectés. La figure 4.7 illustre les clusters de réseaux de mots couramment utilisés ensemble. Pour rendre la visualisation graphique compréhensible, seules les associations les plus fréquentes entre les lemmes sont affichées. Il s'agit des bi-grammes qui ont au moins 60 fréquences dans l'ensemble du corpus.

Nous remarquons que l'adverbe *très* constitue le centre des nœuds : il est souvent suivi d'adjectifs (particulièrement *bon*, et aussi *agréable*, *frais*, *sympathique*, *sympa*, *professionnel*) ou précédé par les noms (*prix*, *accueil*, *restaurant*, *personnel*, *cadre*, *service*). En outre, ces derniers, indiquant les cibles sur lesquelles portent l'évaluation, sont attachés aux lemmes qui les qualifient ; dans la plupart des cas, il s'agit des adjectifs comme *agréable*, *parfait*, *impeccable* pour le *service* ou *inventif*, *raffiné* pour la *cuisine*. Nous observons aussi des paires ou des triades à la périphérie qui constituent de courtes séquences communes telles que la négation (*ne…pas*) ou sont associées à la gastronomie (*amuse-bouche*, *foie gras*, *menu dégustation/prestige*). La visualisation montre ainsi la relation entre deux mots, principalement au sein de l'opinion positive puisque cette dernière constitue la majorité des bi-grammes les plus fréquents[51]. Nous notons, néanmoins, ceux qui représentent la suggestion (*recommander vivement*) et l'intention (*nous revenir*).

[51] Une analyse de notre corpus montre que parmi les 50 fréquents bi-grammes, 35 bi-grammes concernent l'opinion positive.

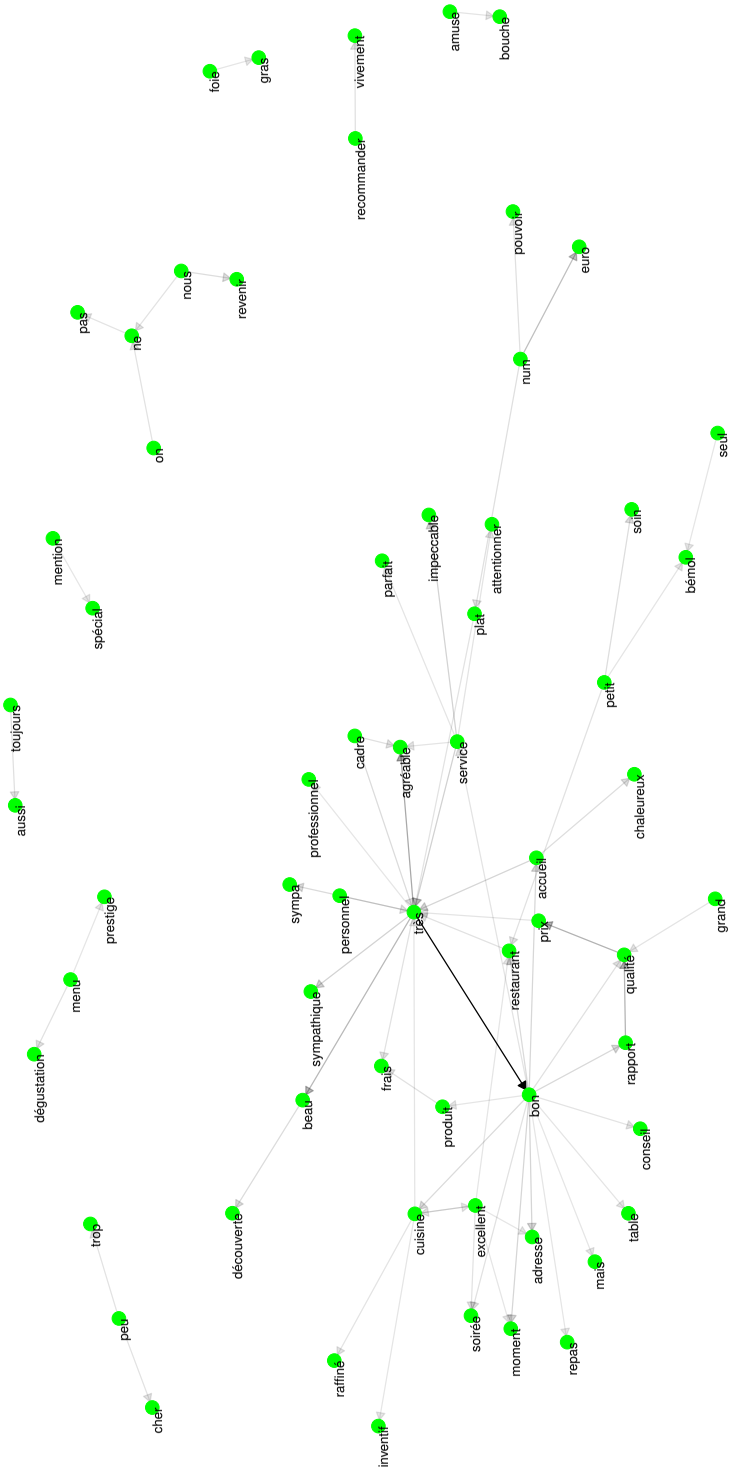


FIG. 4.7 : Clusters de réseaux de mots couramment utilisés ensemble

4.4 Distribution des parties du discours

Dans cette partie, nous étudions les caractéristiques morphologiques et syntaxiques de chaque catégorie d'évaluation. Nous nous focalisons sur les catégories morphosyntaxiques (ADJ, ADV, DET, NOUN, PRON et VERB). Le choix de ces catégories est guidé par la nature du corpus et la tâche du traitement visée. D'après une observation manuelle, les avis ont tendance à adopter une organisation « télégraphique » ou « fragmentée », ce qui a été déjà relevé par Marcoccia (2016). Plus précisément, certaines composantes de la phrase sont absentes et donc « de nombreuses phrases peuvent être composées de clauses ou de propositions incomplètes (qui ne contiennent pas de sujet et de prédicat) » (Marcoccia 2016, loc.1392[52]), comme dans l'exemple [7].

[7] *Magnifique endroit, très bonne cuisine et accueil irréprochable.*

Lorsque nous disposons d'un corpus de données volumineux, il devient difficile de dégager manuellement leurs caractéristiques pertinentes. Pour cette raison, certains chercheurs se tournent vers l'analyse factorielle des correspondances (AFC) (Desagulier 2020 ; Poudat et Landragin 2017). Cette méthode d'origine française (Benzécri *et al.* 1973) permet de représenter un vaste tableau des données (figure 4.8) dans un espace de dimension réduit ; elle résume le tableau de contingence et exprime les différentes associations intéressantes par le biais d'une visualisation graphique en deux dimensions.

[52] [version Kindle iOS].

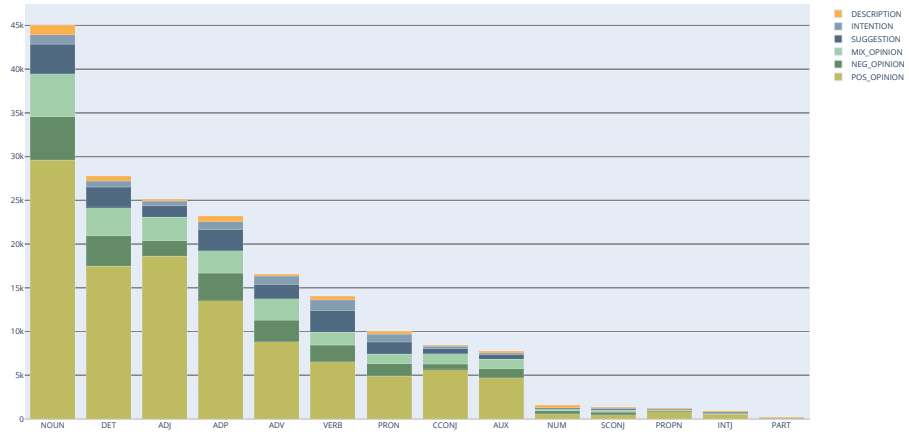


FIG. 4.8 : Diagramme en bâtons des deux variables (la catégories d'évaluation et la position)

Un test du khi-deux d'indépendance est conçu pour vérifier l'existence d'une liaison entre deux variables qualitatives au sein d'une population (Desagulier 2017). Ce test a été effectué dans le but d'examiner l'association entre les catégories morphosyntaxiques et les différentes catégories d'évaluation. Une dépendance significative entre ces dernières et les catégories morphosyntaxiques a été observée avec un risque d'erreur α = 5 % ($\chi 2$ = 12 318 ; ddl[53] = 35 ; p < 2,2e-16). L'AFC a été effectuée ensuite en utilisant le package FactoMineR (Lê *et al.* 2008), disponible dans le langage R, pour l'analyse. Les valeurs propres (*eigenvalue*)[54] associées à chaque dimension sont présentées dans la figure 4.9 : les dimensions 1 et 2 expliquent respectivement environ 84,2 % et 7,7 % de l'inertie totale[55]. L'inertie projetée cumulée nous indique que les deux premières dimensions rendent compte à elles seules de 91,9 % des variations observées dans nos données.

La figure 4.10 illustre une analyse symétrique standard de l'analyse factorielle des correspondances : il s'agit d'un nuage de points dans lequel les deux variables sont représentées simultanément dans un même espace. Les catégories d'évaluation sont présentées par des points bleus et

[53]Le degré de liberté.
[54]La quantité d'informations retenue par chaque axe.
[55]La somme des variances des variables considérées (Baccini 2010).

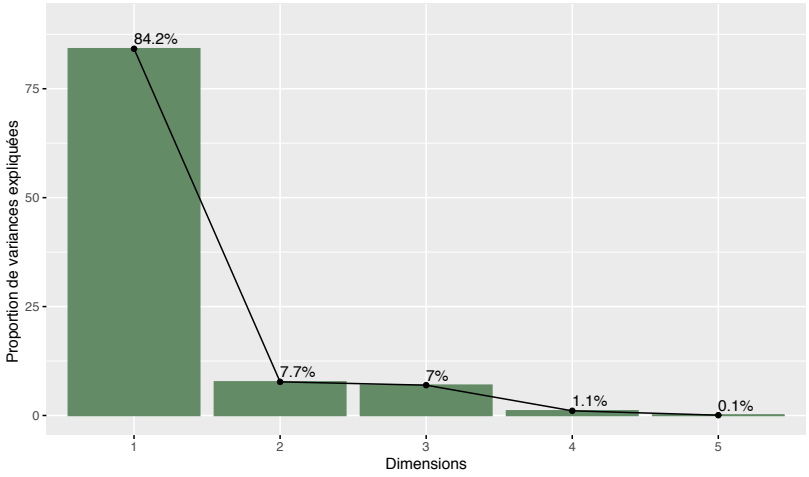


FIG. 4.9 : Valeurs propres associées à chaque dimension

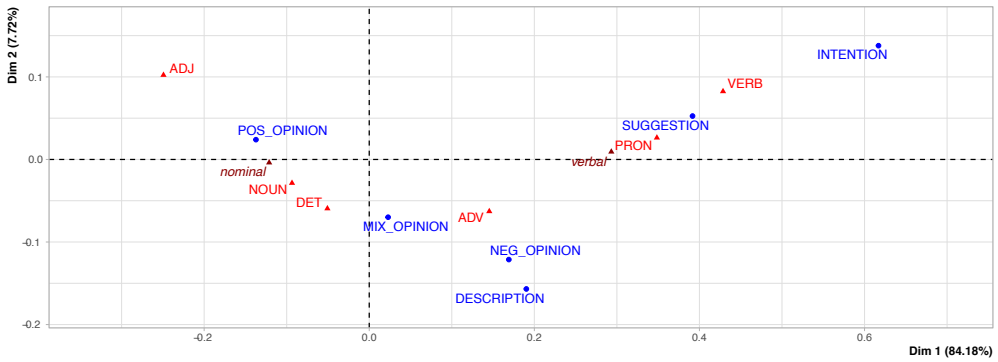


FIG. 4.10 : Analyse symétrique standard de l'analyse factorielle des correspondances

les catégories morphosyntaxiques par des triangles rouges. Les éléments supplémentaires sont également projetés, permettant de regrouper les informations pertinentes : les valeurs totales de l'adjectif, du déterminant et du nom sont indiquées sous le nom « nominal » et celles de l'adverbe, du pronom et du verbe comme « verbal ». Cependant, ces éléments ne contribuent pas à la constitution des dimensions. Le centre du graphique correspond au profil moyen des deux variables. Un point donné s'écarte du centre de gravité s'il contribue aux axes davantage que le profil moyen. Les points en corrélation négative sont éloignés, tandis que ceux qui ont

des profils similaires se situent de façon relativement proche.

Une coupure est observée entre l'opinion positive, qui se trouve à gauche du graphique, et les catégories restantes à droite. Parmi ces dernières, l'intention est davantage éloignée du centre de gravité par rapport aux autres catégories. En s'appuyant sur le premier axe, les adjectifs, les noms et les déterminants sont associés à l'opinion positive, alors que les verbes et pronoms sont étroitement liés à l'intention et à la suggestion. Les adverbes se situent à proximité de la description, de l'opinion négative et de l'opinion mixte. Le graphique montre que l'opinion positive contient beaucoup de syntagmes nominaux, ce qui est observable dans les données (voir les exemples [7]-[9]). En revanche, les syntagmes verbaux sont notamment associés aux suggestions et aux intentions. Ces dernières sont toutes deux liées à l'action, comme nous l'avons remarqué lors de l'analyse qualitative et de l'analyse lexicométrique. La suggestion proposée par un client implique une action future de la part du restaurateur et de ses futurs clients, et l'intention concerne l'action qui sera effectuée par le client lui-même. Notons également que les catégories décrivant les syntagmes verbaux sont légèrement associées à la description et à l'opinion négative : la première décrit parfois une situation dynamique [10] ou un état [11] dans lesquels l'expérience se déroule, fonctionnant donc comme l'arrière-plan de l'expérience au restaurant, et la seconde est souvent exprimée par la négation d'un prédicat. Enfin, l'opinion mixte se place relativement au milieu de la figure, car elle présente des caractéristiques de l'opinion positive et de l'opinion négative.

[8] *Très bon rapport qualité prix et service très aimable et efficace.*

[9] *Cuisine originale avec de bons produits frais.*

[10] *Nous sommes arrivés à 21h 30 et avons quitté le restaurant les tout derniers.*

[11] *Il faisait 7 degrés avec un petit crachin.*

4.5 Position : l'ouverture et la clôture

Selon Polanyi et Zaenen (2006), il est crucial d'examiner la position d'une information dans un avis parce que les remarques qui se situent tout au début ou à la toute fin d'un avis ont des valeurs différentes de celles qui se trouvent dans les autres positions. Taboada (2011) analyse, en ce sens, le début, le milieu et la fin des avis de films. Plus précisément, elle propose plusieurs étapes (*stages*) qui se composent du thème, de l'intrigue, des personnages, de l'arrière-plan et de l'évaluation, et constate que les avis commencent principalement par l'arrière-plan ou l'évaluation, et finissent par l'évaluation.

L'objectif de cette partie est ainsi d'observer l'ouverture et la clôture de l'évaluation rédigée par les clients afin de vérifier s'il existe une liaison entre la catégorie d'évaluation et sa position dans un avis. C'est pourquoi nous étudierons dans quelle mesure la catégorie dépend de la position. D'après une observation manuelle, la description se manifeste souvent au début d'un avis, alors que la suggestion et l'intention ont une tendance à apparaître vers la fin, comme on peut l'observer dans les exemples suivants :

[12] ***Nous hésitions entre plusieurs grands restaurants sur Paris pour célébrer nos 15 ans de mariage et avons opté pour ce restaurant après avoir lu plusieurs critiques sur internet.*** **[DESCRIPTION]** *Bien nous en a pris ! Nous avons opté pour le menu Prestige : quelle expérience fabuleuse : une pure merveille et un voyage incroyable des sens ! Nous avons apprécié l'attention du personnel, l'espace entre chaque table. La sommelière a fait un superbe travail pour accompagner chaque plat, le service impeccable et chaleureux ce qui n'est pas le cas dans tous les restaurants 2 ou 3 étoiles à Paris.* ***Nous avons hâte d'y retourner :)*** **[INTENTION]**

[13] ***Menu déjeuner + accord mets/vin à 63 euros.*** **[DESCRIPTION]** *Plats savoureux et intéressant. Service efficace, décor sobre et moderne.* ***Il faut y aller le soir pour les menus au prix juste.*** **[SUGGESTION]**

[14] *Restaurant très bien situé dans Paris, une carte qui donne envie et un cadre des plus fabuleux accompagné d'un personnel réactif et efficace.* ***À refaire sans hésiter !*** **[INTENTION]**

[15] *Très bon accueil par un personnel très professionnel. Très bonne cuisine. Soirée inoubliable sur la terrasse : des mets de qualité (menu du marché accessible et très fin) et service hors pair !* ***À recommander pour les amoureux les soirs d'été...*** **[SUGGESTION]**

Ces observations manuelles ont été vérifiées automatiquement. Pour ce faire, l'ouverture et la clôture (première et dernière phrase) ont été considérées comme les positions les plus caractéristiques. Les avis composés d'une seule phrase n'ont pas été pris en compte. Les données relatives aux observations qui concernent les deux variables (la catégorie d'évaluation et la position) se présentent sous la forme d'un tableau de contingence (ou un tableau croisé), comme dans le tableau 4.3. Ce tableau peut être également visualisé au moyen d'un diagramme en bâtons (figure 4.11) qui facilite l'observation de la tendance générale. Bien que sa fréquence diminue en dernière position, l'opinion positive apparaît fréquemment dans les deux positions. L'intention, l'opinion négative et la suggestion finissent souvent les évaluations, tandis que la description, peu présente d'une manière générale, semble être placée plutôt au début. L'opinion mixte apparaît avec une tendance similaire dans les deux positions.

Or, pour vérifier la liaison entre la catégorie et la position d'une manière plus précise, nous avons produit le graphique d'association (*association plot*) entre les deux. Avant de réaliser le graphique, nous avons d'abord effectué un test du khi-deux afin de vérifier si les catégories qui se

TAB. 4.3 : Tableau de contingence des deux variables (la catégories d'évaluation et la position)

Catégorie	Première	Dernière	Total
POS_OPINION	3790	2253	6043
NEG_OPINION	195	505	700
MIX_OPINION	269	244	513
SUGGESTION	131	919	1050
INTENTION	50	611	661
DESCRIPTION	117	20	137
Total	4552	4552	9104

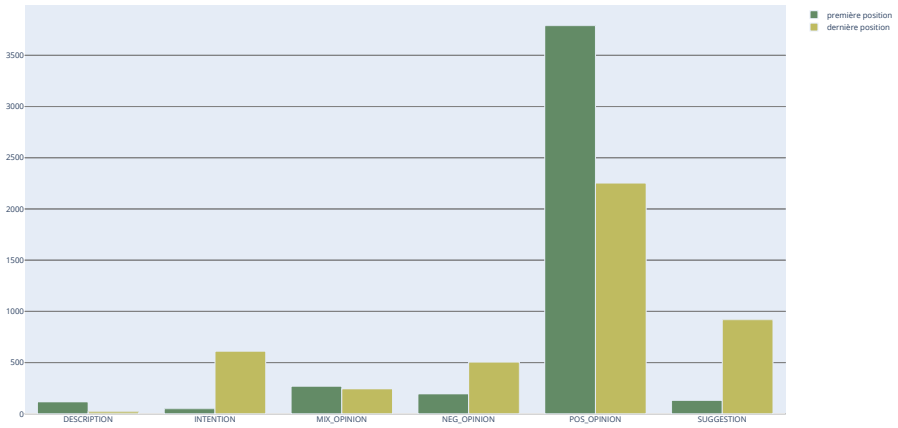


FIG. 4.11 : Diagramme en bâtons des deux variables (la catégories d'évaluation et la position)

situent en première ou en dernière position sont dues au hasard ou à l'interdépendance. Le test a été effectué montrant une dépendance entre la catégorie et sa position avec un risque d'erreur α = 5 % ($\chi 2$ = 1665,6 ; ddl = 5 ; p < 2,2e-16).

Si le test examine l'existence de la liaison entre les deux variables, il ne renseigne pas sur l'intensité de l'association (Desagulier 2017). Les résidus de Pearson, écarts standardisés obtenus en comparant la valeur observée et la valeur attendue, permettent d'analyser plus finement cette

association. Nous pouvons produire un tableau des résidus de Pearson par le biais du graphique d'association de Cohen-Friendly (Cohen 1980 ; Friendly 1992), présenté dans la figure 4.12.

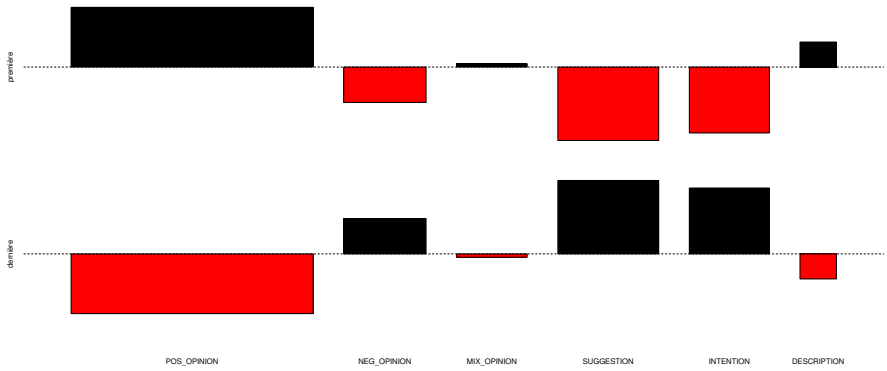


FIG. 4.12 : Graphique d'association entre la catégorie d'évaluation et la position

Le principe de lecture du graphique est le suivant : chaque cellule est représentée par un rectangle dont la hauteur est proportionnelle au résidu de Pearson, tandis que la largeur est proportionnelle à la racine carrée de l'effectif attendu. La ligne pointillée indique la ligne de référence (*baseline*) représentant l'indépendance. Si la valeur du résidu est ainsi positive (négative), la cellule se situe au-dessus (au-dessous) de la ligne de référence.

Nous constatons dans la figure 4.12 que l'opinion positive et la description sont significativement placées en première position d'un avis alors que la suggestion, l'intention et l'opinion négative sont placées en dernière position. Les clients commencent fréquemment leur avis par un arrière-plan (la description) ou directement par une opinion positive. Cette dernière semble naturelle car la fonction principale d'une évaluation est de juger la valeur d'une entité donnée. Par ailleurs, les avis sont biaisés en faveur de la polarité positive, comme nous l'avons vu précédemment (voir chapitre 1). Les avis ont une tendance à finir par une impression globale en exprimant leur souhait de retourner ou non au restaurant (l'intention)

ou en faisant des suggestions aux lecteurs (les clients potentiels et le restaurateur).

4.6 Conclusion

Nous venons de décrire différentes analyses quantitatives réalisées en nous appuyant sur le corpus annoté. Pour ce faire, nous avons, d'abord, prétraité le corpus afin de faciliter son exploration automatique. Chaque catégorie est analysée et décrite à travers les mots-clés et les n-grammes ayant des scores TF-IDF élevés, la distribution des parties du discours, et sa position au sein d'un avis. Les diverses analyses quantitatives ont permis de décrire les évaluations sous un angle nouveau (tableau 4.4). L'opinion se caractérise ainsi par le lexique évaluatif et les cibles sur lesquelles elle porte, et se manifeste à travers les syntagmes nominaux. La description, l'intention et la suggestion impliquent les syntagmes verbaux qui permettent de décrire l'arrière-plan de l'expérience (la description), d'effectuer une action par le client lui-même (l'intention) ou de suggérer une action future au restaurant ou aux futurs clients (la suggestion). En ce qui concerne la position des catégories dans un avis, l'opinion et la description sont placées majoritairement au début d'un avis, alors que la suggestion et l'intention sont positionnées à la fin. L'analyse menée dans cette section a ainsi permis de mieux comprendre, d'un point de vue quantitatif, la façon dont les clients évaluent les restaurants. En outre, elle a également apporté une vision plus précise des spécificités de chaque catégorie d'évaluation proposée, faisant de chacune d'elle une source d'information utile tant pour les clients que pour les restaurateurs. Nous présenterons, dans le chapitre suivant, la détection automatique de ces catégories qui tient compte des particularités linguistiques relevées et décrites ci-dessus.

TAB. 4.4 : Synthèse de l'analyse quantitative

Catégorie	**Lexicométriques**	**Catégories morphosyntaxique**	**Position**
Opinion positive	• Les lexiques évaluatifs ayant une polarité positive • Les cibles sur lesquelles elle porte • L'adverbe *très* qui renforce l'opinion positive	Les syntagmes nominaux	L'ouverture
Opinion négative	• Les lexiques évaluatifs ayant une polarité négative • La négation	Les syntagmes verbaux (peu fréquent)	La clôture
Opinion mixte	• La conjonction *mais* • L'adverbe *peu* (dérivé d'*un peu*)		L'ouverture (rare)
Suggestion	• Les verbes d'ordre, de conseil • Les pronoms (*vous*, *nous*/*on*)	Les syntagmes verbaux	La clôture
Intention	• Les verbes itératifs ayant le préfix *re-* • Les pronoms (*je*, *nous*/*on*)		
Description	• Les lexiques concernant les informations factuelles (notamment les chiffres)	Les syntagmes verbaux (peu fréquent)	L'ouverture

Chapitre 5

Détection automatique de l'évaluation

5.1 Introduction

Comme nous l'avons observé dans l'état de l'art (chapitre 2), les recherches dans le domaine du TAL consacrées à la fouille d'opinions se limitent majoritairement à la detection des categories d'opinions positives et negatives et à celle de leur cible. Cependant, certaines recherches se sont dernièrement réorientées vers l'extraction de suggestions (Brun et Hagège 2013 ; Negi *et al.* 2016 ; Negi et Buitelaar 2015 ; Negi *et al.* 2018). Ainsi l'atelier SemEval 2019 par exemple, proposait récemment des tâches visant à extraire des suggestions de forums et d'avis portant sur les hôtels. Le travail de Benamara *et al.* (2017a) propose de traiter l'évaluation d'un produit grâce à la notion d'intention. Cette tendance montre que l'intérêt des chercheurs pour l'évaluation sur le web dépasse largement la notion d'opinion positive ou négative. En ce sens, nous nous intéresserons, dans ce chapitre, à la détection automatique d'évaluation, sans pour autant nous limiter à la valeur positive ou négative. Nous présenterons la détection automatique des six catégories d'évaluation proposées précédemment (opinion positive/négative/mixte, suggestion, intention et description) en exploitant deux techniques : l'apprentissage de surface et l'apprentissage profond.

L'un des problèmes que nous avons rencontrés concerne le déséquilibre dans la répartition des classes d'évaluation créées. Lorsque les données subissent une répartition biaisée de leurs classes, les classifieurs peuvent considérer que les observations des classes minoritaires ne sont pas signi-

fiantes. Ils ignoreront alors celles-ci, et donneront des résultats performants en général, mais très mauvais concernant les classes minoritaires. Le déséquilibre des classes a été pris en compte en choisissant les mesures d'évaluation appropriées, en adoptant l'approche Cost-Sensitive Learning (CSL) sur les algorithmes ou en choisissant enfin l'échantillonnage des données.

Notre objectif au cours des expériences a été d'obtenir de bonnes performances aussi bien pour les classes minoritaires que pour les classes majoritaires. Dans un premier temps, nous présenterons la méthodologie générale, puis les démarches réalisées lors des expériences, la préparation des données, la classification et le traitement des données déséquilibrées ainsi que l'évaluation du module automatique développé.

5.2 Méthodologie générale

La fouille d'opinions est souvent assimilée à un défi de classification automatique des textes qui consiste à associer automatiquement une classe à chaque donnée d'entrée. Les apprentissages supervisés (apprentissage de surface et l'apprentissage profond) sont aujourd'hui les approches dominantes pour ce type de tâches. Elles nécessitent toutes les deux des étiquettes préalablement associées aux données, par exemples « spam » ou « non-spam » pour identifier la nature d'un email, « positive » ou « négative » pour classifier les documents en polarité. Dans le cadre de cette étude, nous disposons de données étiquetées selon la typologie d'évaluation élaborée (POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION et DESCRIPTION).

Afin de détecter automatiquement l'évaluation modélisée dans le corpus d'avis, nous avons suivi plusieurs étapes (figure 5.1). Les avis ont d'abord été prétraités, segmentés et transformés en vecteurs. Nous avons testé différents classifieurs utilisés pour l'apprentissage de surface et l'apprentissage profond, tout en prenant en compte le déséquilibre dans la distribution des classes. Chacun des classifieurs a produit un modèle global, qui effectue une prédiction parmi les six possibilités. Nous avons cherché à obtenir de bonnes performances à la fois sur les classes minoritaires et sur les classes majoritaires. Dans la section qui suit, nous exposerons

plus en détails les étapes suivies pour réaliser les expériences.

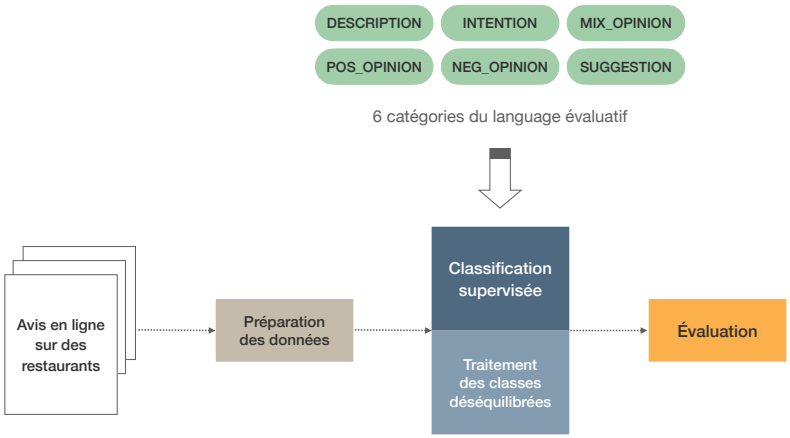


FIG. 5.1 : Processus d'apprentissage supervisé

5.3 Préparation des données

5.3.1 *Normalisation*

Le prétraitement des données textuelles consiste à proposer pour chaque mot du texte « sa forme normativement correcte ou une forme qui lui est flexionnellement liée » (Baranes et Sagot 2014, p.137). Pour les expériences de la détection automatique, les données textuelles normalisées dans le cadre de l'analyse quantitative (voir chapitre 4) ont été utilisées à nouveau. Cette étape de prétraitement est essentielle, puisque les résultats des expériences dépendent considérablement de la qualité du prétraitement.

5.3.2 *Représentation vectorielle*

Afin de représenter le mieux possible les données, il convient de prendre en compte leurs spécificités ainsi que celles du modèle de la tâche. Puisque les classifieurs d'apprentissage supervisé reposent sur la représentation numérique, les données textuelles doivent être transformées en vecteurs,

c'est-à-dire en points dans un espace vectoriel de très grande taille auxquels nous pourrons appliquer les classifieurs. Dans cette partie, nous explorerons certaines méthodes de représentation des données textuelles : le sac de mots et le TF-IDF d'une part, et le plongement de mots (*word embedding*), d'autre part.

5.3.2.1 Représentation en sac de mots et TF-IDF

La représentation en sac de mots est la méthode la plus simple et consiste à représenter chaque document par un vecteur de fréquence de mots, dit « sac de mots » (*bag of words* ou BOW ; Salton et Buckley 1988). Un document est représenté par les vecteurs de fréquence des mots qui le composent, permettant, pour chaque mot, de voir sa fréquence au sein de l'ensemble de données. Ce dernier est ainsi représenté sous forme de matrice utilisée comme entrée du classifieur et composée de l'ensemble des N phrases ou documents, chacun représenté par un vecteur de dimension égale à la taille du vocabulaire. La figure 5.2 offre un exemple de cette représentation. Dans l'exemple [1], le token *plat* est présent une fois, *très* deux fois, et ainsi de suite. Le token *service* n'est pas mentionné dans cette phrase, mais il l'est dans les autres ; par conséquent, dans le vecteur de cette phrase, l'entrée de ce token correspond à 0.

[1] *plat très bon et cadre très agréable*

Cependant, la faiblesse de cette représentation se trouve dans les données clairsemées (*sparse*). Puisque la dimension du vecteur est égale au nombre de tous les mots distincts qui y apparaissent, qui est finalement grand, la représentation est clairsemée (elle contient des zéros). Un autre défaut réside dans la négligence des informations sur la structure des données, c'est-à-dire qu'une phrase est considérée comme un ensemble de mots non ordonnés dont la position n'est pas prise en compte, ce qui entraîne une perte d'informations sur les relations sémantiques entre les mots.

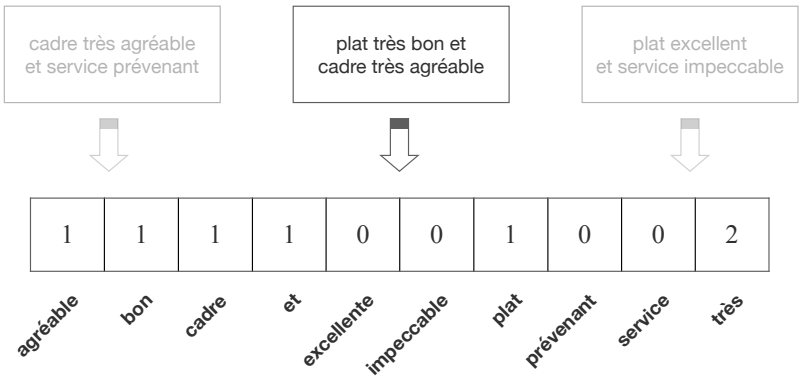


FIG. 5.2 : Représentation en sac de mots

À titre d'exemple, *pas mal* signifie *plaisant* ou *agréable*[56], tandis que *pas* et *mal* sont respectivement liés à la négation et la connotation négative. Néanmoins, dans la pratique, la représentation en sac de mots est largement utilisée, car elle donne des résultats satisfaisants. Une approche avancée consiste à pondérer les mots distinctement plus fréquents dans un document spécifique que dans d'autres documents : il s'agit de la méthode TF-IDF, que nous avons déjà exploitée dans l'analyse quantitative (chapitre 4).

Dans le but de combler les lacunes créées par la perte d'informations sur les relations sémantiques entre les mots, nous pouvons utiliser aussi les n-grammes, ce qui a été déjà démontré dans le chapitre précédent. Les n-grammes retiennent la structure séquentielle des données d'origines ; par conséquent, leur application dans la représentation rend les données plus informatives. Le désavantage de ce modèle est que plus n est élevé, plus la représentation aboutit à un vecteur de grande dimension et à un vecteur creux, ce qui augmente le coût calculatoire. De ce fait, les n-grammes ayant une valeur de n élevée ne contribuent pas nécessairement à une amélioration de la performance.

Les données textuelles ont été représentées sous forme de sac de

[56] https://www.larousse.fr/dictionnaires/francais/mal/48790?q=pas+mal#48706 [consulté le 28 septembre 2020].

mots et de TF-IDF, au moyen de la bibliothèque scikit-learn[57] (Pedregosa *et al.* 2011). Deux versions des données lemmatisées ont été testées : l'une comprenant les mots-outils et l'autre sans les mots-outils. Ce choix a permis d'observer la manière dont les mots-outils sont traités afin d'obtenir le meilleur résultat. Les paramètres utilisés sont :

- **`use_idf`** : pour choisir entre la représentation basée sur la fréquence (`CountVectorizer`) et la représentation pondérée par TF-IDF (`TfidfVectorizer`) ;
- **`min_df`** : pour ignorer les mots dont la fréquence est inférieure au chiffre choisi – ici nous avons choisi un seuil de 2 ;
- **`ngram_range`** : pour déterminer le nombre des séquences de mots (les n-grammes), pour n allant de 1 à 3.

Afin de déterminer les meilleures combinaisons des paramètres, nous avons utilisé la technique `GridSearch`[58] dont dispose la bibliothèque scikit-learn. Les résultats obtenus selon les versions du corpus et la combinaison des paramètres choisis seront présentées dans la section 5.6.2.

5.3.2.2 Plongement de mots : Word2vec

Les plongements lexicaux font partie d'une approche basée sur les réseaux de neurones et visant à créer une représentation vectorielle du vocabulaire. À la différence des méthodes précédentes, celle-ci prend en compte une certaine similarité sémantique entre les mots. Il existe différents modèles de cette approche, tels que Word2vec (Mikolov *et al.* 2013a,b), GloVe (Pennington *et al.* 2014), FastText (Bojanowski *et al.* 2017), WordRank (Ji *et al.* 2015), Varembed (Bhatia *et al.* 2016) et Poincaré (Nickel et Kiela 2017). Dans cette étude, nous avons utilisé le modèle Word2vec, élaboré par les chercheurs de Google, qui a donné naissance à de nouveaux standards dans les applications du TAL (Brownlee 2017). À l'aide de grandes quantités de données non-étiquetées, il permet de représenter un mot par un vecteur dans le but de capturer une certaine

[57] `http://scikit-learn.org/stable/` [consulté le 28 septembre 2020].

[58] Il s'agit d'une méthode qui permet d'améliorer la performance du modèle en affinant ses paramètres. Elle cherche les meilleurs hyperparamètres du modèle en essayant toutes les combinaisons possibles de ceux-ci.

sémantique lexicale. Deux mots ayant une signification proche ont alors une représentation similaire, c'est-à-dire qu'ils sont projetés à proximité dans l'espace vectoriel.

Il existe deux architectures de la représentation Word2vec : *continuous bag of words* (CBOW) et *skip-gram*. CBOW prédit le mot cible en s'appuyant sur le contexte local, à savoir les mots qui l'entourent. Pour [2] par exemple, le mot cible *vin* est prédit en fournissant les mots constituant son contexte (*carte*, *des*, *est* et *limitée*) en entrée. A contrario, le *skip-gram* se base sur le mot cible pour prédire les mots qui l'entourent. Cet exemple, avec chacune des architectures, est illustré dans la figure 5.3.

[2] *Seul petit bémol, la carte des vins est limitée et un peu chère.*

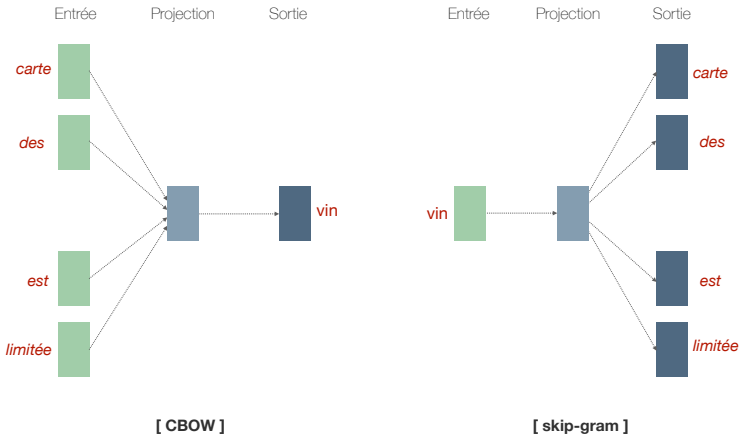


FIG. 5.3 : Exemple de CBOW et *skip-gram*

La représentation Word2vec a été créée à l'aide de la bibliothèque Gensim (Řehůřek et Sojka 2010). Les paramètres configurés sont :

- `size` : le nombre de dimensions, c'est-à-dire la longueur du vecteur qui représente les mots – la valeur par défaut (100 dimensions) a été sélectionnée ;

- `window` : la taille de la fenêtre, où pour une taille égale à n, n mots précèdent le mot noyau et n mots le suivent – n a été fixé à 1 ;
- `min_count` : le nombre minimum de mots à prendre en compte lors de l'entraînement – un seuil de 2 a été défini ;
- `workers` : le nombre de nœuds de calcul – 4 nœuds ont été choisis ;
- `sg` : l'architecture d'entraînement, soit CBOW (0) ou *skip-gram* (1) – nous les avons fait varier au cours des expériences.

Par ailleurs, nous n'avons exploité qu'une seule version des données, dans laquelle les données ont été lemmatisées et les mots-outils ont été conservés. En effet, ces derniers n'ont pas été supprimés en raison du principe sur lequel s'appuie la représentation Word2vec, qui encode les propriétés grammaticales et sémantiques du contexte dans lequel un mot est apparu. Si les mots-outils sont peu significatifs pour la représentation en sac de mots, ils permettent d'apprendre le contexte dans la représentation Word2vec.

Quelques exemples du Word2vec (de type CBOW) réalisés sur nos données sont présentés comme suit :

```
> model.wv.similarity('délicieux', 'succulent')
0.933402371778884

> model.wv.most_similar('sympathique')
[('attentif', 0.988103985786438),
('accueillant', 0.9850661754608154),
('sympa', 0.9849212765693665),
('attentionner', 0.9818465709686279),
('chaleureux', 0.9814795851707458),
('aimable', 0.9797272682189941),
('professionnel', 0.9783313274383545),
('souriant', 0.9781066179275513),
('discret', 0.9778811931610107),
('impeccable', 0.9769046902656555)]

> model.wv.most_similar(positive=['service',
'délicieux'], negative=['plat'])
[('sympathique', 0.9466606378555298),
('souriant', 0.932893693447113),
('sympa', 0.9324531555175781),
('chaleureux', 0.9305021166801453),
('attentif', 0.9296811819076538),
('accueillant', 0.9294965267181396),
('attentionner', 0.9278693199157715),
('agréable', 0.9224386215209961),
('professionnel', 0.9213786721229553),
('discret', 0.9179787635803223)]

> model.wv.doesnt_match(['bon', 'sympathique',
'déçu', 'agréable'])
'déçu'
```

Nous observons dans ces exemples que les adjectifs *délicieux* et *succulent* sont rapprochés dans nos données, avec un score de similarité de 0,93. Le modèle propose également des mots sémantiquement proches (*attentif*, *accueillant*, ...) pour un mot donné (*sympathique*). En outre, Word2vec peut retrouver des relations d'analogie entre les mots, comme le fameux exemple : « king » - « man » + « woman » = « queen » (Mikolov *et al.* 2013a). De même, face à une relation « délicieux » - « plat » + « service » dans nos données, des mots tels que *sympathique*, *souriant* et *sympa* sont proposés, les associant ainsi à l'aspect du service. Enfin, nous pouvons également identifier les mots qui ne sont pas étroitement liés. Ici, il semble que le modèle ait compris que *bon*, *sympathique* et *agréable* sont associés, probablement par le biais de leur polarité positive, alors que *déçu* est éloigné de ceux-ci puisqu'il est lié à la polarité négative.

Par ailleurs, il existe des représentations Word2vec déjà entraînées sur des données massives et disponibles en ligne, notamment pour l'anglais. Grâce aux modèles pré-entraînés, il n'est pas nécessaire de chercher un corpus de grande taille, d'une part, et, d'autre part, le temps d'entraînement est économisé. Cependant, la représentation pré-entraînée peut ne pas être appropriée pour tous les types de corpus, car la nature des données doit être également prise en compte. Prenons par exemple deux représentations Word2vec entraînées sur le corpus frWiki (Wikipédia) par Fauconnier (2015). Le premier modèle est de type CBOW à 700 dimensions, et le second de type *skip-gram* à 1 000 dimensions. Les résultats de quelques mots de notre corpus sont présentés dans les pages qui suivent :

[modèle 1 : CBOW]

```
> model.wv.similarity('délicieux', 'succulent')
0.04999870899649351

> model.wv.most_similar('sympathique')
[('excellent', 0.5289491415023804),
('bravo', 0.4202755391597748),
('mauvais', 0.4147752523422241),
('continuation', 0.4032968580722809),
('je', 0.4022369980812073),
('voilà', 0.3956649899482727),
('sympa', 0.39212387800216675),
('espérer', 0.39003652334213257),
('boulot', 0.3858136534690857),
('ça', 0.38267141580581665)]

> model.wv.most_similar(positive=['service',
'délicieux'], negative=['plat'])
[('renseignement', 0.28794270753860474),
('prestataire', 0.21320557594299316),
('gaieté', 0.20634810626506805),
('fourniture', 0.19591951370239258),
('surveillance', 0.194951593875885),
('voyageur', 0.1948515921831131),
('bienfaisant', 0.19447195529937744),
('fournisseur', 0.19438087940216064),
('opérationnel', 0.19044512510299683),
('charger', 0.18514513969421387)]

> model.wv.doesnt_match(['bon', 'sympathique',
'déçu', 'agréable'])
'bon'
```

[modèle 2 : *skip-gram*]

```
> model.wv.similarity('délicieux', 'succulent')
0.20180939616189358

> model.wv.most_similar('sympathique')
[('attachant', 0.43631160259246826),
('jovial', 0.434454083442688),
('timide', 0.4242999851703644),
('serviable', 0.4227880835533142),
('sérieux', 0.4165791869163513),
('nerveux', 0.4146566390991211),
('prétentieux', 0.4126540422439575),
('gentil', 0.40739548206329346),
('humeur', 0.4062051773071289),
('amusant', 0.40524473786354065)]

> model.wv.most_similar(positive=['service',
'délicieux'], negative=['plat'])
[('gérer', 0.2977357506752014),
('réseau', 0.2743857204914093),
('public', 0.2554211914539337),
('prestataire', 0.25289666652679443),
('clientèle', 0.25243550539016724),
('offre', 0.24943317472934723),
('client', 0.2481207251548767),
('renseignement', 0.2467653751373291),
('infrastructure', 0.24324527382850647),
('disponibilité', 0.23766371607780457)]

> model.wv.doesnt_match(['bon', 'sympathique',
'déçu', 'agréable'])
'sympathique''
```

Les scores sont en général très faibles pour les exemples de ces deux modèles entraînés sur frWiki. Ces derniers n'ont pas perçu la similarité entre *délicieux* et *succulent*, qui représentent un score bien inférieur à 0,5, tandis que notre propre modèle a produit un score de 0,9334. Par ailleurs, en ce qui concerne les mots similaires à *sympathique*, les deux modèles comprennent de nombreux mots ayant très peu de liens avec cet adjectif, tels que *mauvais*, *continuation*, *je*, *boulot*, *ça*, *nerveux* et *humeur*. De même, les relations entre les mots *service*, *délicieux* et *plat*, ou *bon*, *sympathique*, *déçu* et *agréable* n'ont pas été saisies.

La nature de données peut expliquer ce résultat : le registre de langue utilisé sur le Wikipédia est relativement soutenu, alors que nos données relèvent d'un registre familier, avec des caractéristiques de l'oral et de la communication numérique. Le choix s'est porté alors sur l'utilisation de modèles entraînés directement sur le corpus traité qui permettent de prendre en compte ses spécificités linguistiques.

5.3.2.3 Traits retenus lors de l'observation manuelle

En complément des données textuelles, les traits observés et retenus lors de l'analyse de notre corpus (voir chapitres 3 et 4) ont été intégrés à l'expérience. Les traits ont été extraits de manière automatique ; il s'agit des éléments suivants :

- les **catégories morphosyntaxiques** : toutes les catégories morphosyntaxiques disponibles sur l'étiqueteur Stanford CoreNLP (nom, déterminant, adjectif, adverbe, préposition, adverbe, verbe, pronom, auxiliaire, conjonction de coordination, chiffre, conjonction de subordination, particule, interjection) et, en particulier, les pronoms personnels ou possessifs et les adjectifs possessifs de la deuxième personne (*vous*, *vôtre* et *votre*) ;

- les **diverses formes verbales** : puisque l'étiqueteur Stanford CoreNLP se limite à la simple catégories du verbe, des catégories plus détaillées ont été extraites au moyen de l'étiqueteur TreeTagger (Schmid 1994), telles que le conditionnel, le subjonctif, le futur, l'infinitif, l'impératif. De plus, les verbes avec le préfixe *re-*, indiquant l'itération d'une action ont été extraits également à l'aide des expressions régulières ;

- les **modifieurs** : les éléments qui changent la valeur initiale d'un mot en agissant directement sur l'orientation de la polarité, tels que la négation (*ne*, *jamais*, *aucun*, etc.) et la conjonction *mais* ;

- la **polarité** : le score de polarité, le score de subjectivité, la fréquence des mots positifs et négatifs ont été obtenus grâce à l'utilisation de la bibliothèque TextBlob[59] et les émoticônes ;

- la **position** : la position de la phrase dans un avis. Il s'agit d'une variable catégorielle qui prend pour valeur F (première phrase), L (dernière phrase), M (phrase au milieu) ou S (un avis composé d'une seule phrase) ;

- la **forme de surface** : le nombre de caractères, la longueur de la phrase, la diversité[60] et la densité des mots[61], les ponctuations multiples (⁉) et les majuscules en cascade (*EXCELLENT*), ainsi que l'unité monétaire (*euro* et €).

Nous avons ainsi obtenu un total de 35 traits. La prise en compte de tous ces traits peut accroître la complexité du modèle, ce qui augmente le risque de sur-apprentissage. Pour cette raison, nous avons tenté de diminuer le nombre de traits en conservant seulement les plus pertinents, ce qui a abouti à des modèles plus simples et plus généralisables. Nous avons alors appliqué la fonction `SelectPercentile` de scikit-learn, avec un centile de 90[62], qui permet d'évaluer la portée de chaque trait individuellement. Plus précisément, cette fonction sélectionne les meilleures caractéristiques en se fondant sur des mesures statistiques, permettant de réduire leur dimension en éliminant celles qui ne sont pas informatives. Les traits ayant une valeur de p élevée – autrement dit, ceux qui ne sont pas reliés étroitement – sont supprimés. Le tableau 5.1 montre ainsi le classement des traits par importance[63].

[59] https://github.com/sloria/textblob-fr [consulté le 18 septembre 2020].
[60] Le calcule a été effectué grâce au module proposé sur : https://github.com/kristopherkyle/lexical_diversity [consulté le 18 septembre 2020].
[61] Le nombre de tokens uniques divisé par le nombre total de mots http://michellejm.github.io/GCDRB_Text_Analysis/ [consulté le 18 septembre 2020].
[62] Le centile a été choisi sur la base de plusieurs expériences réalisées.
[63] re:VER (verbes avec le préfixe *re-*), VER:futu (verbes en futur), mais (conjonction

TAB. 5.1 : Classement des traits par importance

Trait	Score	Valeur P
re:VER	4147,40868	0
VER:futu	1066,86591	0
mais	880,404079	0
polarity_score	673,435149	0
VERB	332,329052	0,00e+00
count_positive	303,806417	1,59e-306
ADV	284,679443	3,87e-288
ADJ	228,063071	6,25e-233
sent_length	201,74477	7,39e-207
VER:impf	198,415475	1,53e-203
NUM	188,38383	1,60e-193
count_negative	184,965058	4,26e-190
char_count	178,178991	2,74e-183
negation	159,893953	7,46e-165
CCONJ	148,285227	4,40e-153
PRON	139,302392	6,13e-144
NOUN	138,090845	1,06e-142
DET	123,463105	9,69e-128
subjectivity_score	104,727488	1,86e-108
ADP	102,994931	1,15e-106
AUX	98,7265632	2,98e-102
VER:cond	94,3448736	1,03e-97
euro	87,3376298	1,93e-90
VER:infi	76,9535589	1,24e-79
SCONJ	72,6591426	3,75e-75
lex_diversity	70,4562138	7,49e-73
vous	70,045478	2,01e-72
lex_density	56,2056007	6,20e-58
multi_punc	25,3144865	1,76e-25
VER:subp	21,9543633	5,85e-22
position	19,9086491	8,06e-20
PART	7,76635007	2,64e-07
INTJ	4,05660519	1,11e-03
emoticon	3,6847561	2,47e-03
majuscule	0,42795404	8,29e-01

Selon les expériences menées à l'aide de la fonction `SelectPercentile`, sur 35 traits, quatre traits ont été jugé peu importants : les particules, les interjections, les émoticônes et les majuscules en cascade. Ainsi, les 31 traits restants ont été retenus pour l'apprentissage de surface, qui sera présenté dans la partie suivante.

5.4 Classification supervisée

Afin de détecter les six catégories d'évaluation proposées, nous avons exploité deux techniques d'apprentissage supervisé : l'apprentissage de surface et l'apprentissage profond.

5.4.1 *Apprentissage de surface*

L'apprentissage se déroule en deux phases : l'entraînement, puis la prédiction, comme le montre la figure 5.4. Durant la phase d'entraînement, les données et les étiquettes sont utilisées en tant qu'entrée et sortie, respectivement. Les données textuelles et les traits retenus sont transformés en vecteurs de caractéristiques (*features*). Un modèle d'apprentissage automatique est ensuite construit, auquel est fourni le jeu d'apprentissage. Le classifieur cherche par la suite une relation mathématique entre les caractéristiques et leurs étiquettes, permettant, pour une entrée donnée, de générer la sortie associée. Lors de la phase de prédiction, le classifieur est ensuite capable d'attribuer une étiquette à des échantillons qu'il n'a jamais rencontrés auparavant.

---

mais), polarity_score (scores de polarité), VERB (verbes), count_positive (fréquence des mots positifs), ADV (adverbes), ADJ (adjectifs), sent_length (longueur de la phrase), VER:impf (verbes impératif), NUM (chiffre), count_negative (fréquence des mots négatifs), char_count (nombre de caractères), negation (négation), CCONJ (conjonctions de coordination), PRON (pronoms), NOUN (noms), DET (déterminants), subjectivity_score (scores de subjectivité), ADP (prépositions), AUX (auxiliaires), VER:cond (verbes au conditionnel), euro (unité monétaire), VER:infi (verbes en infinitif), SCONJ (conjonctions de subordination), lex_diversity (diversité des mots), vous (pronom *vous*), lex_density (densité des mots), multi_punc (ponctuations multiples), VER:subp (verbes au subjonctif), position (position dans un avis), PART (particules), INTJ (interjection), emoticon (émoticônes), majuscule (majuscules en cascade).

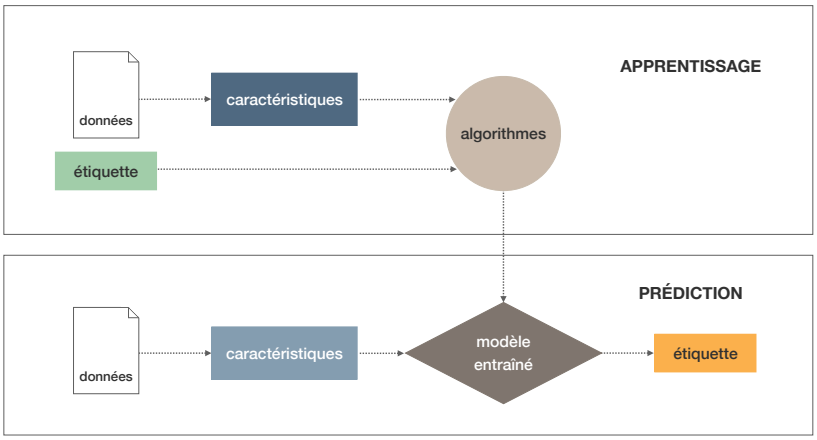


FIG. 5.4 : Processus d'apprentissage de surface

Les classifieurs courants de l'apprentissage de surface visant à classifier les textes sont les modèles linéaires, le modèle génératif et les modèles ensemblistes (Aggarwal et Zhai 2012 ; Bhatia *et al.* 2020 ; Khan *et al.* 2010). Les modèles linéaires visent à chercher un hyperplan qui permet de classifier les échantillons aux propriétés similaires au sein de la même dimension (Logistic Regression et Support Vector Machine). Le modèle génératif repose sur le théorème de Bayes, qui calcule la probabilité conditionnelle de l'occurrence de plusieurs événements en fonction de la probabilité d'occurrence de chaque événement individuel (Naïve Bayes). Les modèles ensemblistes combinent divers modèles d'apprentissage automatique, produisant des modèles plus puissants (Random Forest et XGBoost). Afin de les exploiter, nous avons utilisé la bibliothèque scikit-learn (Pedregosa *et al.* 2011) et la bibliothèque XGBoost[64] (Chen et Guestrin 2016).

5.4.2 *Apprentissage profond*

L'apprentissage profond s'est récemment révélé très prometteur concernant les applications en TAL. Sa puissance provient de sa structure en réseau hiérarchique ou multicouche, qui lui permet d'apprendre la

[64] https://xgboost.readthedocs.io/en/latest/ [consulté le 18 septembre 2020].

représentation des données et d'effectuer des prédictions.

Le réseau de neurones est constitué de trois types de couches : la couche d'entrée, la couche cachée et la couche de sortie (figure 5.5). La couche d'entrée sert à introduire les données dans le réseau, où elles sont représentées sous forme de vecteurs. Elle est suivie d'une ou de plusieurs couches cachées, dans lesquelles les données interagissent avec les poids de connexion et le choix de la fonction d'activation. Enfin, la couche de sortie fait les prédictions, dont la fonction d'activation est choisie selon la configuration des réseaux de neurones. Lorsqu'il s'agit d'un problème de classification, une fonction d'activation de type softmax (pour les multi-classes) est utilisée.

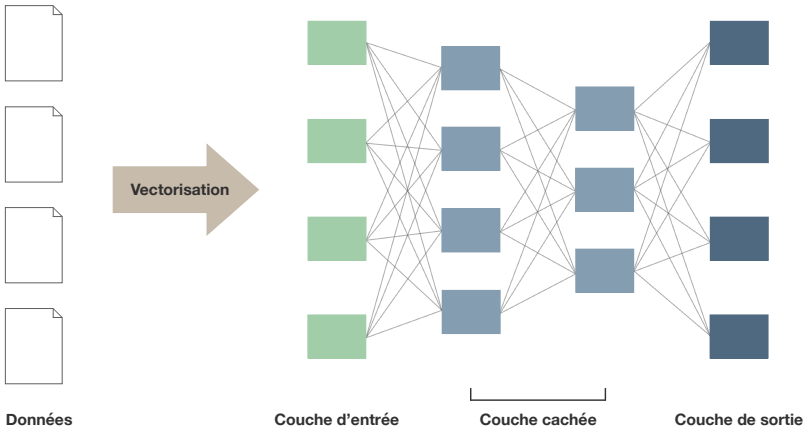


FIG. 5.5 : Processus d'apprentissage profond

Dans le cadre de cette étude, nous avons développé des modèles basés sur les réseaux de neurones convolutifs (*convolutional neural network* ou CNN) et les réseaux récurrents à mémoire court et long terme (*long short-term memory* ou LSTM). Nous présenterons ici ces deux architectures de manière synthétique. Le point fort des CNN réside dans leur capacité à extraire les sous-structures des données utiles pour la prédiction globale. Ainsi, grâce à la prise en compte des mots enchaînés, ils sont fréquemment utilisés dans la classification des textes. Les LSTM disposent d'une cellule de mémoire (*cell state*) composée de trois portes (la porte d'entrée, la

porte d'oubli et la porte de sortie) qui régulent le flot d'informations. Grâce au dispositif de ces portes, les LSTM sont capables d'oublier les tokens les moins importants et de traiter de longues séquences de mots.

Les réseaux de neurones profonds ont été construits à l'aide de la bibliothèque Keras (Chollet *et al.* 2015). Afin de construire une simple architecture CNN, nous avons pris pour entrée une matrice d'*embedding* à l'aide d'un modèle Word2vec de type CBOW. Nous avons utilisé un filtre de convolution de 16 neurones ainsi qu'un filtre de taille 3 suivi par la fonction d'activation `ReLU` (unité de rectification linéaire). Une couche de `MaxPooling1D` a ensuite été appliquée sur la sortie de la couche de convolution, divisant par deux la taille de la sortie de la couche précédente. Enfin, nous avons effectué l'aplatissement et la réduction de la dimension, avant d'exécuter une activation `softmax` sur la couche finale composée de six neurones. Un réseau LSTM a été élaboré, utilisant la matrice d'*embeddings* suivie d'une couche LSTM avec 100 unités, en y appliquant un dropout avec une probabilité de 0,2 et un dropout récurrent de 0,2. Une fonction d'activation softmax a été appliquée sur la couche finale composée de six neurones. Pour les deux architectures, la taille du batch a été déterminée à l'aide de `GridSearch` et le nombre d'époques par la fonction `EarlyStopping`.

La différence entre l'apprentissage de surface et l'apprentissage profond réside dans les caractéristiques. Le premier exige plus de temps et d'efforts, car ce sont les humains qui créent manuellement les caractéristiques, alors que pour le second, les caractéristiques sont apprises de manière automatique, l'intervention humaine n'étant donc pas nécessaire. L'autre particularité qui distingue l'apprentissage profond de l'apprentissage de surface est sa capacité à mémoriser certaines informations. A contrario, l'apprentissage de surface ne possède pas cette capacité et ne peut donc pas être rattaché aux informations passées (Bhatia *et al.* 2020).

5.4.3 *Validation croisée*

Dans le but d'évaluer la capacité de généralisation d'un modèle, c'est-à-dire de demander aux modèles de refaire ses prédictions sur de nouvelles données jamais traitées, la validation croisée a été appliquée. Cependant, lorsque nous disposons de données déséquilibrées, la validation croisée

n'est pas toujours suffisante pour l'évaluation du modèle. Étant donné que la répartition des classes au sein de nos données est disparate, il est probable que certains plis ne fournissent que très peu ou pas d'observation de la classe minoritaire. Pour gérer ce problème, nous avons donc effectué une validation croisée stratifiée qui permet de produire des plis contenant un ratio représentatif de chaque classe. Grâce à celle-ci, chaque jeu de test généré comporte des proportions de classes à peu près identiques à celle du jeu de données dans son ensemble. Au cours de toutes les expériences, la validation croisée stratifiée à cinq plis a été appliquée.

5.5 Traitement des données déséquilibrées

Les modèles développés dans les bibliothèques supposent une répartition équilibrée des classes. Cela signifie qu'une application naïve de ces modèles peut mener à l'apprentissage des caractéristiques de la classe majoritaire (celle disposant du plus grand nombre d'exemples) uniquement. La classe minoritaire est plus difficile à prédire parce qu'elle contient peu d'exemples. De ce fait, le grand nombre d'instances de la classe majoritaire submerge la classe minoritaire, et l'apprentissage devient biaisé en faveur de la classe majoritaire. Ainsi, lorsque nous disposons des données déséquilibrées, il faut non seulement choisir les bonnes métriques d'évaluation, mais également prendre des mesures pertinentes pour l'apprentissage. Dans cette étude, nous avons exploité, d'une part, la technique Cost-Sensitive Learning (CSL) qui consiste à pénaliser les observations dont la prédiction est incorrecte et, d'autre part, les techniques d'échantillonnage des données qui transforment l'ensemble des données d'apprentissage en vue d'équilibrer la répartition des classes.

5.5.1 *Cost-Sensitive Learning (CSL)*

Hormis l'hypothèse d'une distribution des classes équilibrée, la plupart des classifieurs présument également que les erreurs de prédiction faites par un classifieur possèdent toutes la même valeur. Cependant, en pratique, ce n'est généralement pas le cas ; nous disposons souvent de données ayant une distribution des classes déséquilibrée. La technique CSL pénalise les prédictions erronées en leur attribuant des coûts : son objectif n'est

donc pas d'optimiser l'exactitude de la détection, mais de minimiser les erreurs de prédiction.

Afin d'adapter CSL, nous avons ajusté les paramètres nécessaires pour chaque bibliothèque et algorithme (tableau 5.2) :

Tab. 5.2 : Paramètres pour adapter CSL

Bibliothèque	Algorithmes	Paramètre	Valeur du paramètre
scikit-learn	Naïve Bayes, Logistic Regression, Support Vector Machine	`class_weight`	`'balanced'`
	Random Forest		`'balanced_subsample'`
XGBoost	XGBoost	`scale_pos_weight`	les poids de chaque classe*
Keras	CNN, LSTM	`class_weight`	

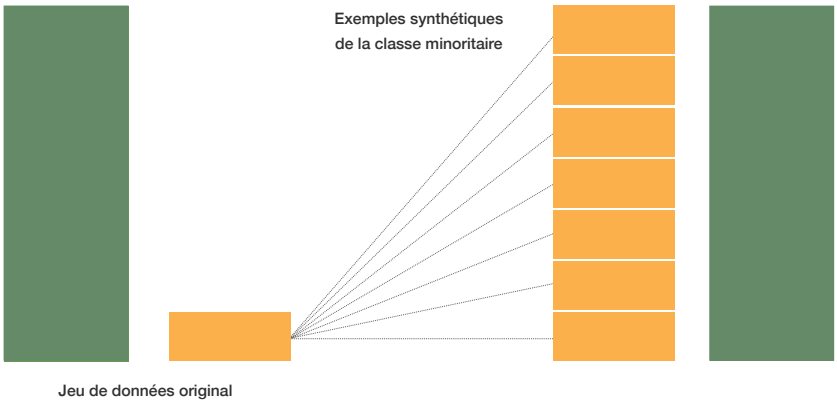
* Obtenus grâce à la fonction `compute_class_weight` de la bibliothèque scikit-learn

5.5.2 *Sur-échantillonnage*

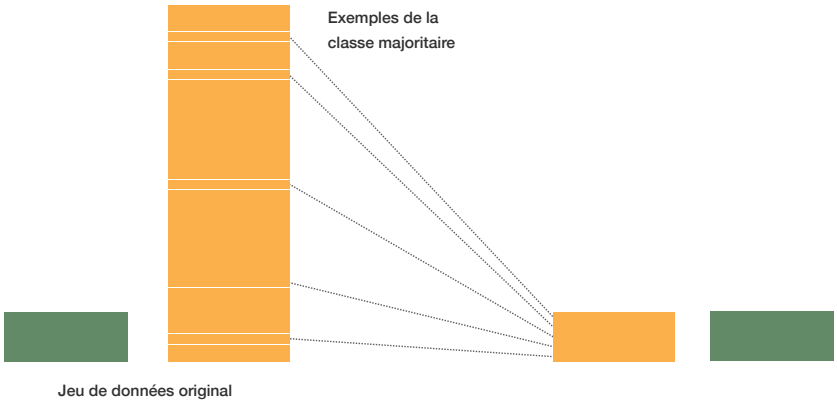
L'échantillonnage des données consiste à modifier la répartition des classes de l'ensemble de données, permettant ainsi de résoudre le problème du déséquilibre. Il existe deux types d'échantillonnage : le sur-échantillonnage (figure 5.6a) et le sous-échantillonnage (figure 5.6b). Le premier reproduit de nouvelles instances issues de la classe minoritaire, alors que le second supprime les instances de la classe majoritaire. Nous avons préféré utiliser le sur-échantillonnage, puisque la classe minoritaire est très peu représentée et qu'il est inenvisageable de supprimer les instances de la classe majoritaire (opinion positive).

Les techniques de sur-échantillonnage ont été utilisées grâce à la bibliothèque Imbalanced-learn[65] (Lemaître *et al.* 2017). Les différentes techniques de sur-échantillonnage, Random Oversampling et Synthetic Minority Oversampling Technique (SMOTE) ont été exploitées. La première est qualifiée de naïve, car elle ne fait aucune supposition sur les données.

[65] https://imbalanced-learn.readthedocs.io/en/stable/ [consulté le 8 novembre 2020].



(a) Sur-échantillonnage



(b) Sous-échantillonnage

FIG. 5.6 : Sur-échantillonnage et sous-échantillonnage

De ce fait, l'approche est simple et rapide à implémenter, ce qui est un avantage lorsqu'il s'agit de données très vastes et complexes. Cependant, cette approche peut inciter au sur-apprentissage, car elle reproduit des instances exactement identiques à ceux de la classe minoritaire. La seconde technique consiste à créer des exemples synthétiques, proches des points de données dans l'espace des caractéristiques. Elle est efficace, car les nouveaux exemples créés sont semblables à ceux faisant réellement partie de la classe minoritaire. Le sur-échantillonnage n'a été effectué que sur le jeu d'apprentissage.

Quant à l'apprentissage de surface, les deux techniques (Random Over-sampling et SMOTE) ont été exploitées, alors que seule la première technique a été utilisée pour l'apprentissage profond, en raison de la survenue d'un problème d'incompatibilité entre les deux bibliothèques (Keras, Imbalanced-learn).

Notre objectif n'est pas d'obtenir le meilleur score possible, mais plutôt d'avoir de bonnes performances aussi bien sur les classes minoritaires que sur les classes majoritaires. Nous avons ainsi appliqué la technique de CSL et celles du sur-échantillonnage. Or, nous ne considérons pas ce dernier comme la meilleure solution, car les nouvelles instances reproduites ne sont pas des données authentiques. De ce fait, nous privilégions les résultats obtenus par le biais de CSL, qui pénalise la classe majoritaire tout en maintenant les données en tant que telles. Ainsi, la technique CSL a été exploitée au cours des expériences, tandis que le sur-échantillonnage a été utilisé uniquement afin d'observer son impact sur la performance.

5.6 Résultats

L'évaluation a été effectuée sur le jeu de test, représentant 33 % des données de manière stratifiée. La métrique d'évaluation utilisée dans cette étude sera présentée dans la section 5.6.1, suivie des résultats des expériences réalisées (les sections 5.6.2 et 5.6.3). Nous proposerons ensuite une observation de l'application du sur-échantillonnage dans la section 5.6.4.

5.6.1 *Métrique d'évaluation*

L'exactitude, c'est-à-dire la proportion des échantillons correctement classés, est une métrique fréquemment utilisée dans l'évaluation des modèles, car facile à calculer. Toutefois, elle ne permet pas d'évaluer correctement les expériences à distribution des classes déséquilibrée. Par exemple, si les données contiennent une observation pour la classe A (classe minoritaire) et 99 observations pour la classe B (classe majoritaire), un modèle prédisant la classe B pour toute observation aura une exactitude de 99 %, ce qui conduit à une conclusion erronée.

Ainsi, nous avons évalué la performance en termes de précision, rappel, F-mesure (ou F-score) et matrice de confusion. La précision indique l'exactitude des prédictions positives, c'est-à-dire qu'elle mesure le taux de prédictions correctes parmi les observations prédites comme positives. Elle est utilisée lorsque le but est de limiter le nombre de faux positifs. Le rappel correspond à la proportion d'échantillons positifs prédits comme positifs. Afin de trouver un compromis entre la précision et le rappel, nous pouvons combiner les deux en une métrique unique, leur moyenne harmonique : la F-mesure. Celle-ci peut être (i) une moyenne micro, qui calcule la moyenne en se basant sur toutes les observations, (ii) une moyenne macro, donnant un poids identique à chaque catégorie sans tenir compte de leur taille, ou (iii) une moyenne pondérée, qui prend en considération la répartition déséquilibrée des classes[66]. Étant donné que la répartition de nos données est asymétrique, nous avons utilisé la moyenne pondérée de la F-mesure sur le jeu de test. La matrice de confusion donne un meilleur aperçu de la performance, mais aussi des classes qui sont prédites correctement ou non et du type d'erreurs qui apparaît fréquemment.

5.6.2 *Apprentissage de surface*

Les expériences exploitant les techniques de l'apprentissage de surface ont été menées en deux étapes : d'abord, seules les données textuelles ont été utilisées, puis les traits retenus lors de l'observation des données y ont

[66] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html [consulté le le 19 octobre 2020].

été ajoutés. Les performances ont été évaluées en fonction de différents classifieurs, tels que Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine linéaire (SVM), Random Forest (RF) et XGBoost (XG). Seul le classifieur Naïve Bayes n'a pas donné de résultats sur l'ensemble des traits linguistiques, à cause de son incapacité à traiter les chiffres négatifs.

Pour les premières expériences réalisées, prenant en compte uniquement les données textuelles, deux versions des données ont été testées (tableau 5.3) : une version des données lemmatisées comprenant les mots-outils (*lemme*), et une autre version dans laquelle les données ont été lemmatisées, mais les mots-outils supprimés (*lemme-stp*). Ces différentes versions des données permettent d'observer le comportement des mots-outils. Pour chacune de ces versions, la meilleure combinaison des méthodes de vectorisation (*Count* contre *TF-IDF*) et la gamme des n-grammes a été proposée par `GridSearch`. Le tableau 5.3 montre ainsi une comparaison des résultats obtenus : pour chaque classifieur, la meilleure F-mesure (pondérée) est présentée en caractères gras.

TAB. 5.3 : Moyenne pondérée de la F-mesure (sac de mots/TF-IDF + apprentissage de surface)

Algorithmes	Corpus	Vectorisation	N-grammes	F-mesure	Différence
NB	lemme	TF-IDF	(1,3)	**0,8222**	0,008
	lemme-stp	Count	(1,3)	0,8141	
LR	lemme	TF-IDF	(1,2)	**0,8562**	0,003
	lemme-stp	TF-IDF	(1,2)	0,8528	
SVM	lemme	TF-IDF	(1,2)	**0,8659**	0,003
	lemme-stp	Count	(1,2)	0,8633	
RF	lemme	Count	(1,1)	0,7812	0,039
	lemme-stp	TF-IDF	(1,1)	**0,8199**	
XG	lemme	TF-IDF	(1,3)	0,8091	0,017
	lemme-stp	Count	(1,3)	**0,8262**	

Les résultats obtenus sur les données textuelles seules, sans prise en compte de traits linguistiques, sont assez satisfaisants. Notons également qu'il y a très peu de différences entre les F-mesures des deux versions du corpus. Quel que soit le corpus utilisé, il semble que les meilleurs résultats

soient obtenus par le contrôle des mots-outils grâce aux méthodes de vectorisation : le corpus lemme a tendance à pénaliser les mots-outils par la méthode TF-IDF, alors que le corpus lemme-stp, dans lequel les mots-outils sont déjà supprimés, ne nécessite plus de traitement supplémentaire.

La représentation Word2vec a également été testée suivant deux architectures différentes : CBOW et *skip-gram*. Nous avons jugé pertinent d'exploiter les données lemmatisées conservant les mots-outils, en raison du principe de Word2vec selon lequel les propriétés grammaticales et sémantiques sont encodées en considérant le contexte dans lequel un mot est apparu. Le tableau 5.4 montre les moyennes pondérées des F-mesures des expériences. En règle générale, la représentation Word2vec mène à de mauvaises performances, avec une F-mesure de 0,7839 en moyenne. Pour cette raison, cette représentation n'a pas été utilisée dans le reste des expériences de l'apprentissage de surface.

TAB. 5.4 : Moyenne pondérée de la F-mesure (Word2vec + apprentissage de surface)

Algorithmes	**Architecture**	**F-mesure**
LR	CBOW	0,7815
	skip-gram	0,7939
SVM	CBOW	0,7873
	skip-gram	0,7869
RF	CBOW	0,7703
	skip-gram	0,7775
XG	CBOW	0,7847
	skip-gram	0,7889

Nous avons ensuite réalisé les expériences en intégrant les 33 traits retenus précédemment (section 5.3.2.3). Ce processus révèle dans quelle mesure les traits linguistiques peuvent améliorer ou dégrader la performance des expériences. La configuration de la représentation des données textuelles (la version du corpus, la méthode de vectorisation, la gamme des n-grammes) a été ajustée sur la base des meilleurs résultats obtenus précédemment. Les valeurs des traits étant numériques, celles-ci ont été

normalisées afin que toutes les valeurs soient à la même échelle, en utilisant `RobustScaler` de scikit-learn. Le tableau 5.5 présente ainsi une synthèse des F-mesures obtenues en fonction de la manière dont les données ont été utilisées : i) les données textuelles seules dans la colonne DT et ii) la combinaison des données textuelles et des traits retenus dans la colonne DT+TL. Afin de mesurer la variation de la F-mesure entre ces deux processus, le taux d'évolution a été calculé – indiqué dans la colonne Taux (%). Le classifieur NB n'affiche pas de résultat dans la colonne DT+TL, puisqu'il ne peut pas traiter les chiffres négatifs présents dans les caractéristiques.

Dans le tableau 5.5, nous constatons qu'à l'exception du modèle RF, les classifieurs ont vu leur performance s'améliorer légèrement (notamment de 5,29 % pour le classifieur XG) avec l'ajout des traits retenus. Le classifieur SVM donne le meilleur score de F-mesure (0,8818), suivi de XG (0,8699) et LR (0,8627).

Tab. 5.5 : Moyenne pondérée de la F-mesure (apprentissage de surface)

Algorithmes	**DT**	**DT+TL**	**Taux (%)**
NB	0,8222	-	-
LR	0,8562	0,8627	0,76
SVM	0,8659	0,8818	1,84
RF	0,8145	0,8076	-0,85
XG	0,8262	0,8699	5,29

Nous avons également examiné la F-mesure, la précision et le rappel de chaque catégorie de notre meilleur classifieur (SVM), comme représenté dans le tableau 5.6. La détection de l'opinion positive obtient la meilleure performance (la F-mesure étant de 0,94), et l'intention et la suggestion obtiennent des F-mesures correspondant à 0,88 et 0,84 respectivement. En revanche, la détection de la description connaît la plus mauvaise performance (0,41), ayant l'écart le plus important (1,0) entre la précision et le rappel. Ce résultat est dû à la faible fréquence de la description dans le corpus ; elle semble donc s'appuyer sur peu de marqueurs lexicaux en raison de la nature hétérogène des informations qui y sont présentes.

L'opinion mixte semble également engendrer des difficultés aux classifieurs. En effet, l'observation manuelle du corpus montre que le changement de polarité est souvent marqué par la conjonction *mais*, comme dans [3]. Étant donné que 57,5 % des opinions mixtes incluent cette conjonction, la segmentation du corpus en fonction de sa présence pourrait améliorer la détection des opinions mixtes. Par exemple, la segmentation de la phrase [3] en deux propositions permet de classer la première proposition (*Le restaurant était complet et très bruyant*) comme opinion négative et la seconde comme opinion positive (*la cuisine excellente*). Intégrer l'approche de Lark (2017), qui emploie les règles lexicosyntaxiques appliquées aux connecteurs, semble résoudre ce problème.

TAB. 5.6 : Précision, rappel et F-mesure de chaque catégorie (SVM)

	POS_OPINION	NEG_OPINION	MIX_OPINION	SUGGESTION	INTENTION	DESCRIPTION
Précision	0,94	0,69	0,70	0,88	0,88	0,46
Rappel	0,94	0,76	0,68	0,82	0,89	0,36
F-mesure	0,94	0,72	0,69	0,84	0,88	0,41

[3] *Le restaurant était complet et très bruyant* ***mais*** *la cuisine excellente.*

Enfin, nous avons représenté le résultat de notre meilleur classifieur (SVM) sous la forme d'une matrice de confusion normalisée, qui offre une vision globale de la performance (figure 5.7). Chaque ligne correspond à la classe réelle et chaque colonne à la classe prédite. Les cellules de la diagonale principale indiquent les classes qui ont été classifiées correctement. Parmi elles, la cellule de la description est légèrement plus claire que les autres, ce qui signifie que cette catégorie a été plus difficile à détecter. Ce résultat s'explique en partie par le manque d'échantillons des descriptions et donc par le faible nombre de caractéristiques fournies durant l'entraînement. Par ailleurs, cette catégorie est très hétérogène et varie en fonction du profil du client, ce qui cause l'apparition d'un large éventail de vocabulaires et de contextes. Par conséquent, la description

a tendance à être rangée dans la classe majoritaire, c'est-à-dire l'opinion positive (0,37), et occasionnellement dans l'opinion négative (0,18). Nous observons une tendance similaire pour l'opinion mixte, dont le mauvais score (0,68) vient du fait que la catégorie implique à la fois l'opinion positive (0,16) et l'opinion négative (0,13).



FIG. 5.7 : Matrice de confusion normalisée (SVM)

5.6.3 *Apprentissage profond*

L'apprentissage profond ne nécessitant pas d'intervention humaine en ce qui concerne les caractéristiques fournies, les expériences n'ont été réalisées que sur les données textuelles. Nous avons ainsi testé deux différentes architectures d'apprentissage profond : CNN et LSTM. La performance de cette approche est illustrée dans le tableau 5.7. L'architecture LSTM donne le meilleur résultat avec une moyenne pondérée des F-mesures égale à 0,8497, très légèrement supérieure à celle de CNN (0,8429). Par ailleurs, en comparant le tableau 5.7 avec le tableau 5.5, nous remarquons que le

meilleur résultat de l'apprentissage profond (LSTM) est inférieur à celui de l'apprentissage de surface (SVM) – 0,8818.

TAB. 5.7 : Moyenne pondérée de la F-mesure (apprentissage profond)

Algorithmes	**F-mesure**
CNN	0,8429
LSTM	0,8497

Le tableau 5.8 présente la F-mesure, la précision et le rappel pour chaque catégorie de l'architecture LSTM. Les comportements de LSTM ont été très semblables à ceux de SVM, mais moins performants : l'opinion positive a produit de nouveau le meilleur résultat, avec une F-mesure de 0,92, suivie de l'intention et de la suggestion, qui ont obtenu respectivement les scores de 0,88 et de 0,77. La description a toujours eu la moins bonne performance (0,33) – plus mauvaise que celle de l'apprentissage de surface (0,41) –, avec le plus grand écart entre la précision et le rappel (0,15).

TAB. 5.8 : Précision, rappel et F-mesure de chaque catégorie (LSTM)

	POS_OPINION	NEG_OPINION	MIX_OPINION	SUGGESTION	INTENTION	DESCRIPTION
Précision	0,94	0,64	0,62	0,74	0,88	0,27
Rappel	0,90	0,65	0,66	0,81	0,89	0,42
F-mesure	0,92	0,64	0,64	0,77	0,88	0,33

Cette catégorie détient également la plus mauvaise performance dans la matrice de confusion (figure 5.8). Nous constatons à nouveau que SVM et LSTM produisent des matrices plus ou moins similaires : la détection de la description est toujours une tâche difficile. L'opinion mixte (0,64) et l'opinion négative (0,64) donnent des scores inférieurs à ceux de SVM.

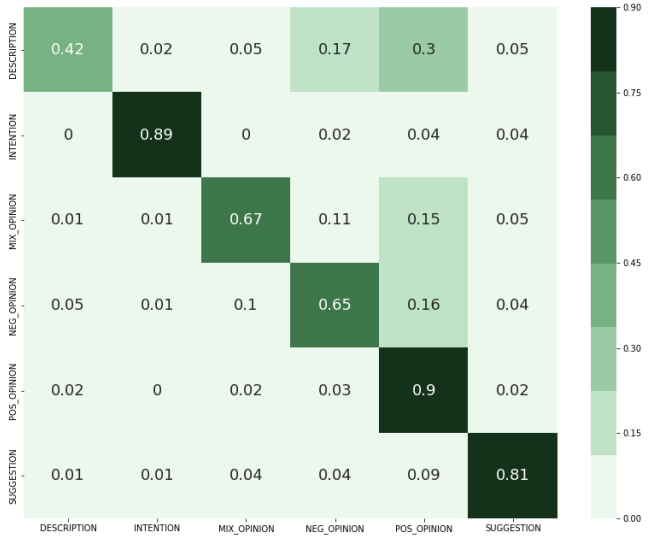


FIG. 5.8 : Matrice de confusion normalisée (LSTM)

5.6.4 *Impact du sur-échantillonnage*

Au cours des expériences réalisées précédemment, la technique CSL a été appliquée afin de résoudre le problème du déséquilibre des classes. Ici, les techniques de sur-apprentissage (Random Oversampling et SMOTE) ont été exploitées afin d'observer leur efficacité, en particulier sur les classes minoritaires. Quant à l'apprentissage profond, seule la technique du Random Oversampling a été employée, en raison d'un problème d'incompatibilité des fonctions entre les bibliothèques (Keras, Imbalanced-learn).

Afin d'observer l'impact du sur-échantillonnage, les expériences concernant l'apprentissage de surface ont été réalisées en deux étapes, comme précédemment. En premier lieu, seules les données textuelles ont été utilisées pour déterminer la meilleure combinaison de configuration (la méthode de représentation vectorielle, la version du corpus[67], les gammes

[67]Une version des données lemmatisées comprenant les mots-outils, et une autre version dans laquelle les données ont été lemmatisées, mais les mots-outils supprimés.

de n-grammes). Son résultat est présenté dans le tableau 5.10 ; pour chaque méthode d'échantillonnage d'un classifieur, la meilleure F-mesure (pondérée) est présentée en caractères gras.

Nous constatons de nouveau que la combinaison de configuration de ces données produit peu de différences sur le plan des performances. Par ailleurs, les traits retenus ont été intégrés aux expériences : les performances s'améliorent grâce à l'ajout des traits linguistiques (voir tableaux 5.10 et 5.9).

Une comparaison entre les performances obtenues en utilisant la technique CSL et celle du sur-échantillonnage est affichée dans le tableau 5.9, où la meilleure F-mesure est présentée en caractères gras. Notons que les F-mesures de sur-échantillonnage chutent généralement par rapport à celles de CSL, à l'exception du classifieur RF, qui a progressé de 3,62 %.

TAB. 5.9 : Moyenne pondérée de la F-mesure (apprentissage de surface) : CSL contre sur-échantillonnage

Algorithmes	**CSL**	**Échantillonnage**		**Taux (%)**
		Random	**SMOTE**	
NB	0,8222	**0,7605**	0,7553	-7,50
LR	0,8642	**0,8589**	0,8536	-0,17
SVM	0,8818	0,8652	**0,8658**	-1,88
RF	0,8076	0,8343	**0,8368**	3,62
XG	0,8699	0,8697	**0,8699**	0,00

TAB. 5.10 : Moyenne pondérée de la F-mesure selon différentes méthodes d'échantillonnage (sac de mots/TF-IDF et apprentissage de surface)

Algorithmes	Échantillonnage	Vectorisation	Corpus	N-grammes	F-mesure	Différence (valeur absolue)
NB	Random	Count	lemme	(1,2)	**0,7605**	0,0116
		Count	lemme-stpw	(1,2)	0,7489	
	SMOTE	TF-IDF	lemme	(1,2)	0,7282	0,0116
		Count	lemme-stpw	(1,3)	**0,7553**	
LR	Random	TF-IDF	lemme	(1,3)	**0,8508**	0,0271
		TF-IDF	lemme-stpw	(1,2)	0,8486	
	SMOTE	TF-IDF	lemme	(1,2)	**0,8445**	0,0022
		TF-IDF	lemme-stpw	(1,2)	0,8360	
SVM	Random	Count	lemme	(1,3)	**0,8591**	0,0085
		Count	lemme-stpw	(1,2)	0,8528	
	SMOTE	Count	lemme	(1,2)	**0,8559**	0,0063
		Count	lemme-stpw	(1,2)	0,8507	
RF	Random	TF-IDF	lemme	(1,2)	0,8131	0,0075
		TF-IDF	lemme-stpw	(1,1)	**0,8206**	
	SMOTE	TF-IDF	lemme	(1,2)	0,8058	0,0132
		Count	lemme-stpw	(1,3)	**0,8190**	
XG	Random	TF-IDF	lemme	(1,2)	**0,8416**	0,0015
		TF-IDF	lemme-stpw	(1,2)	0,8401	
	SMOTE	TF-IDF	lemme	(1,3)	**0,8540**	0,0145
		TF-IDF	lemme-stpw	(1,2)	0,8395	

En ce qui concerne l'apprentissage profond, nous avons pris en entrée une matrice d'*embedding* à l'aide de Word2vec entraîné précédemment. Les F-mesures des deux architectures (CNN et LSTM) ont légèrement augmenté grâce au sur-échantillonnage, la meilleure étant 0,8604 (LSTM) (tableau 5.11). Cette amélioration peut s'expliquer par le principe de l'apprentissage profond, qui nécessite généralement beaucoup plus de données pour fournir de meilleures prédictions (Domingos 2012). Ce score augmenté est, néanmoins, inférieur à notre meilleur classifieur au cours de cette étude, la technique CSL appliquée sur le classifieur SVM (0,8818).

TAB. 5.11 : Moyenne pondérée de la F-mesure (apprentissage profond) : CSL contre sur-échantillonnage

Algorithmes	**CSL**	**Échantillonnage**	**Taux (%)**
		Random	
CNN	0,8429	0,8530	1,21
LSTM	0,8497	0,8604	1,25

Afin d'observer la portée du sur-échantillonnage, notamment sur les classes minoritaires, nous comparons les performances de chaque catégorie obtenue pour les deux méthodes (CSL et sur-échantillonnage). Pour l'apprentissage de surface, nous examinons les performances du classifieur SVM (tableau 5.12), qui produit le meilleur résultat avec CSL, et du classifieur RF (tableau 5.13), car celui-ci a vu une plus forte amélioration (3,62 %). Pour l'apprentissage profond, nous observons LSTM, dans lequel les catégories ont été mieux détectées grâce au sur-échantillonnage (tableau 5.14).

Tab. 5.12 : Précision, rappel et F-mesure de chaque catégorie (SVM) : CSL contre SMOTE

	POS_OPINION		MIX_OPINION		NEG_OPINION		SUGGESTION		INTENTION		DESCRIPTION	
	CSL	**SMOTE**	**CSL**	**SMOTE**	**CSL**	**SMOTE**	**CSL**	**SMOTE**	**CSL**	**SMOTE**	**CSL**	**SMOTE**
Précision	0,94	0,93	0,71	0,70	0,69	0,67	0,87	0,82	0,88	0,89	0,48	0,37
Rappel	0,94	0,94	0,68	0,61	0,76	0,69	0,82	0,81	0,89	0,86	0,37	0,36
F-mesure	0,94	0,93	0,70	0,65	0,72	0,68	0,84	0,82	0,88	0,87	0,42	0,36

Tab. 5.13 : Précision, rappel et F-mesure de chaque catégorie (RF) : CSL contre SMOTE

	POS_OPINION		MIX_OPINION		NEG_OPINION		SUGGESTION		INTENTION		DESCRIPTION	
	CSL	**SMOTE**	**CSL**	**SMOTE**	**CSL**	**SMOTE**	**CSL**	**SMOTE**	**CSL**	**SMOTE**	**CSL**	**SMOTE**
Précision	0,95	0,91	0,52	0,63	0,49	0,60	0,76	0,80	0,84	0,88	0,21	0,42
Rappel	0,82	0,91	0,70	0,62	0,75	0,71	0,69	0,62	0,88	0,87	0,51	0,42
F-mesure	0,88	0,91	0,60	0,63	0,60	0,65	0,72	0,70	0,86	0,87	0,30	0,42

Tab. 5.14 : Précision, rappel et F-mesure de chaque catégorie (LSTM) : CSL contre Random

	POS_OPINION		MIX_OPINION		NEG_OPINION		SUGGESTION		INTENTION		DESCRIPTION	
	CSL	**Random**	**CSL**	**Random**	**CSL**	**Random**	**CSL**	**Random**	**CSL**	**Random**	**CSL**	**Random**
Précision	0,94	0,92	0,62	0,71	0,64	0,63	0,74	0,84	0,88	0,92	0,27	0,31
Rappel	0,90	0,92	0,66	0,65	0,65	0,67	0,81	0,79	0,89	0,90	0,42	0,40
F-mesure	0,92	0,92	0,64	0,68	0,64	0,65	0,77	0,81	0,88	0,91	0,33	0.,35

En effet, la performance de SVM s'est dégradée pour presque toutes les catégories lorsque le sur-échantillonnage a été appliqué. En revanche, avec RF et LSTM, les catégories ont, en règle générale, progressé dans leurs performances. Le premier a notamment vu une amélioration concernant la description, mais avec une précision améliorée et un rappel diminué. En dépit de leur progrès grâce au sur-échantillonnage, le résultat général de RF et LSTM reste toujours inférieur à celui du classifieur de SVM pour lequel la technique CSL est utilisée. Par ailleurs, puisque nous privilégions CSL, qui maintient les données telles quelles, le sur-échantillonnage appliqué a peu d'intérêt, particulièrement pour les classes minoritaires.

Ainsi, le sur-échantillonnage peut être employé lorsque peu de données sont disponibles. Cependant, la description est très hétérogène et peut varier selon les clients, ce qui donne lieu à un large éventail de vocabulaires et de contextes. En outre, les nouvelles instances reproduites par le sur-échantillonnage ne sont pas des données authentiques, il ne s'agit que d'exemples dupliqués issus des données originales. Pour ces raisons, dans le cadre de cette étude, nous ne considérons pas le sur-échantillonnage comme la meilleure solution.

5.7 Conclusion

Ce chapitre a décrit les expériences portant sur la détection automatique de l'évaluation selon six catégories proposées en exploitant deux techniques de l'apprentissage supervisée : l'apprentissage de surface et l'apprentissage profond. Le premier exige la création manuelle des caractéristiques par les humains, alors que le second ne nécessite pas d'intervention humaine. Pour les premières expériences, le problème du déséquilibre des classes a été pris en compte par l'application de la technique CSL, pénalisant la classe majoritaire. Les performances de l'apprentissage de surface ont été améliorées par l'ajout de traits linguistiques. La meilleure moyenne pondérée de la F-mesure (0,8818) a été obtenue grâce au classifieur SVM et est supérieure à celle de l'apprentissage profond (LSTM : 0,8497). Néanmoins, ce dernier est intéressant en termes de temps et d'efforts, car il est moins coûteux que l'apprentissage de surface. Les moyennes pondérées de la F-mesure pour chaque catégorie, avec le meilleur classifieur (SVM),

sont : POS_OPINION (0,94), NEG_OPINION (0,72), MIX_OPINION (0,69), SUGGESTION (0,84), INTENTION (0,88) et DESCRIPTION (0,41). Parmi ces catégories, la description a été la plus difficile à détecter, donnant les résultats les moins satisfaisants.

Le sur-échantillonnage a été appliqué au cours des dernières expériences, ce qui a permis de résoudre le problème du manque d'échantillons des classes minoritaires. La figure 5.9 illustre une synthèse des résultats au moyen de la boîte à moustaches. Le point clair en étoile correspond à la meilleure F-mesure lorsque le classifieur donné a été utilisé avec application de la technique CSL, tandis que les points restants affichent les F-mesures obtenues en exploitant le sur-échantillonnage. Grâce au sur-échantillonnage, les classifieurs RF, CNN et LSTM ont connu une amélioration de leur performance, alors que les classifieurs NB, LR et SVM ont connu une chute, et que XG présente de très faibles différences.

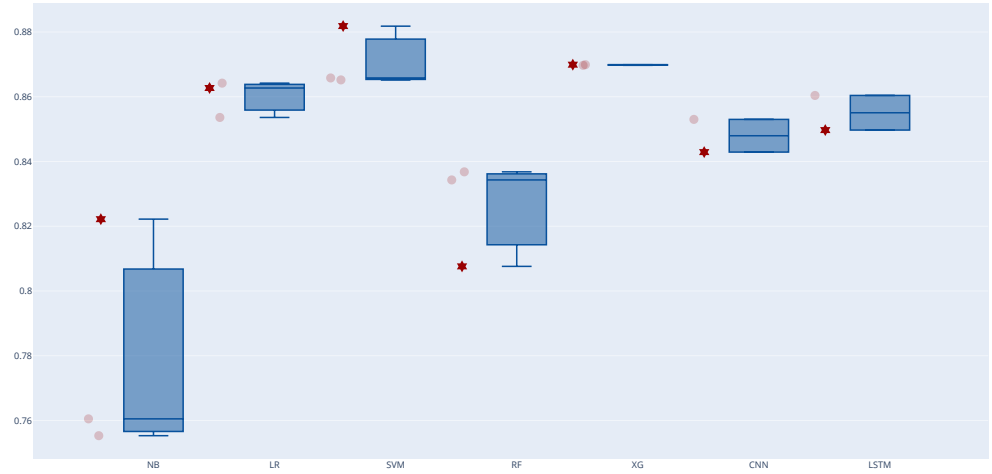


FIG. 5.9 : Synthèse de résultats (CSL, sur-échantillonnage)

Malgré sa légère amélioration grâce au sur-échantillonnage (RF, CNN et LSTM), la description a toujours obtenu la plus mauvaise performance, ce qui signifie que la détection de cette catégorie représente une tâche difficile. L'augmentation de la taille du corpus de référence contenant plus d'exemples des descriptions pourrait améliorer les résultats de sa détection. Cependant, la collecte des données (notamment celles de cette catégorie) n'est pas évidente, car seule une petite portion des données

supplémentaires recueillies concernerait la description. Ainsi, bien que le sur-échantillonnage soit pratique lorsque peu de données sont disponibles, dans le cadre de cette étude, il semble qu'il ne présente pas un grand d'intérêt pour les classes minoritaires.

Troisième partie

APPLICATION

Chapitre 6

Application de la typologie proposée

6.1 Introduction

Le modèle d'évaluation proposé dans ce travail, qui comprend six catégories (opinion positive/négative/mixte, suggestion, intention et description), s'est basé sur une observation manuelle du corpus constitué d'avis sur les restaurants rédigés en français. Ce corpus a servi également de référence pour le développement du module de la détection automatique de ces six catégories. L'une des étapes essentielles lors de l'élaboration d'un nouveau module consiste à évaluer sa généralisabilité (Clark et Watson 2016). Le travail présenté dans ce chapitre vise à vérifier si le modèle proposé peut s'appliquer à d'autres données. Afin d'éprouver sa généralisabilité, nous l'avons testé sur deux corpus supplémentaires. Le premier est un corpus qui fait varier le domaine puisqu'il est constitué d'avis sur des hôtels (HotelFR), quant au second il fait varier la langue (moins étudiée) puisque le corpus est composé d'avis sur des restaurants, mais en coréen (RestoKR).

6.2 Application du modèle au corpus portant sur l'hôtellerie

La réussite de la généralisation d'une approche dans un autre domaine dépend principalement de deux éléments (Szarvas *et al.* 2012) : d'abord, les domaines source et cible doivent être étroitement liés pour permettre le partage des connaissances ; deuxièmement, l'adaptation doit se fondre

sur les points communs des deux domaines, tout en garantissant les caractéristiques particulières du domaine cible. Bien que l'hôtellerie et la restauration utilisent des lexiques différents, il s'agit dans les deux domaines de situations vécues dans un lieu, dont l'évaluation porte sur le processus de satisfaction et la valorisation de l'expérience. Ainsi, nous nous intéresserons aux procédés linguistiques utilisés pour exprimer l'évaluation dans le corpus portant sur le domaine de l'hôtellerie, ce qui permet également de vérifier l'applicabilité du modèle proposé sur un autre corpus et de le généraliser ainsi à d'autres domaines. Nous décrirons d'abord le processus d'annotation manuelle, puis nous présenterons des analyses qualitatives et quantitatives du corpus annoté en comparant les résultats obtenus avec ceux issus de l'analyse du corpus RestoFR. Enfin, nous développerons le module de la détection automatique de six catégories d'évaluation en nous basant sur le corpus HotelFR annoté en suivant les mêmes démarches que pour le corpus RestoFR (voir chapitre 5).

6.2.1 *Corpus d'étude*

Un corpus français portant sur les hôtels a été collecté sur un site web dédié aux avis à propos d'hôtels, de restaurants, de villes et de régions ainsi que de lieux de loisirs visités[68]. Trois hôtels situés à Paris ont été choisis, et pour chacun de ceux-ci, les avis ont été extraits selon leur date de parution : ceux qui sont antérieurs au mois de février 2020 ont été collectés, dans la limite de 33 avis. Au total, le corpus est constitué de 97 avis, ce qui correspond à 296 phrases. Comme pour RestoFR, les avis qui se rapportent à HotelFR peuvent s'adresser à la fois à ceux qui gèrent l'hôtel concerné (afin qu'ils prennent conscience des problèmes qui se posent) et aux clients potentiels (futurs clients).

6.2.2 *Annotation manuelle*

L'annotation manuelle a été réalisée par deux doctorantes en linguistique[69]. Ces dernières ont suivi le guide d'annotation qui préconise d'attribuer une étiquette à chaque phrase, parmi six étiquettes propo-

[68] https://www.tripadvisor.com [consulté le 1er août 2020].
[69] Laura Noreskal et l'auteure.

sées : POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION et DESCRIPTION. Quelques exemples de l'annotation manuelle réalisée sont présentés ci-dessous (voir tableau 6.1).

TAB. 6.1 : Exemple d'annotation manuelle (HotelFR)

Phrases	**Catégories**
Le studio est très grand et bien aménagé.	**POS_OPINION**
Literie en mauvaise état.	**NEG_OPINION**
Le quartier Montmartre est très agréable et vivant même si le soir, on entendait un peu les gens attablés aux cafés en face mais rien de trop gênant non plus.	**MIX_OPINION**
Juste ajoutez une machine de séchage de vêtement.	**SUGGESTION**
Nous reviendrons avec plaisir.	**INTENTION**
La nuit avec le petit déjeuner est à 159 €.	**DESCRIPTION**

L'AIA a été calculé sur le corpus de 100 phrases en utilisant la mesure Kappa de Cohen (Cohen 1960). Nous avons obtenu un score de 0,83, moins satisfaisant que celui obtenu pour RestoFR (0,90 – voir chapitre 5), néanmoins considéré comme « presque parfait » selon l'échelle proposée par Landis et Koch (1977). Le reste du corpus (196 phrases) a été annoté par une des annotatrices, ce qui donne 296 phrases annotées au total. La répartition des catégories dans ce corpus est présentée dans la figure 6.1. Notons qu'en comparaison à RestoFR (cf. figure 3.4 du chapitre 3), la proportion de la description a augmenté de manière significative puisqu'elle passe de 1,85 % à 10,47 %. Nous détaillerons cette augmentation dans la section suivante.

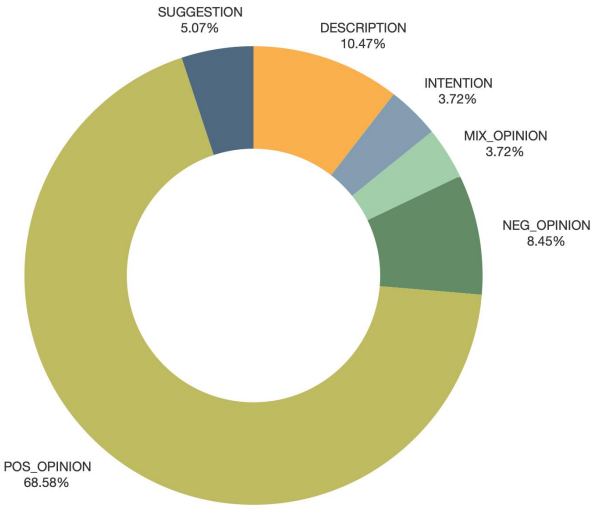


FIG. 6.1 : Répartition des catégories d'évaluation (HotelFR)

6.2.3 *Analyse du corpus annoté*

Au cours de cette étude, nous avons montré que l'évaluation peut être exprimée au moyen de différents niveaux linguistiques : lexical, syntaxique, énonciatif, etc. Le lexique d'évaluation peut être de deux types : certains lexiques évaluatifs sont spécifiques au domaine concerné (*délicieux*, *savoureux*, *sucré*), d'autres sont plus généraux et peuvent être employés dans n'importe quel domaine (*excellent*, *super*, *sympa*). Par ailleurs, des structures syntactiques ou textuelles, indépendantes du domaine donné, peuvent aussi contribuer à l'expression de l'évaluation (plus d'usage des syntagmes nominaux par exemple). Dans cette optique, nous présenterons une analyse qualitative et quantitative sur le corpus HotelFR, ce qui nous permettra de dégager à la fois les points communs et les différences entre le domaine de l'hôtellerie et celui de la restauration.

6.2.3.1 Analyse qualitative

Les corpus HotelFR et RestoFR se rapportent toutes les deux à l'évaluation. Bien que les domaines traités soient différents (l'hôtellerie et la restauration), il s'agit de décrire des situations vécues dans un lieu,

dont l'évaluation porte davantage sur la satisfaction et la valorisation de l'expérience. Or, les normes et valeurs en vigueur dans chaque domaine sont différentes. Selon Galatanu (2003), celles qui constituent l'évaluation sont notamment esthétiques, pragmatiques, cognitives, intellectuelles et hédoniques-affectives (chapitre 2). Parmi ces valeurs, certaines sont davantage exigées dans un domaine que dans l'autre. Ainsi, dans le domaine de l'hôtellerie, le confort est un aspect primordial au cours d'un voyage : ce sont les équipements, les installations et la localisation qui permettent de l'assurer. Il est donc naturel que les avis émis sur les hôtels fassent davantage mention de ces éléments. En conséquence, HotelFR a tendance à inclure de nombreuses informations factuelles (la description) sur l'hôtel, comme on peut le voir dans [1] ou [2].

[1] *Petit coin cuisine équipée avec micro-ondes, frigo, lave-vaisselle.*

[2] *Un écran pour regarder un peu la TV.*

Même si la description est moins présente dans le corpus RestoFR, elle comprend davantage d'informations factuelles possédant des caractéristiques anecdotiques comme dans [3]. Cette observation se vérifie sur un autre corpus RestoKR [4] lui aussi relatif au domaine de la restauration, mais dans une langue différente. En dépit de sa nature factuelle, le client fournit au lecteur un contexte supplémentaire au travers du récit de son expérience personnelle et révèle davantage d'informations sur lui. Il s'agit d'une façon de nouer des liens interpersonnels avec les lecteurs. Par conséquent, bien que la description soit essentiellement factuelle, nous constatons que sa nature et sa proportion dans un corpus peuvent différer selon les domaines.

[3] *Un choix pour un dîner en famille, afin de fêter les 18 ans de notre dernière fille.*

[4] 오랜만에 아는 동생을 만나서 간 식당.
(*C'est le restaurant où je suis allé(e) avec un(e) ami(e) que je n'avais pas rencontré(e) depuis longtemps.*)

Par ailleurs, il semble que la description est plus ambiguë dans le domaine de l'hôtellerie et peut être confondue avec deux autres catégories : l'opinion positive et l'opinion négative. Cette difficulté a été observée lors de l'annotation manuelle par les deux annotatrices. Prenons le cas de l'exemple précédent [1]. Celui-ci a été annoté comme la description par la première annotatrice (A1) car, d'après elle, cet exemple ne fournit que des informations sur les équipements à disposition de la cuisine. En revanche, l'autre annotatrice (A2) l'a interprété comme une opinion positive, car pour elle, la cuisine était équipée de tout ce dont le client avait besoin pour un séjour. Comme nous l'avons vu dans le chapitre 3, les clients peuvent exprimer leur satisfaction ou leur mécontentement à travers une représentation de la situation concernée, jugée favorable ou défavorable selon l'interprétation axiologique du locuteur qui lui est propre ou personnel. En effet, Martin et White (2003) affirment qu'une interprétation subjective de l'évaluation est inévitable, car celle-ci est le fruit d'une négociation entre les auteurs des textes (donc le client) et les lecteurs, sans compter que la position interprétative de ces derniers est déterminée par les contextes culturels et les expériences personnelles. Martin et White (2003, p.62) appellent ce type d'interprétation une évaluation « impliquée » (*invoked*). Cette dernière fait appel à des choix idéologiques qui sont superficiellement neutres, mais qui ont le potentiel d'évoquer des réactions de jugement envers ceux qui partagent un ensemble spécifique de valeurs idéologiques. Par conséquent, ils rendent difficile la distinction entre la description et les opinions car même de simples informations factuelles peuvent être perçues différemment d'un lecteur à l'autre.

Étant donné cette difficulté d'interprétation, nous avons opté pour une approche restreinte : si la phrase ne contient pas d'indice axiologique de manière limpide et explicite, nous l'avons considérée comme une description [1], [2] ; dans le cas contraire, comme une opinion [5].

[5] *La cuisine a* ***tout*** *le nécessaire (assiettes, couverts, verres, casseroles jusqu'au seau à champagne et j'en passe), une machine Nespresso, lave-vaisselle.*

6.2.3.2 Analyse quantitative

Avant de procéder à l'analyse quantitative, les données ont été nettoyées et prétraitées pour faciliter l'exploration de manière automatique. Elles sont également exploitées plus tard dans la section suivante pour la détection automatique. Tout comme le chapitre 4, l'analyse quantitative passe par l'extraction de mots-clés, la répartition en catégories morphosyntaxiques et l'analyse de la position des catégories au sein d'un avis. L'observation a permis de révéler à la fois des ressemblances et des différences entre les deux corpus (HotelFR et RestoFR).

Les prétraitements réalisés sont les mêmes que ceux décrits dans le chapitre 4. Nous nous contentons de présenter ici leur synthèse. Afin de nettoyer et prétraiter le corpus, certaines démarches ont été effectuées, comme le remplacement des émoticônes par « emoPOS » ou « emoNEG » selon leur polarité, le passage des mots en minuscules et leur normalisation, le développement des abréviations pour donner le mot correspondant en toutes lettres (*resto* en *restaurant*), le remplacement des chiffres par « NUM » et la lemmatisation en utilisant StanfordCoreNLP.

Pour obtenir un aperçu sur la similarité sémantique entre HotelFR et RestoFR, nous avons procédé au calcul de la similarité du cosinus. Cette métrique, fréquemment exploitée en recherche d'informations (RI) ou en fouille de textes, permet de déterminer dans quelle mesure les deux documents sont similaires au niveau de leur thème, quelle que soit leur taille (Singhal et Google 2001). Elle mesure le cosinus de l'angle entre les deux documents, représentés en vecteurs et projetés dans un espace multidimensionnel. Chaque dimension correspond à un mot du document, et la similarité du cosinus saisit l'orientation (l'angle) des documents : plus l'angle est petit, plus la valeur de cosinus est grande (plus proche de 1), tout comme la similarité. Afin de calculer la similarité du cosinus entre les deux corpus, les fonctions `cosine_similarity` et `TfidfVectorizer`, disponibles dans la bibliothèque scikit-learn, ont été exploitées sur les corpus lemmatisés avec les mots-outils supprimés. Nous avons obtenu un score de 0,63, ce qui montre que les deux corpus sont modérément similaires. Autrement dit, bien qu'il s'agisse de deux domaines différents, certains termes utilisés dans l'évaluation de l'hôtellerie sont également présents dans celle de la restauration.

Afin de mieux cerner les lexiques utilisés dans HotelFR, nous avons d'abord examiné les mots-clés par rapport à chaque catégorie d'évaluation à l'aide du logiciel AntConc, illustrés dans la figure 6.2. En raison de la taille relativement petite du corpus, très peu de mots-clés ont été extraits par le logiciel. Notons que la majorité de ceux-ci figurent également dans RestoFR (cf. chapitre 4), par exemple *NUM* dans la description, *revenir* et *retourner* dans l'intention, *mais* dans l'opinion mixte, *pas* dans l'opinion négative, *très* dans l'opinion positive, et enfin *recommander*, *vivement* ou *conseiller* dans la suggestion. Bien que les lexiques utilisés dans chaque domaine soient différents, les mots-clés obtenus montrent qu'il existe des termes systématiquement exploités pour exprimer chaque catégorie de l'évaluation.

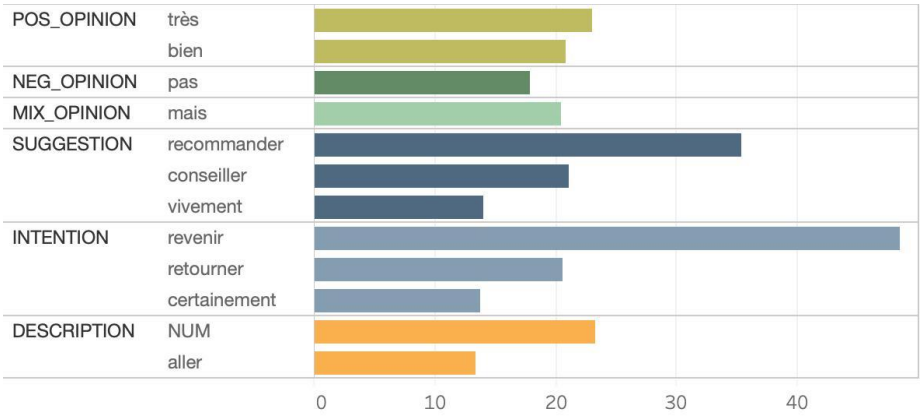


FIG. 6.2 : Mots-clés de chaque catégorie (HotelFR)

La répartition en catégories morphosyntaxiques a ensuite été analysée par le biais de l'AFC. Avant la réalisation de celle-ci, l'existence d'une relation entre les catégories morphosyntaxiques et chaque catégorie d'évaluation a été vérifiée grâce au test du khi-deux d'indépendance, ce qui donne un risque d'erreur $\alpha = 5\ \%$ ($\chi 2 = 184{,}48$; $ddl = 35$; $p < 2{,}2\text{e-}16$). L'AFC est illustrée dans la figure 6.3 grâce au langage R et au package FactoMineR. Les deux axes (ou dimensions) représentent respectivement environ 47,67 % et 44,79 % de l'inertie totale, ce qui représente 92,46 % des variations observées dans nos données.

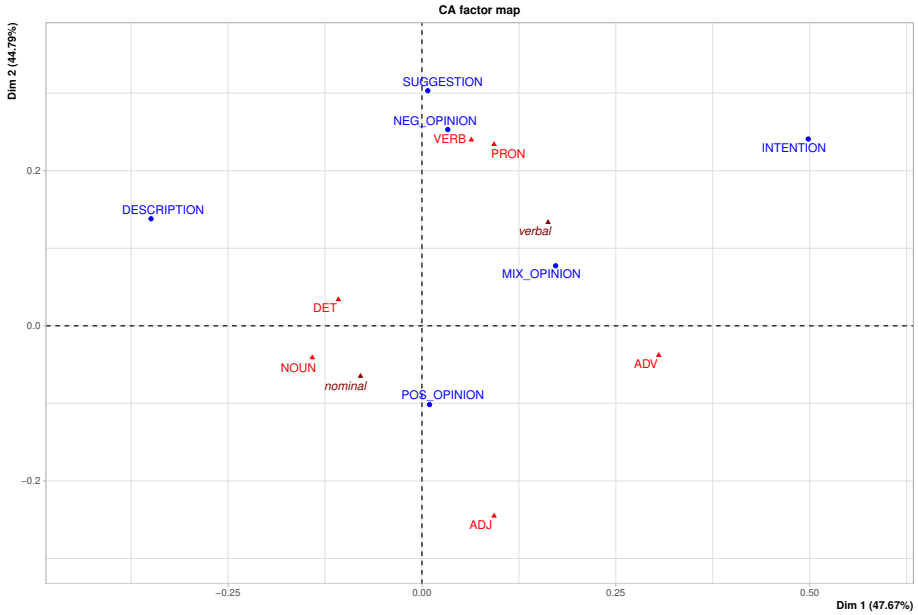


FIG. 6.3 : Analyse symétrique standard de l'analyse factorielle des correspondances (HotelFR)

Si l'on observe le premier axe, on constate une coupure entre la description, les noms et les déterminants situés à gauche du graphique et d'autres catégories morphosyntaxiques et d'évaluation figurant à droite. Ce résultat peut être expliqué par la spécificité du domaine hôtellier que nous avons constatée précédemment : les avis émis comportent plus d'informations factuelles (la description) sur les équipements disponibles, les installations et la localisation dans l'établissement, comme [1] représentés sous forme de syntagmes nominaux dans le corpus.

Sur le second axe, une tendance analogue est observée entre HotelFR et RestoFR, avec toutefois une légère différence : les catégories faisant partie d'un syntagme nominal (adjectifs et noms) se trouvent en dessous de l'axe en bas du graphique, tout comme l'opinion positive. Autrement dit, cette dernière se rattache fréquemment aux syntagmes nominaux, comme dans [6] ou [7].

[6] *Un très bel appartement, très grand, parfaitement bien soigné, fournis, propre, élégant, un accueil impeccable.*

[7] *La salle de bain est parfaite aussi avec peignoirs, chaussons, produits de base.*

Les verbes et les pronoms, en revanche, sont davantage liés aux syntagmes verbaux se situant en haut du graphique avec les autres catégories d'évaluation (suggestions, intentions, opinions négatives et mixtes). Tout comme dans le corpus RestoFR, la suggestion et l'intention sont toujours associées à l'action : la première incite l'hôtel ou les futurs clients à effectuer une action [8], [9], la seconde concerne l'action qui sera réitérée à l'avenir par le client lui-même [10].

[8] *Juste ajoutez une machine de séchage.*

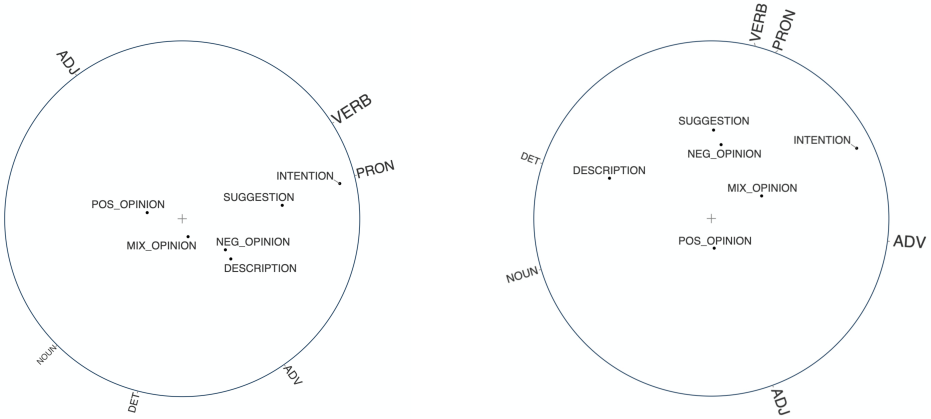
[9] *Que ce soit pour un court ou un long séjour, nous conseillons vivement cet endroit !*

[10] *Nous reviendrons avec plaisir.*

Afin d'avoir une meilleure vue d'ensemble des deux corpus (HotelFR et RestoFR), nous avons produit un graphique de lune (*moonplot*) pour chacun d'eux (figure 6.4). Développé par Bock (2011), il est disponible sur une plateforme dédiée à la visualisation des données[70]. Le principe de lecture de ce graphique est le suivant : les catégories d'évaluation sont affichées à l'intérieur du cercle dans lequel la distance qui les sépare est significative, alors que les catégories morphosyntaxiques sont indiquées à la périphérie du cercle se situant dans le même angle que dans l'AFC. Par ailleurs, la taille de police représente la distance du centre de gravité à l'intérieur de l'AFC : la taille différente montre en effet que les catégories morphosyntaxiques ne se valent pas. À titre d'exemple, la taille relativement petite des noms (figure 6.4a) ou des déterminants (figure 6.4b) témoigne du fait que ces catégories morphosyntaxiques sont peu signi-

[70] https://www.displayr.com/ [consulté le 6 septembre 2020]).

ficatives dans chacun des corpus. Contrairement à l'AFC, ce graphique rend possible et plus simple l'interprétation de la distance entre les deux variables (catégories d'évaluation et catégories morphosyntaxiques)[71].



(a) RestoFR (b) HotelFR

FIG. 6.4 : Graphique de lune

Le graphique met en lumière les points communs et spécificités observés dans HotelFR et RestoFR. Dans les deux corpus, l'opinion positive est davantage associée aux syntagmes nominaux, tandis que la suggestion et l'intention sont liées aux syntagmes verbaux, dont la taille de la police des verbes et des pronoms est grande. Toutefois, l'intention dans HotelFR est plus proche des adverbes, comme le montrent certains exemples [11]-[13] dans lesquels les adverbes se trouvent liés à des verbes par le préfixe *re-*. Le corpus HotelFR présente, cependant, une différence significative par rapport à RestoFR quant à la présence de la catégorie de description. Puisque cette catégorie englobe non seulement l'arrière-plan de l'expérience (la raison pour laquelle les clients se rendent dans cet hôtel, les personnes qui les accompagnent, etc.), mais également l'énumération des équipements disponibles et des installations dans l'établissement, elle

[71] Sur le graphique de l'AFC, il s'agit de la distance entre les catégories d'évaluation présentées par des points bleus et les catégories morphosyntaxiques symbolisées par des triangles rouges, qui n'a pas de sens dans l'absolue (http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/74-afc-analyse-factorielle-des-correspondances-avec-r-l-essentiel/ [consulté le 19 octobre 2020]).

est davantage associée aux noms et déterminants. En ce sens, notons par ailleurs, que la taille de la police des noms est plus grande que RestoFR dans la figure 6.4b.

[11] *Nous reviendrons certainement.*

[12] *Nous y retournerons évidemment aussi en couple.*

[13] *Nous reviendrons assurément.*

Enfin, un graphique d'association a été produit pour vérifier le lien entre une catégorie d'évaluation et sa position dans un avis posté. Selon le test du khi-deux, la relation entre les deux est significative avec un risque d'erreur $\alpha = 5\ \%$ ($\chi 2 = 21{,}812$; $ddl = 5$; $p = 0{,}0005685$). Les avis ne comportant qu'une seule phrase ont été exclus. En observant la figure 6.5, nous constatons une tendance assez similaire entre RestoFR (voir chapitre 4) et HotelFR, à l'exception de l'opinion mixte. En effet, au niveau de ce dernier, les clients sont plutôt enclins à commencer une évaluation par un arrière-plan (la description) ou directement par une opinion positive. En revanche, les clients ont tendance à finir soit par une suggestion aux lecteurs (les clients potentiels et l'hôtel), soit par une intention du client de revenir ou pas à l'hôtel, soit par une opinion négative concernant un des aspects du lieu visité. La seule différence entre RestoFR et HotelFR réside dans le fait que, dans le cas de ce dernier, les opinions mixtes apparaissent davantage à la fin d'une évaluation. Bien que nous ne puissions rien conclure en raison de la faible quantité de données à notre disposition, nous avons constaté que les avis finissent par une opinion mixte qui, soit nuance un avis positif avec un détail négatif, soit résume une expérience mitigée.

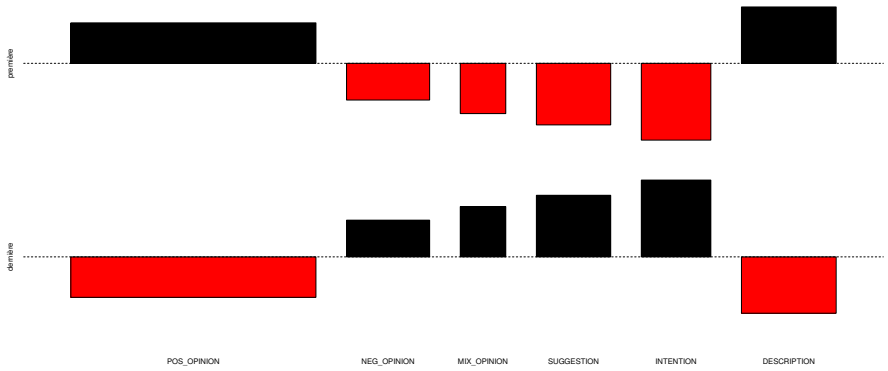


FIG. 6.5 : Graphique d'association entre la catégorie d'évaluation et la position (HotelFR)

L'analyse quantitative a ainsi montré comment et en quoi les deux corpus (RestoFR et HotelFR) se ressemblent ou se différencient. Malgré des différences observées, il semble que notre typologie d'évaluation soit généralisable également au secteur de l'hôtellerie. Chaque corpus contient des termes spécifiques au domaine qu'il représente ; les mots-clés obtenus montrent néanmoins que certains termes sont systématiquement employés pour exprimer chaque catégorie de l'évaluation (les chiffres pour la description, *revenir*, *retourner* pour l'intention, *mais* pour l'opinion mixte, *pas* pour l'opinion négative, *très* pour l'opinion positive et *recommander*, *conseiller* et *vivement* pour la suggestion). En ce qui concerne les catégories morphosyntaxiques, la plupart des catégories d'évaluation, notamment l'intention et la suggestion, ont tendance à employer plus de syntagmes verbaux pour décrire les actions du client ou des lecteurs. L'opinion positive, quant à elle, est davantage exprimée avec des syntagmes nominaux incluant les termes évaluatifs (souvent les adjectifs) et les cibles sur lesquelles elle porte (les noms). En comparaison au corpus RestoFR, dans HotelFR la description concerne davantage d'informations factuelles relatives aux équipements et à l'établissement qui résulte du recours à des noms et à des déterminants. Par ailleurs, les catégories d'évaluation contenues dans les deux corpus occupent des positions similaires au sein d'un avis : la description et l'opinion positive sont placées majoritairement

au début d'une évaluation tandis que l'intention et la suggestion sont positionnées à la fin.

6.2.4 *Annotation automatique*

Dans cette partie, nous nous intéressons à la détection automatique de l'évaluation modélisée à travers les six catégories dans le corpus HotelFR. Nous avons adopté la méthodologie exploitée précédemment pour le corpus RestoFR (voir chapitre 5). Nous considérons trois jeux différents d'entraînement que nous avons représentés par les lettres H, R et H+R :

- H : 67 % de HotelFR de manière stratifiée ;
- R : l'ensemble de RestoFR ;
- H+R : l'ensemble de RestoFR et 67 % de HotelFR de manière stratifiée.

Pour les trois jeux d'entraînement (H, R et H+R), l'évaluation a été effectuée sur le même jeu de test qui représente 33 % de HotelFR de manière stratifiée (98 phrases).

6.2.4.1 Expériences

Les données prétraitées lors de l'analyse quantitative ont été exploitées. Parmi les divers algorithmes, nous avons choisi ceux ayant produit les meilleurs résultats pour l'apprentissage de surface et l'apprentissage profond sur le corpus RestoFR – SVM et LSTM respectivement. Le premier a été réalisé à l'aide de la bibliothèque scikit-learn et le second avec les bibliothèques Keras et TensorFlow.

Afin d'exploiter le classifieur SVM, les données textuelles ont d'abord été représentées en vecteurs en choisissant la version de données textuelles (lemmatisée, lemmatisée ayant les mots outils supprimés) et en ajustant la méthode de représentation vectorielle (`CountVectorizer` ou `TfidfVectorizer`) et le paramètre `n_gram`. Leur meilleure combinaison a été obtenue grâce à une procédure de grille de recherche (`GridSearch`). Les caractéristiques prises en compte lors de l'apprentissage sont (voir section 5.3.2.3) : les catégories morphosyntaxiques jugées pertinentes proposées par StanfordCoreNLP et TreeTagger, les différentes variations des verbes, la négation, la conjonction *mais*, les mots positifs et négatifs, les

scores de polarité et de subjectivité obtenus par TextBlob, la position de la phrase dans un avis, le nombre de caractères, la longueur de la phrase, la diversité et la densité lexicales, les ponctuations multiples et l'unité monétaire (*euro* et €). Les valeurs des traits ont été normalisées en utilisant `RobustScaler` de scikit-learn.

En ce qui concerne la technique des LSTM, nous avons pris en entrée une matrice d'*embedding* à l'aide de Word2vec. Ce dernier a été créé à nouveau pour chaque jeu d'entraînement, puisque nous disposions de données différant de celles des expériences réalisées dans le chapitre précédent, tout en fixant les mêmes paramètres.

Pour toutes les expériences, le paramètre `class_weight` a été ajusté selon la méthode CSL afin de résoudre le problème de la disproportion des classes. Enfin, pour évaluer toutes les expériences, une validation croisée stratifiée à cinq plis a été effectuée.

6.2.4.2 Résultats

Les performances des différents jeux d'entraînement ont été évaluées en calculant pour chacun d'eux la moyenne pondérée de la précision, du rappel et de la F-mesure. Le tableau 6.2 illustre la moyenne pondérée de la F-mesure lorsque SVM et LSTM ont été appliqués aux trois différents jeux d'entraînement.

Tab. 6.2 : Moyenne pondérée de la F-mesure (HotelFR)

Jeu d'entraînement	**F-mesure**	
	SVM	**LSTM**
H	0,6366	0,0189
R	0,7992	0,5160
H+R	**0,8064**	**0,6684**

Le classifieur SVM entraîné avec H+R semble fournir le meilleur score, qui est la F-mesure de 0,8064. Les résultats obtenus sont généralement faibles par rapport aux expériences que nous avons effectuées sur le corpus RestoFR dans le chapitre précédent et à l'issue desquelles le score

moyen était de 0,8488. Notre meilleur résultat n'est cependant pas très éloigné de l'accord inter-annotateurs (0,83). Les performances de SVM et LSTM se sont considérablement améliorées grâce à la combinaison des données (HotelFR et RestoFR), passant respectivement de 0,6366 à 0,8064, et de 0,0189 à 0,6684. Il semble que les données de RestoFR aient permis de combler le manque d'informations relatives à HotelFR. Ainsi, bien que les deux ensembles de données soient issus de domaines différents (hôtellerie et restauration), ils partagent certains points communs. Aussi, plus nous disposons de données, meilleures seront les performances.

La précision et le rappel du meilleur résultat (H+R avec SVM) pour chaque catégorie sont présentés dans le tableau 6.3. L'intention est détectée avec la meilleure performance (la F-mesure étant 1,00), suivie par l'opinion positive et la suggestion dont la F-mesure se situe autour de 0,90. En revanche, la description possède la plus mauvaise performance (0,40). En comparaison au corpus RestoFR (cf. chapitre 5), nous observons également un écart plus important entre la précision et le rappel (hormis pour l'intention, ayant une précision, un rappel et une F-mesure identiques).

TAB. 6.3 : Précision, rappel et F-mesure de chaque catégorie (HotelFR-SVM)

	POS_OPINION	NEG_OPINION	MIX_OPINION	SUGGESTION	INTENTION	DESCRIPTION
Précision	0,84	0,67	0,67	1,00	1,00	0,60
Rappel	0,97	0,44	0,50	0,80	1,00	0,30
F-mesure	0,90	0,53	0,57	0,89	1,00	0,40

Ainsi, les performances observées au cours des expériences réalisées dans cette section sur le corpus HotelFR sont inférieures à celles effectuées sur le corpus RestoFR. Néanmoins, ces résultats ont permis de constater que notre modèle d'évaluation proposé sous forme de six catégories ainsi que le module développé de leur détection automatique sont généralisables et reproductibles sur d'autres données.

6.2.5 *Pour conclure*

Cette section a soulevé la question de l'application du modèle d'évaluation élaboré pour le corpus RestoFR sur le corpus issu d'un autre domaine, l'hôtellerie, HotelFR. L'analyse qualitative a permis de mettre en évidence les spécificités de l'évaluation dans le secteur de l'hôtellerie. Nous observons davantage de descriptions qui sont parfois difficiles à différencier des opinions. Cette difficulté est liée à la perception des valeurs qui est dépendante des idéaux et des préférences de chacun (Bréchon 2012). Par ailleurs, en dépit de la différence lexicale propre à chaque domaine (hôtellerie et restauration), l'analyse quantitative a montré qu'il existe des moyens systématiques relativement indépendants du domaine pour exprimer une évaluation. Cela signifie que notre typologie d'évaluation peut être généralisée à d'autres domaines tant que l'évaluation porte sur des expériences à un endroit donné (un lieu touristique ou un magasin par exemple). Par ailleurs, des données qui viennent du domaine de la restauration peuvent améliorer la performance de la détection automatique appliquée au domaine hôtelier grâce à des constructions linguistiques communes. Ainsi, lorsque nous ne disposons pas de suffisamment de données annotées pour un domaine en particulier, celles relatives à un domaine qui s'en rapproche peuvent pallier ce déficit et donc permettre de réduire les coûts d'annotation nécessaires pour couvrir le manque des données.

6.3 Application du modèle sur le corpus en langue coréenne

La question à laquelle répond le travail présenté dans cette section est de savoir si notre module, développé par rapport à l'évaluation, peut être perçu et interprété de la même manière dans une culture et une langue différentes, car l'interprétation axiologique de la valeur – ce qui est bon ou mauvais, agréable ou désagréable, satisfaisant ou insatisfaisant, souhaitable ou à éviter – peut varier d'une culture à l'autre. Par ailleurs, dans le paradigme de la mondialisation, notamment sous l'effet de la révolution numérique, nous ne pouvons plus nous limiter à des documents rédigés en une seule langue. Notre objectif est donc d'évaluer la possibilité de tester

notre module sur une autre langue : le coréen. Si notre démarche est jugée fiable, les différentes catégories devraient également être identifiées dans le corpus de cette langue. Le prétraitement du coréen étant très différent et compliqué par rapport à celui du français, comme nous le montrerons par la suite, nous n'effectuerons pas pour ce corpus d'analyse quantitative et de détection automatique, nous nous en tiendrons à son annotation manuelle. En premier lieu, nous présenterons des travaux en coréen sur la fouille d'opinions. Ensuite, nous décrirons le corpus d'étude ainsi que les processus d'annotation manuelle. Enfin, nous présenterons une analyse générale du corpus annoté qui permettra d'examiner la compatibilité du module que nous proposons sur une autre langue que le français.

6.3.1 *Travaux en coréen*

Peu d'études ont été menées sur la fouille d'opinions en coréen, car la langue coréenne fait partie des langues agglutinantes, ce qui rend notamment difficile l'analyse des écrits numériques (Lim et Kim 2014).

Le contour des notions liées à la subjectivité et l'évaluation n'est donc pas précisé. De plus, la plupart des travaux sur le coréen est abordée dans le domaine du management du système d'information (*Management Information System* ou MIS) ou de l'informatique, davantage que dans celui de la linguistique. À notre connaissance, l'équipe rassemblée autour du linguiste Hyopil Shin[72] constitue l'une des rares équipes de recherche qui travaille sur la fouille d'opinions en coréen et qui prend en considération les particularités linguistiques (Jang *et al.* 2013 ; Jang et Shin 2010a,b ; Kim et Shin 2017 ; Shin 2013 ; Shin *et al.* 2016). Dans Jang et Shin (2010) par exemple, les chercheurs montrent comment les modifieurs (*valence shifters*) coréens peuvent agir sur l'orientation de la polarité et le prennent en compte dans la réalisation d'outils de fouille d'opinions.

Comme les travaux en anglais et en français, les travaux en coréen exploitent différentes méthodes de classification : l'apprentissage non supervisé (Jang et Shin 2010a ; Kim *et al.* 2018 ; Kim et Shin 2017 ; Myung *et al.* 2008 ; Shin *et al.* 2016), l'apprentissage supervisé de surface (Jang *et al.* 2013 ; Jang et Shin 2010b ; Kim *et al.* 2011 ; Kim, Jeong *et al.*

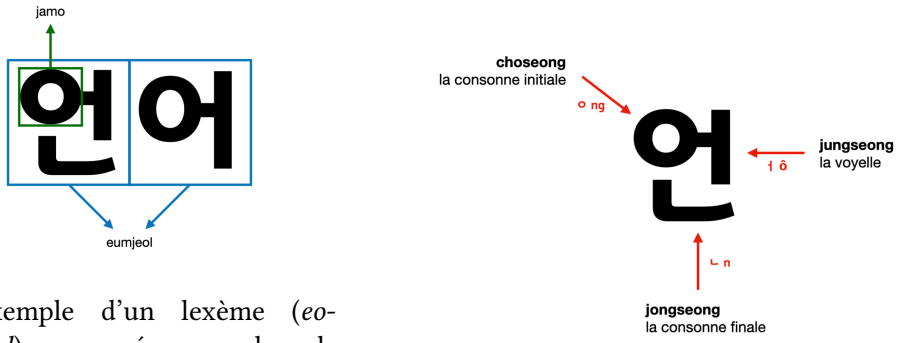
[72] http://knlp.snu.ac.kr/ [consulté le 6 août 2020].

2015 ; Lee *et al.* 2018 ; Lim et Kim 2014) et l'apprentissage profond (Park *et al.* 2018a ; Park *et al.* 2018b ; Song *et al.* 2019 ; Sung *et al.* 2020).

Cependant, en raison de la particularité du coréen par rapport aux langues latines, différentes démarches ont été mises en œuvre dans le cadre de la fouille d'opinions. En premier lieu, le coréen est une langue agglutinante qui utilise des morphèmes (suffixes ou clitiques) pour marquer des fonctions syntaxiques et pragmatiques, des temps, des modes, des niveaux de politesse, etc. Il s'agit en effet d'une langue dont les suffixes sont prédominants au regard des autres langues parlées dans le monde (Dryer 2013). Le token coréen ne correspond donc pas à un mot français : il s'agit d'une unité plus grande que le mot et que l'on appelle *eojeol*. C'est ce dernier qui entraîne la complexité et l'ambiguïté de l'analyse (Paumier et Nam 2014).

Jang et Shin (2010) réalisent l'étiquetage sur le chunking qui s'appuie sur les dépendances morphologiques. Leur méthode permet de procéder à une analyse efficace en fournissant des informations structurelles limitant la portée sémantique des éléments fonctionnels tels que la négation et les intensifieurs. Song *et al.* (2019) entraînent un modèle Word2vec en se basant non sur le token, mais sur les morphèmes – plus précisément certaines racines des adverbes, adjectifs et verbes.

Une autre caractéristique unique que possède le coréen réside dans le fait que les mots sont formés par une structure hiérarchique explicite. Chaque lexème (*eojeol*) peut être décomposé en une séquence de syllabes (*eumjeols*), qui à leur tour peuvent être décomposés en *jamos*, un ensemble d'unités phonétiques de base représentant les consonnes et les voyelles de la langue (figure 6.6a). Chaque *eumjeol* est composé de deux ou trois caractères (figure 6.6b) : le *choseong* (consonne initiale), le *jungseong* (la voyelle) et le *jongseong* (optionnel – la consonne finale).



(a) Exemple d'un lexème (*eojeol*) : une séquence de syllabes (*eumjeols*)

(b) Structure d'une syllabes (*eumjoel*)

FIG. 6.6 : Système d'écriture du coréen

Stratos (2017) a été, à notre connaissance, le premier à exploiter les *jamos* dans le traitement. Malgré le rôle important que jouent ces derniers dans le codage des informations syntaxiques et sémantiques des mots, ils ont été négligés dans les algorithmes traitant le coréen. Ils ont depuis attiré l'attention des Coréens : par exemple, Park *et al.* (2018) entraînent un modèle *skip-gram* au niveau du *jamo* sur le Wikipédia coréen, les articles de journaux et le corpus Sejong[73] (Kim *et al.* 2007). Ils exploitent ce modèle sur le corpus d'avis en ligne de films. De même, Sung *et al.* (2020) entraînent un modèle *skip-gram* de *jamo* sur un corpus d'articles de journaux avant de réaliser la fouille d'opinions qui portent sur des films.

Ainsi, la différence du coréen par rapport à d'autres langues, comme le français ou l'anglais, a ralenti les recherches dans la tâche de fouille d'opinions. Par ailleurs, le coréen connaît une extrême dispersion des données au niveau des mots (Song *et al.* 2019 ; Stratos 2017). De ce fait, un grand nombre d'études adoptent des méthodes particulières en s'appuyant non sur des mots, mais sur d'autres unités comme les morphèmes ou les *jamos*.

[73] https://github.korean.go.kr/user/main.do [consulté le 6 août 2020].

6.3.2 *Corpus d'étude*

Le corpus d'étude a été collecté sur le même site que le corpus HotelFR[74]. Puisqu'un grand nombre de restaurants se situent à Séoul, la capitale, nous en avons choisi six qui s'y trouvent et proposent de la nourriture coréenne. Les avis sur les restaurants ont été sélectionnés aussi en fonction de leur date de parution : pour chaque restaurant, nous avons extrait dix des avis les plus récents du mois de mai 2020. Ainsi, nous avons extrait 60 avis qui correspondent à 246 phrases.

6.3.3 *Annotation manuelle*

Deux annotatrices coréennes ont effectué l'annotation manuelle sur le corpus RestoKR : une doctorante en linguistique[75] et une autre en littérature[76]. L'une des six étiquettes (POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION et DESCRIPTION) a été attribuée à chaque phrase segmentée, comme dans le tableau 6.4. En utilisant la mesure Kappa de Cohen (Cohen 1960), nous avons obtenu 0,92 – meilleur AIA dans le cadre de cette étude.

[74] https://www.tripadvisor.com [consulté le 6 août 2020].
[75] L'auteure.
[76] Inhye Hong.

TAB. 6.4 : Exemple d'annotation manuelle (RestoKR)

Phrases	Catégories
직원분들도 친절하셔서 굉장히 만족했던 곳이에요. (*Le personnel s'est montré gentil, j'étais donc très satisfait(e) de cet endroit.*)	**POS_OPINION**
단지 조금 양이 적은 느낌이라 그게 조금 아쉬웠어요. (*La portion était petite, ce que j'ai trouvé un peu dommage.*)	**NEG_OPINION**
갈비찜은 부드럽고 맛있었지만 양이 너무 적었어요. (*Le Galbi-jjim* était doux et délicieux, mais la portion était trop petite.*)	**MIX_OPINION**
서울에 가게되면 한번 들릴만 한곳이라고 추천하고 싶다. (*Si on va à Séoul, je recommande de visiter cet endroit.*)	**SUGGESTION**
두 번 다시 가고 싶지는 않음. (*Je ne veux pas y retourner.*)	**INTENTION**
서울 광장시장에 있는 빈대떡 집입니다. (*Il s'agit d'un restaurant de Bindae-tteok** sur le marché de Gwangjang, à Séoul.*)	**DESCRIPTION**

* Les courtes côtes de bœuf braisées
** Les crêpes de pâte d'haricots mung

Si l'on observe la répartition de chaque catégorie (figure 6.7), elle s'avère relativement plus équilibrée que celle dans le corpus RestoFR (cf. figure 3.4 dans le chapitre 3). L'opinion positive représente toujours la majorité des catégories dans le corpus RestoKR (29,27 %), mais elle est moins importante que dans RestoFR (68,05 %). En revanche, l'opinion négative est passée de 9,5 % à 23,58 %. Si nous reconnaissons que ce résultat peut être dû à d'autres paramètres fixés lors de la sélection des données (tels que la taille réduite de ces derniers), nous observons que le principe de pollyanna – une tendance générale des personnes à préférer l'information positive – ne s'applique pas à ce corpus coréen. De plus, Jurafsky *et al.* (2014) admettent dans la conclusion de leur travail, que ce principe, dans les

avis émis par les clients, peut s'appliquer uniquement aux données en langue anglaise issues des États-Unis. Par ailleurs, la proportion d'opinions (positives/négatives/mixtes) a chuté, passant de 84,36 % à 59,35 %, tandis que celle de descriptions a augmenté, passant de 1, 85 % à 21,14 %. La répartition des catégories est analysée dans la section suivante.

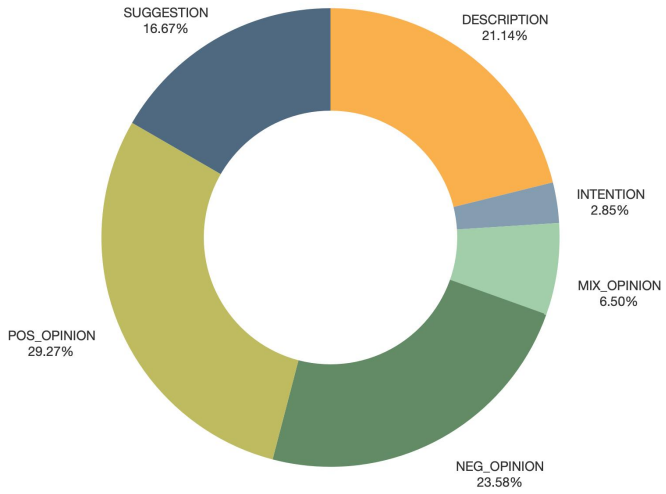


FIG. 6.7 : Répartition des catégories d'évaluation (RestoKR)

6.3.4 *Analyse du corpus annoté*

Nous avons réalisé l'annotation manuelle en nous appuyant sur la typologie proposée pour le corpus RestoKR. Cependant, l'analyse du corpus annoté montre la difficulté que les annotateurs ont rencontrée ; il s'agit de la distinction des catégories de description et d'opinion positive et négative, ce qui a été également le cas pour le corpus HotelFR. Prenons la phrase [14] : la première annotatrice (A1) l'a considérée comme relevant de la description, estimant qu'il s'agissait simplement d'un inventaire des farces dans le ravioli. La deuxième annotatrice (A2), en revanche, a relié cette phrase à l'opinion négative, parce qu'elle considère que les raviolis au bœuf devraient être principalement farcis de viande et non de tofu ou de légumes. Ainsi, tout comme pour le corpus HotelFR, les deux

annotatrices ont choisi différentes étiquettes, car elles se sont basées sur des valeurs qui leur sont propres.

[14] 고기만두라고 하지만, 두부와 야채를 많이 넣은 만두에요.
(*Ce sont des raviolis au bœuf, mais ils sont remplis de beaucoup de tofu et de légumes.*)

L'évaluation axiologique de valeur s'appuie par ailleurs sur les valeurs culturelles spécifiques. À titre d'exemple, la capacité à percevoir la texture collante ou gluante des aliments est universelle ; pourtant, bien que les Coréens apprécient une telle texture, les Occidentaux ne l'aiment pas[77]. De même, Hartmann *et al.* (2015) affirment que les clients chinois sont plus favorables que les clients allemands à la nourriture à base d'insectes, en raison de sa valeur nutritive, son goût, sa familiarité et son acceptation sociale.

Par ailleurs, les valeurs hédoniques et utilitaires – fréquemment prises en compte dans l'étude de la satisfaction des clients dans le contexte de la restauration (Ha et Jang 2010) – semblent varier d'une culture à l'autre en raison des différents systèmes d'interprétation. La valeur hédonique renvoie aux aspects affectifs, émotionnels, ainsi qu'à la signification symbolique concernant l'expérience au restaurant, tels que le plaisir et la joie (Holbrook et Hirschman 1982 ; Spangenberg *et al.* 1997), tandis que la valeur utilitaire concerne les aspects fonctionnels et économiques (Jarvenpaa et Todd 1996 ; Teo 2001). Les deux valeurs conditionnent à la fois les jugements évaluatifs des clients (Oliver 2014) et leur satisfaction (Babin *et al.* 1994 ; Jones *et al.* 2006). En ce sens, bien qu'il s'agisse dans les deux cas de personnes d'origine asiatique, certaines études (Cha *et al.* 2019 ; Luo et Wang 2014) ont permis de constater que les clients chinois sont plus susceptibles de réagir aux valeurs utilitaires d'un restaurant que les clients coréens, puisque le pragmatisme est une des priorités absolues en Chine.

[77] https://koreajoongangdaily.joins.com/2020/04/23/etc/Bringing-diners-to-a-specific-time-and-place-Chef-Joseph-Lidgerwood-adds-a-new-perspective-to-Koreas-favorite-ingredients/3076423.html [consulté le 13 septembre 2020].

Cependant, ce n'est pas seulement une question d'origine ; d'autres facteurs tels que la motivation du visiteur ou le type de restaurant peuvent également contribuer à influencer les valeurs priorisées. Les clients qui se rendent au restaurant pour le plaisir ou l'interaction sociale se concentreront sur la valeur hédonique, tandis que ceux qui y vont pour des raisons pratiques ou budgétaires prioriseront la valeur utilitaire (Park 2004). De même, les clients qui se rendent dans un restaurant exotique et ceux qui visitent un restaurant servant la cuisine locale traditionnelle ne privilégient pas les mêmes valeurs (Ha et Jang 2010). Malgré toutes ces différences, quels que soient la culture, la motivation, le type de restaurant ou autre, les évaluations peuvent être identifiées selon la typologie que nous proposons ; seule la répartition sera différente.

6.3.5 *Pour conclure*

Tout comme pour RestoFR, la description et l'opinion ont été difficiles à distinguer dans le cas du corpus RestoKR. En outre, l'interprétation axiologique peut différer d'une culture à l'autre. La typologie d'évaluation que nous avons développée démontre néanmoins qu'elle représente assez fidèlement en quoi consiste l'évaluation. Si chaque langue emploie des indices linguistiques différents pour exprimer leur point de vue axiologique, cela est dû en grande partie au fait que les deux langues (coréen, français) ne sont pas apparentées. L'analyse menée sur le corpus RestoKR a ainsi permis d'appuyer la reproductibilité du module que nous proposons sur d'autres données.

6.4 Conclusion

Nous venons de décrire l'application de notre approche à deux corpus : l'un relevant de l'hôtellerie ; l'autre, rédigé en coréen et portant sur la restauration. L'annotation manuelle ainsi que les analyses qualitatives et quantitatives sur les deux corpus ont permis de vérifier la généralisabilité de notre modèle conceptuel proposé. En ce qui concerne le module de la détection automatique, son évaluation sur le corpus HotelFR a donné le meilleur résultat, avec une moyenne pondérée de F-mesure égale à 0,81.

Nos conclusions principales se résument en deux points. Premièrement, le modèle que nous avons développé en nous appuyant sur le corpus RestoFR peut être généralisé et reproduit, soit sur un autre domaine proche dont l'évaluation concerne l'expérience dans un lieu (au restaurant ou à l'hôtel), soit dans une autre langue (coréen). Deuxièmement, il semble que l'évaluation est exprimée de façon systématique, comme l'ont notamment confirmé les analyses fondées sur les mots-clés, l'AFC des catégories morphosyntaxiques et les positions de celles-ci.

Nos annotations ont particulièrement mis en lumière les spécificités de la description. Malgré sa nature objective, cette dernière montre certaines particularités du domaine étudié : en ce qui concerne l'hôtellerie, la description énumère les équipements, les installations et la localisation ; pour ce qui est de la restauration, elle contient davantage d'informations factuelles possédant des caractéristiques anecdotiques. De plus, la distinction entre les opinions (positive ou négative) et la description peut s'avérer quelque peu délicate, car si certains acceptent la description pour ce qu'elle est, d'autres vont plus loin et la perçoivent comme un état souhaitable ou non pour eux. Par ailleurs, contrairement au corpus RestoFR, RestoKR n'est pas nécessairement orienté vers la polarité positive et présente une répartition des catégories relativement homogène. Cela montre que l'évaluation dépend d'un système de normes et de valeurs propre à chaque locuteur et à chaque culture.

De nos jours, de nombreuses expériences vécues dans différents domaines sont évaluées en ligne. De plus, les évaluations sont rédigées par des clients issus de divers milieux culturels et qui parlent différentes langues. Une étude approfondie sur les caractéristiques de ces évaluations présente donc un intérêt considérable puisqu'elle permet de mieux comprendre ces particularités d'évaluation.

Quatrième partie

CONCLUSION ET PERSPECTIVES

Chapitre 7

Conclusion et perspectives

7.1 Contributions

L'étude présentée dans cette thèse consiste en une analyse de l'évaluation portant sur des restaurants et publiée en ligne par des clients. Elle s'inscrit dans le domaine du TAL et relève de la fouille d'opinions, devenue essentielle dans de nombreux champs d'application tels que la veille d'opinions et l'analyse de tendances ou de marchés. En effet, le volume massif de données textuelles publiées sur Internet incite aujourd'hui les entreprises à considérer avec attention les évaluations qui circulent et influencent leur réputation et leurs ventes. Conscients de cette nécessité, les professionnels se heurtent cependant à des difficultés considérables dans l'analyse de l'évaluation, en raison de l'étendue et de la complexité de cette notion.

L'évaluation est étudiée dans de nombreux travaux (Lark *et al.* 2016 ; Poria *et al.* 2016) suivant la polarité qu'elle exprime. Notre ambition était de mieux répondre aux enjeux que pose la fouille d'opinions par le biais d'une étude du fonctionnement de l'évaluation.

Nous avons montré que l'évaluation des restaurants visités par les clients ne se limite pas à des avis positifs ou négatifs, mais peut avoir d'autres fonctions : le visiteur laisse son avis pour donner ses opinions, faire des suggestions, exprimer ses intentions ou décrire son expérience. Toutes ces informations reflètent les différents objectifs que poursuit l'internaute lorsqu'il laisse une évaluation. L'hétérogénéité de ces objectifs et fonctions s'explique aussi par la nature double des destinataires visés par les évaluations en ligne : une entité évaluée, un lieu visité (un restaurateur, dans notre corpus), d'une part, et client potentiel d'autre part. Par ailleurs,

nous avons montré que l'évaluation ne décrit pas seulement l'expérience ponctuelle, mais se situe sur une échelle temporelle.

Nous avons également développé un module de détection automatique de chacune de ces fonctions et proposé leur étude quantitative dans le corpus.

Enfin, nous avons montré que l'approche utilisée est reproductible car elle peut être appliquée à d'autres données, telles que le corpus composé d'évaluations portant sur un autre sujet ou sur le corpus provenant d'une autre langue.

Nous reprendrons et synthétiserons à présent les contributions principales composant cette thèse de doctorat, puis nous évoquerons les perspectives envisagées.

7.1.1 *De l'élaboration du modèle à la détection automatique*

7.1.1.1 Proposition d'une typologie

L'observation manuelle du corpus a montré que les évaluations en ligne peuvent avoir plusieurs fonctions (expression d'une opinion positive/négative/mixte, suggestion, intention et description). Suite à ces observations, une modélisation de l'évaluation a été proposée sous forme de typologie comprenant six catégories (POS_OPINION, NEG_OPINION, MIX_OPINION, INTENTION, SUGGESTION et DESCRIPTION).

Dans le cadre de cette typologie, l'OPINION est définie selon sa dimension axiologique : il s'agit des évaluations de la valeur d'une entité par la révélation de sentiments. Cette catégorie peut être caractérisée en fonction de trois polarités : positive, négative et mixte. La polarité est déterminée sur un axe dont le pôle est bon ou mauvais, agréable ou désagréable, satisfaisant ou insatisfaisant, souhaitable ou à éviter. La SUGGESTION, quant à elle, s'adresse à la fois aux restaurateurs, à qui elle fait prendre conscience des problèmes de leur établissement, et aux clients potentiels (il s'agit alors de conseils). L'INTENTION correspond au souhait d'un client de retourner ou non dans un restaurant. La DESCRIPTION, souvent négligée dans la fouille d'opinions, désigne les informations factuelles associées à l'expérience vécue et permettant aux lecteurs d'en découvrir

l'arrière-plan. Ces informations permettent de replacer l'expérience individuelle dans un contexte plus large : avant, pendant et après la visite du restaurant.

L'annotation manuelle du corpus a été effectuée sur la base de la typologie proposée, ce qui nous a permis d'explorer ce dernier d'un point de vue quantitatif et de nourrir un système d'apprentissage automatique.

7.1.1.2 Analyse quantitative

Une chaîne d'analyses quantitatives a été réalisée par le biais d'une méthode linguistique de corpus comprenant l'étude des mots clefs et des n-grammes ayant des scores TF-IDF élevés, la distribution des parties du discours et leur position au sein d'un avis. Pour chacune des catégories d'évaluation, nous avons constaté des mots récurrents. L'OPINION est marquée par le lexique évaluatif ainsi que les cibles auxquelles il se rattache et apparaît à travers les syntagmes nominaux. La DESCRIPTION, l'INTENTION et la SUGGESTION sont représentées par des syntagmes verbaux permettant d'exposer l'exécution d'une action par le client lui-même (l'INTENTION) ou l'incitation à la réalisation d'une action future par le restaurateur ou les futurs clients (la SUGGESTION). Quant à la position des catégories dans un avis, l'OPINION et la DESCRIPTION figurent principalement au début de celui-ci, alors que la SUGGESTION et l'INTENTION se trouvent à la fin.

Ces analyses nous ont permis de mieux comprendre la façon dont les clients évaluent les restaurants. Elles ont apporté, par ailleurs, une vision plus précise des spécificités de chaque catégorie d'évaluation proposée, faisant de chacune d'elles une source d'information utile, tant pour les clients que pour les restaurateurs.

7.1.1.3 Détection automatique

Nous avons présenté la détection automatique des catégories d'évaluation fondée sur deux techniques d'apprentissage supervisé : l'apprentissage de surface et l'apprentissage profond. Notre objectif était d'obtenir de bonnes performances, aussi bien sur les classes minoritaires que sur les classes majoritaires. La technique CSL et le sur-échantillonnage ont permis de résoudre le problème des classes minoritaires. Cependant, étant donné que

les nouvelles instances reproduites par le sur-échantillonnage ne sont pas des données authentiques, nous avons considéré la technique CSL comme la meilleure solution.

Au cours des expériences, nous avons obtenu la meilleure moyenne pondérée de la F-mesure grâce au classifieur SVM (apprentissage de surface), donnant un score de 0,8818. La F-mesure obtenue à l'aide de LSTM (apprentissage profond) – qui a connu un véritable succès dans les applications en TAL – est inférieure au meilleur résultat, avec un score de 0,8497. Ce dernier a néanmoins été intéressant, car les caractéristiques sont apprises de manière automatique ; par conséquent, il nécessite moins de temps et d'efforts. Lorsque le meilleur classifieur (SVM) est appliqué, les moyennes pondérées de la F-mesure pour chaque catégorie sont : POS_OPINION (0,94), NEG_OPINION (0,72), MIX_OPINION (0,69), SUGGESTION (0,84), INTENTION (0,88) et DESCRIPTION (0,41). Le résultat montre ainsi que la description a été la plus difficile à détecter, ayant une tendance à être rangée dans la classe majoritaire – l'opinion positive. La mauvaise performance de cette catégorie résulte de sa nature hétérogène et de sa variation selon le profil du client, donnant lieu à un large éventail de vocabulaires et de contextes.

7.1.2 *Application de la typologie proposée*

Afin de généraliser la typologie d'évaluation proposée (POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION et DESCRIPTION), celle-ci a été appliquée à d'autres corpus comparables : un corpus relevant d'un autre domaine (l'hôtellerie) et un corpus écrit dans une autre langue (le coréen). Les analyses réalisées montrent qu'il existe des moyens systématiques relativement indépendants du domaine pour exprimer une évaluation. Le module développé et appris sur le corpus d'évaluations de restaurants peut être utilisé sur d'autres corpus comparables, comme celui comprenant les évaluations d'hôtels, sans passer par le stade d'apprentissage sur ces données. L'analyse du corpus coréen annoté portant sur des restaurants a démontré la généralisabilité de la typologie proposée, malgré une différence importante entre les deux langues (coréen, français) du point de vue linguistique et culturel.

Cependant, chaque corpus analysé présente ses propres particularités

quant aux évaluations qui le constituent. Les annotations effectuées sur ces deux nouveaux corpus ont permis de souligner les spécificités de la DESCRIPTION. En dépit de sa nature objective, elle présente certaines particularités du domaine étudié : pour ce qui est de l'hôtellerie, la DESCRIPTION énumère les équipements, les installations et l'emplacement ; en ce qui concerne la restauration, elle contient davantage d'informations factuelles possédant des caractéristiques anecdotiques. De plus, la distinction entre l'OPINION et la DESCRIPTION est parfois ambiguë, car si certaines personnes acceptent les informations factuelles pour ce qu'elles sont, d'autres cherchent à aller plus loin et les interprètent d'une manière plus subjective. De plus, à la différence du corpus français, le corpus coréen n'est pas nécessairement orienté vers une polarité positive et présente une répartition relativement homogène des catégories.

7.2 Perspectives

En premier lieu, le travail réalisé dans le cadre de cette thèse pourra être appliqué à l'ABSA (*Aspect-Based Sentiment Analysis*). En effet, ce niveau d'analyse permet de couvrir plusieurs applications, telles que l'extraction des opinions et de leurs cibles, le résumé automatique et la visualisation. Bien que nous ayons appliqué notre module sur deux corpus supplémentaires, notre travail de thèse devra également évaluer sa généralisabilité à plus grande échelle. Par ailleurs, il est important de prendre en considération le fait qu'une langue peut évoluer avec le temps, tout comme les dispositifs utilisés pour communiquer l'évaluation. Dans la présente section, nous exposerons ces perspectives plus en détails.

7.2.1 *Détection des cibles et des évaluations*

L'ajout de la détection de la cible d'évaluation pourrait améliorer les résultats du module développé. À titre d'exemple, dans le cadre de cette étude, dans la figure 7.1, la phrase est détectée comme MIX_OPINION. Pourtant, si nous ajoutons la détection de la cible, nous pouvons envisager la suppression de cette catégorie : *accueil* (cible) – *très sympathique* (POS_OPINION), *cuisine* (cible) – *décevante* (NEG_OPINION).

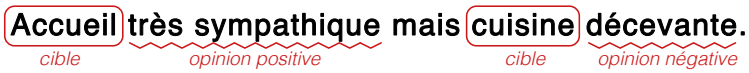


FIG. 7.1 : Exemple de la détection de la cible et de l'opinion

Les améliorations pourraient concerner la détection d'autres catégories, comme la SUGGESTION, par exemple. Pour la première phrase, dans la figure 7.2, la cible détectée sera *le magret de canard*, et la SUGGESTION portant sur cette cible sera *recommande* (figure 7.2a). Pour la deuxième phrase, la cible détectée sera *une carte*, et la SUGGESTION sur cette cible correspondra à *aurait souhaité* [...] *plus originale* (figure 7.2b).



(a) Premier exemple



(b) Deuxième exemple

FIG. 7.2 : Exemple de la détection de la cible et de la suggestion

7.2.2 *Résumé automatique*

Face à la quantité importante d'avis disponibles en ligne, il serait intéressant de réaliser un résumé automatique des évaluations recueillies pour chacune des catégories d'évaluation. Celui-ci permettrait de réduire la quantité d'informations redondantes tout en préservant les informations saillantes. Par ailleurs, il pourrait être utile de comparer la proportion d'intentions de répéter l'expérience (et donc de revenir à l'endroit visité) à celles des intentions de ne pas la répéter.

7.2.3 *Visualisation*

La visualisation des avis émis par les clients peut être intéressante dans la mesure où elle aide à explorer et à hiérarchiser une grande

quantité de données. Plus précisément, la hiérarchisation des aspects (caractéristiques) d'une entité permet, non seulement de fournir un aperçu plus structuré des évaluations, mais aussi de comparer les différents aspects du restaurant. Particulièrement grâce à la disponibilité de grandes quantités d'informations, la visualisation de ces masses de données est actuellement axée sur la présentation des tendances générales au niveau des sentiments ; à savoir, comment ceux-ci (positifs ou négatifs) évoluent dans le temps (Gamon *et al.* 2005 ; Marcus *et al.* 2011) et de quelle manière ils se propagent parmi les utilisateurs (Wu *et al.* 2014). En outre, l'interactivité de la visualisation permet de surmonter la difficulté rencontrée pour afficher d'un seul coup la totalité du grand volume de données, passant alors d'une vue globale à une observation plus détaillée par une manipulation directe, le filtrage et l'agrandissement (ou réduction) de la visualisation. Ainsi, cette dernière permet d'observer efficacement les tendances et les particularités de l'évaluation réalisée par les clients.

7.2.4 *Développement d'un outil de la fouille d'opinion*

Les clients publient sans cesse leurs évaluations sur Internet, il est donc indispensable pour les entreprises de se tenir au courant de tous les avis qui les concernent. Cependant, pour éviter d'avoir à parcourir manuellement chaque avis un par un, un outil de fouille d'opinions permet de traiter en toute rapidité les évaluations des clients et d'analyser celles-ci de manière automatique. Pour cette raison, de plus en plus d'entreprises se tournent vers des instruments d'analyse puissants et faciles à mettre en œuvre. Ainsi, les outils de fouille d'opinions peuvent aider l'entreprise à comprendre en temps réel ce que les clients pensent d'elle et de son produit (ou service) par le biais d'une analyse des avis et métadonnées (notes attribuées par les clients, localisation, date, etc.).

Divers outils sur le marché permettent de réaliser une fouille d'opinions, dont quelques-uns sont illustrés dans l'annexe B. Cependant, en ce qui concerne les données textuelles, ces outils se limitent notamment à la détection et l'évolution de la polarité (positive/négative) du sentiment, à l'extraction des mots positifs/négatifs, au nuage de mots et à la découverte de thèmes (*topic modeling*). De plus, ces outils sont fréquemment destinés à traiter les données en anglais.

Nous proposons un outil plus riche et détaillé de fouille d'opinions, comme l'illustre la figure 7.3. Les catégories d'évaluation proposées permettent de fournir des informations plus saillantes. À titre d'exemple, grâce à la suggestion détectée, les restaurateurs pourront se renseigner en temps réel sur les suggestions d'amélioration proposées par leurs clients. Par ailleurs, il est possible d'intégrer aux résultats des métadonnées telles que les notes attribuées par les clients, la localisation ou encore la date, ainsi que d'adapter l'outil aux évaluations portant sur d'autres domaines ou rédigées dans d'autres langues.

[78]Dans cette étude, l'adverbe *très* n'a pas été considéré comme un mot-outil.

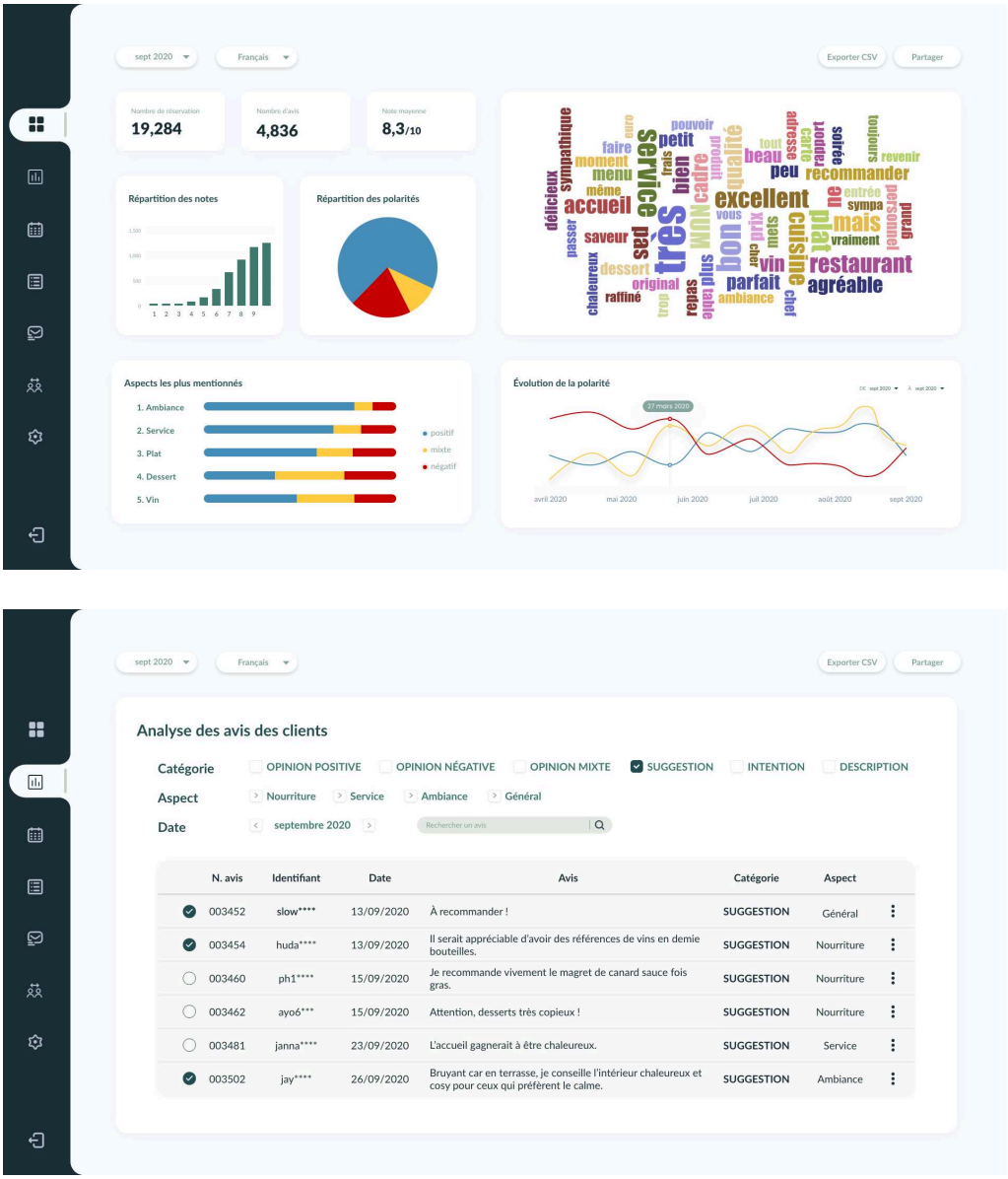


FIG. 7.3 : Exemples de l'outil que nous proposons[78]

Bien que nous nous soyons focalisée davantage sur le point de vue de fournisseurs (entreprises, restaurants, etc.), notre étude peut cibler également les clients. Étant donné la quantité importante de données disponibles en ligne, ce ne sont pas seulement les entreprises qui connaissent des difficultés à obtenir des informations bien structurées ; c'est également le cas des clients. De nos jours, il existe des sites comme Tripadvisor[79] ou Yelp[80] dont l'objectif est de fournir toutes les informations nécessaires aux clients. Le site Amazon.fr[81], par exemple, illustre les n-grammes les plus mentionnés avant de présenter les avis publiés par les clients, comme illustré dans la figure 7.4. Cependant, un simple n-grammes ne fournit pas nécessairement d'informations aux clients potentiels concernant le produit (*qualité prix* ou *rapport qualité* dans la figure 7.4). À cet égard, l'ajout d'un résumé, d'une visualisation, ou encore des n-grammes s'appuyant sur notre typologie d'évaluation semblent plus intéressants et utiles aux clients potentiels.

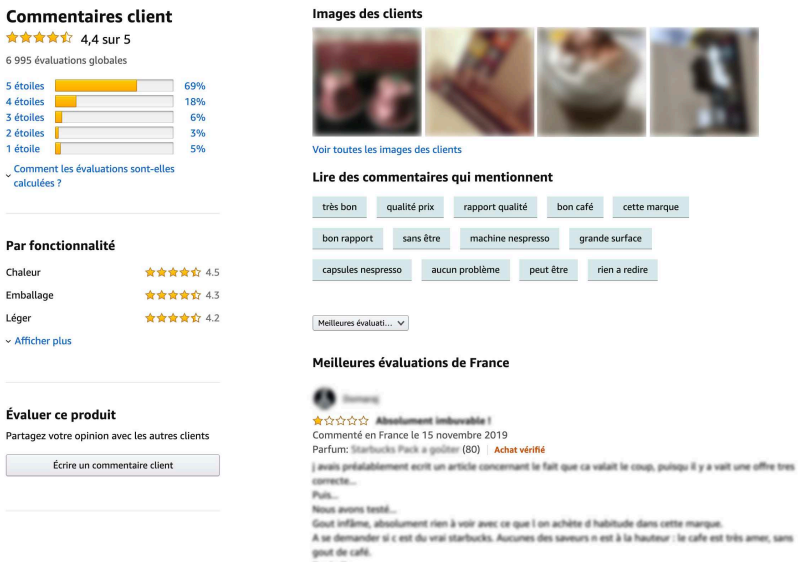


FIG. 7.4 : Exemple de la détection de la cible et de l'opinion

[79] https://www.tripadvisor.fr/ [consulté le 6 octobre 2020].
[80] https://www.yelp.fr/ [consulté le 6 octobre 2020].
[81] https://www.amazon.fr/ [consulté le 6 octobre 2020].

7.2.5 *Application à plus grande échelle*

Notre modèle d'évaluation est reproductible aux corpus comparables (liés à l'expérience vécue dans un lieu) comme les corpus portant sur l'hôtellerie, le tourisme (musées, châteaux, parcs d'attractions, festivals, etc.) et le marketing (magasin, boutique, etc.). Pour ce qui est du domaine du tourisme, par exemple, cette étude peut apporter des informations sur l'expérience et la satisfaction des clients. Une expérience positive amènera un nombre croissant de voyageurs satisfaits, qui souhaiteront revenir et recommanderont la destination à d'autres. Considérant que de nombreuses personnes sont de nos jours influencées par les avis publiés sur Internet, le modèle que nous proposons permettra, d'une part de découvrir les divers motifs et tendances qui sous-tendent la sélection des destinations touristiques, et d'autre part de prendre en considération les facteurs susceptibles d'améliorer le niveau de satisfaction des touristes.

Les avis sont par ailleurs publiés par des clients issus de divers milieux culturels et parlant différentes langues. Étant donné que chaque langue et chaque culture possède un système de normes et de valeurs qui lui est propre, l'étude sur des corpus d'autres langues permettra de mettre en évidence ce qui est spécifique à chaque culture et chaque langue. Si les avis rédigés présentent toujours les différentes catégories d'évaluation quelle que soit la langue, la différence réside, non seulement dans la façon d'exprimer l'évaluation, mais aussi dans les valeurs privilégiées par chaque culture et dans chaque langue. Bien que nous ayons étudié des corpus en deux langues (le français et le coréen), la généralisation devrait être évaluée sur une plus grande variété de langues pour compléter et valider les études menées dans cette thèse. De plus, en raison de la difficulté de prétraitement du corpus coréen, nous n'avons pas pu mener d'analyse quantitative et développer le module de détection automatique sur celui-ci, ce qu'il serait intéressant d'effectuer.

**

ANNEXES

Annexe A

Études en TAL

A.1 Synthèse des tâches proposées aux campagnes d'évaluation

A.1.1 *Défi Fouille de Textes (DEFT)*

Tab. A.1 : Tâches proposées au DEFT

Campagne	Langue	Tâche
DEFT 07 (Grouin *et al.* 2007)	français	Détection de l'opinion exprimée dans un texte de retranscription de débats parlementaires (projets de loi relatifs à l'énergie)
DEFT 09 (Grouin *et al.* 2009)	anglais, français, italien	Fouille d'opinion (objectif/subjectif) en corpus multilingues (journaux et débats européens)
DEFT 15 (Hamon *et al.* 2015)	français	Fouille d'opinion, de sentiment et d'émotion dans des messages postés sur Twitter
DEFT 17 (Benamara *et al.* 2017a)	français	Fouille d'opinion dans des messages postés sur Twitter
DEFT 18 (Paroubek *et al.* 2018)	français	Recherche d'information et analyse de sentiments dans des tweets sur les transports en Île-de-France

A.1.2 *Semantic Evaluation (SemEval)*

TAB. A.2 : Tâches proposées au SemEval

Campagne	Langue	Tâche
SemEval-2013 Tâche 2 (Nakov *et al.* 2013)	anglais	Analyse de sentiments dans des messages postés sur Twitter
SemEval-2014 Tâche 4 (Pontiki *et al.* 2014)	anglais	Analyse de sentiments dans des avis sur les ordinateurs portables et les restaurants
SemEval-2014 Tâche 9 (Rosenthal *et al.* 2014)	anglais	Analyse de sentiments dans des messages postés sur Twitter
SemEval-2015 Tâche 10 (Rosenthal *et al.* 2015)	Anglais	Analyse de sentiments dans des messages postés sur Twitter
SemEval-2015 Tâche 11 (Ghosh *et al.* 2015)	anglais	Analyse de sentiments dans des messages figuratifs postés sur Twitter
SemEval-2015 Tâche 12 (Pontiki *et al.* 2015)	anglais	Analyse de sentiments dans des avis sur les ordinateurs portables et les restaurants
SemEval-2016 Tâche 4 (Nakov *et al.* 2016)	anglais	Analyse de sentiments dans des messages postés sur Twitter
SemEval-2016 Tâche 5 (Pontiki *et al.* 2016)	anglais, arabe, chinois, espagnol, français, néerlandais, russe, turc	Analyse de sentiments dans des avis sur divers objets (ordinateurs portables, restaurants, hôtels, appareils électroniques)
SemEval-2017 Tâche 4 (Rosenthal *et al.* 2017)	anglais, arabe	Analyse de sentiments dans des messages postés sur Twitter

Campagne	Langue	Tâche
SemEval-2017 Tâche 5 (Cortis *et al.* 2017)	anglais	Analyse fine des sentiments dans les microblogs et dans la presse du secteur financier
SemEval-2018 Tâche 1 (Mohammad *et al.* 2018)	anglais, arabe, espagnol	Analyse d'affects dans des messages postés sur Twitter
SemEval-2019 Tâche 5 (Basile *et al.* 2019)	anglais, espagnol	Détection des discours multilingues de haine contre les immigrants et les femmes sur Twitter
SemEval-2020 Tâche 9 (Patwa *et al.* 2020)	Spanglish*, hinglish**	Analyse de sentiments dans des messages multilingues (code-mixing) postés sur Internet

* Spanglish : une combinaison d'anglais et d'espagnol
** Hinglish : une combinaison d'anglais et d'hindi

A.2 Synthèse des études dans la fouille d'opinions

Tab. A.3 : Synthèse des études dans la fouille d'opinions

Étude	Langue	Domaine	Niveau	Technique
Pang *et al.* (2002)	anglais	film	document	Naïve Bayes, SVM
Turney (2002)	anglais	produit, banque, film, destination de voyage	document	non supervisé
Mullen et Collier (2004)	anglais	film	document	SVM
Kennedy et Inkpen (2006)	anglais	film	document	non supervisé
Taboada *et al.* (2006), Taboada *et al.* (2011)	anglais	produit, hôtel, film, musique	document	non supervisé
Bestgen (2008), Bestgen (2013)	anglais	film	document	non supervisé
Poirier *et al.* (2009)	anglais	film	document	non supervisé, SVM, Naïve Bayes
Vernier et Monceaux (2010)	français	développement durable, film, politique etc.	document	non supervisé
Yessenalina *et al.* (2010)	anglais	film, débat politique	document	SVM

Étude	Langue	Domaine	Niveau	Technique
Marchand (2013)	anglais, français	jeux, film, livre, théâtre	document	SVM
Moraes *et al.* (2013)	anglais	film, produit	document	SVM, Naïve Bayes
Vincent et Winterstein (2013)	français	hôtel, film, roman	document	Logistic Regression, SVM
Eensoo et Valette (2015)	français	forum	document	Naïve Bayes, SVM
Tang *et al.* (2015)	anglais	film, restaurant	document	CNN
Chen *et al.* (2016)	anglais	film, restaurant	document	LSTM
Ankit et Saleena (2018)	anglais	Twitter	document	Naïve Bayes, SVM, Random Forest
Wiebe *et al.* (1999)	anglais	presse	phrase (subjectivité)	Naïve Bayes
Wiebe (2000)	anglais	presse	phrase (subjectivité)	non supervisé
Yu et Hatzivassiloglou (2003)	anglais	presse	phrase (subjectivité)	Naïve Bayes

Étude	Langue	Domaine	Niveau	Technique
Kim *et al.* (2004)	anglais	presse	phrase (subjectivité)	non supervisé
Wilson *et al.* (2004)	anglais	presse	phrase (subjectivité)	SVM
Pang et Lee (2005)	anglais	film	phrase (polarité)	SVM
Wiebe et Riloff (2005)	anglais	presse	phrase (subjectivité)	Naïve Bayes
Barbosa et Feng (2010)	anglais	Twitter	phrase (subjectivité)	SVM
Benamara *et al.* (2011)	anglais	film	phrase (subjectivité)	SVM
Kalchbrenner *et al.* (2014)	anglais	film, Twitter	phrase (polarité)	CNN
Qian *et al.* (2017)	anglais	film	phrase (polarité)	LSTM
Hew *et al.* (2020)	anglais	éducation	phrase (polarité)	SVM, Naïve Bayes Logistic Regression
Hu et Liu (2004)	anglais	produit	aspect	non supervisé

Étude	Langue	Domaine	Niveau	Technique
Ding *et al.* (2008)	anglais	produit	aspect	non supervisé
Jiang *et al.* (2011)	anglais	Twitter	aspect	SVM
Lark *et al.* (2015)	français	Twitter	aspect	SVM
Severyn et Moschitti (2015)	anglais	Twitter	aspect	CNN
Lark *et al.* (2016)	français	banque, restaurant, produit, parc	aspect	SVM
Poria *et al.* (2016)	anglais	produit, restaurant	aspect	CNN
Wang *et al.* (2016)	anglais	restaurant	aspect	LSTM
Bhoi et Joshi (2018)	anglais	Twitter	aspect	Naïve Bayes, SVM, Random Forest, XGBoost, LSTM
Pecore et Villaneau (2018), Villaneau *et al.* (2019)	français	film, livre	aspect	SVM, Random Forest

Annexe B

Exemples des outils de la fouille d'opinions

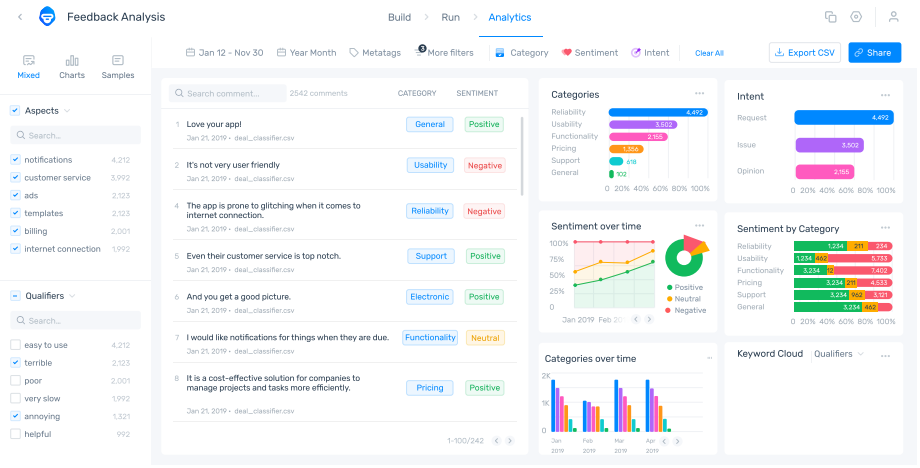


FIG. B.1 : Capture d'écran de l'outil Monkey Learn[82]

[82] https://monkeylearn.com/ [consulté le 6 octobre 2020].
[83] https://www.lexalytics.com/ssv [consulté le 6 octobre 2020].
[84] https://www.social-searcher.com/ [consulté le 6 octobre 2020].

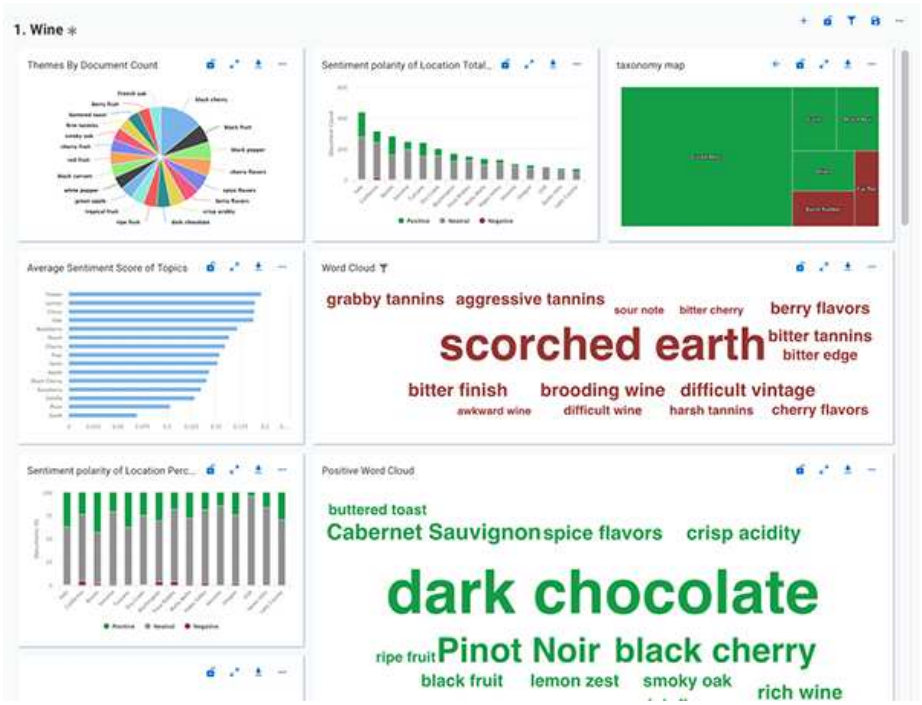


FIG. B.2 : Capture d'écran de l'outil Lexalytics[83]

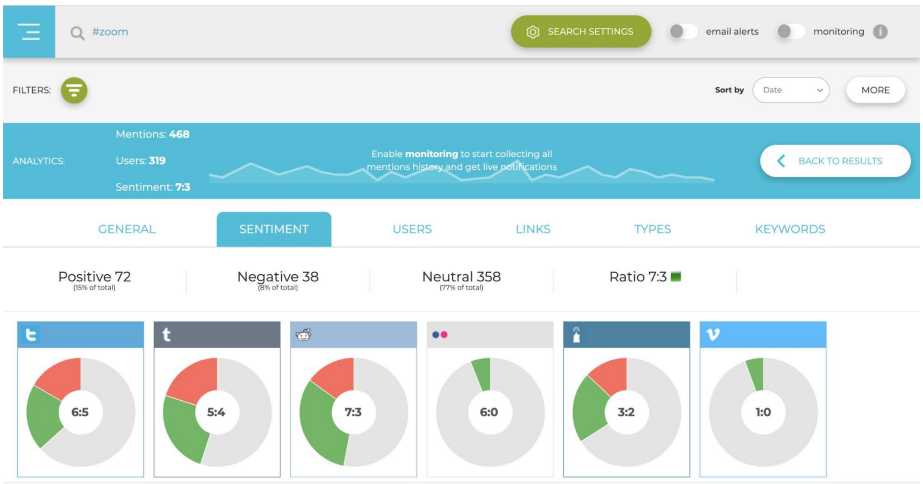


FIG. B.3 : Capture d'écran de l'outil Social Searcher[84]

Bibliographie

Abdaoui, Amine, Jérôme Azé, Sandra Bringay et Pascal Poncelet (2017), "Feel : a French Expanded Emotion Lexicon", *Language Resources and Evaluation*, 51, 3, p. 833-855.

Aggarwal, Charu C. et ChengXiang Zhai (2012), *Mining text data*, Springer.

Amiot, Dany (2012), "De la localisation à l'évaluation : des verbes préfixés évaluatifs au sens bien particulier", *CORELA - COgnition, REprésentation, LAngage*.

Amiot, Dany et Dejan Stosic (2015), "Morphologie aspectuelle et évaluative en français et en serbe", *Lexique*, 22, p. 111-142.

Ananko, Tatiana (2017), "The category of evaluation in political discourse", *Advanced education*, 4, 8, p. 128-137.

Ankit et Nabizath Saleena (2018), "An Ensemble Classification System for Twitter Sentiment Analysis", *Procedia Computer Science*, 132, p. 937-946.

Anthony, Laurence (2019), "AntConc (version 3.5. 8)[computer software]", *Tokyo, Japan : Waseda University*.

Arutyunova, N.D. (1988), *Tipy iazykovyh znacheniiy. Otsenka. Sobytie. Fakt [Types of Language Meanings. Evaluation. Event. Fact]*, Nauka.

Asher, Nicholas et Alex Lascarides (2003), *Logics of conversation*, Cambridge University Press.

Aull, Laura (2017), "Corpus analysis of argumentative versus explanatory discourse in writing task genres", *Journal of Writing Analytics*, 1.

Austin, John Langshaw (1975), *How to Do Things with Words*, Oxford university press, t. 88.

Babin, Barry J., William R. Darden et Mitch Griffin (1994), "Work and/or fun : measuring hedonic and utilitarian shopping value", *Journal of consumer research*, 20, 4, p. 644-656.

Baccianella, Stefano, Andrea Esuli et Fabrizio Sebastiani (2010), "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining", in *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 2010, t. 10, p. 2200-2204.

Baccini, Alain (2010), *Statistique Descriptive Multidimensionnelle (pour les nuls)*.

Baker, Paul (2010), *Sociolinguistics and corpus linguistics*, Edinburgh University Press.

Bakhtin, Mikhail Mikhaïlovich (1978), *Esthétique et théorie du roman*, trad. par Daria Olivier, Gallimard.

Bakhtin, Mikhail Mikhaïlovich, Valentin N. Voloshinov, Roman Jakobson et Marina Yaguello (1979), *Le Marxisme Et la Philosophie du Langage Essai d'Application de la Méthode Sociologique En Linguistique*, trad. par Marina Yaguello.

Baranes, Marion et Benoît Sagot (2014), "Normalisation de textes par analogie : le cas des mots inconnus", in *Actes de la conférence TALN 2014*, p. 137-148.

Barbaray, Christian (2016), *Satisfaction, fidélité et expérience client*, Dunod.

Barbosa, Luciano et Junlan Feng (2010), "Robust Sentiment Detection on Twitter from Biased and Noisy Data", in *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, p. 36-44.

Baron, Alistair, Paul Rayson et Dawn Archer (2009), "Word frequency and key word statistics in corpus linguistics", *Anglistik : International Journal of English Studies*, 20, 1, p. 41-67.

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso et Manuela Sanguinetti (2019), "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter", in *Proceedings of the 13th International Workshop on Semantic Evaluation*, p. 54-63.

Batson, C. Daniel, Laura L. Shaw et Kathryn C. Oleson (1992), "Differentiating affect, mood, and emotion : Toward functionally based conceptual distinctions", in *Review of personality and social psychology*, Sage Publications, t. 13, p. 294-326.

Bauer, L. (1997), "Evaluative Morphology : In Search of Universals", *Studies in Language*, 21, p. 533-575.

Bednarek, Monika (2009), "Dimensions of evaluation : Cognitive and linguistic perspectives", *Pragmatics & Cognition*, 17, 1, p. 146-175.

Ben-Ze'ev, Aaron et Aharon Ben-Ze'ev (2001), *The subtlety of emotions*, MIT press.

Benamara, Farah, Nicholas Asher, Yvette Yannick Mathieu, Vladimir Popescu et Baptiste Chardon (2016), "Evaluation in discourse : A corpus-based study", *Dialogue & Discourse*, 7, 1, p. 1-49.

Benamara, Farah, Baptiste Chardon, Yannick Mathieu et Vladimir Popescu (2011), "Towards Context-Based Subjectivity Analysis", in *Proceedings of 5th International Joint Conference on Natural Language Processing*, p. 1180-1188.

Benamara, Farah, Cyril Grouin, Jihen Karoui, Véronique Moriceau et Isabelle Robba (2017a), "Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du Défi Fouille de Textes DEFT2017", in *Actes de l'atelier DÉfi Fouille de Textes*, p. 1-12.

Benamara, Farah, Maite Taboada et Yannick Mathieu (2017b), "Evaluative language beyond bags of words : Linguistic insights and computational applications", *Computational Linguistics*, 43, 1, p. 201-264.

Bensoussan, Bernard (2009), "Qualifier l'habitation : déterminants socio-culturels, propriétés de l'habitat, lexique de l'immobilier", *Mots : les langages du politique*, 91.

Benveniste, Emile (1966), *Problèmes de linguistique générale*, Gallimard, t. 2.

Benzécri, Jean-Paul *et al.* (1973), *L'analyse des données*, Dunod, t. 2.

Bestgen, Yves (2008), "Building Affective Lexicons from Specific Corpora for Automatic Sentiment Analysis", in *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

– (2013), "Construction automatique de ressources lexicales pour la fouille d'opinion", in *Actes de CORIA 2013 : Conférence en recherche d'information et applications"*, p. 101-116.

Bezuhla, L.R. et I.O. Romanchenko (2013), "Linhvoprahmatyka dyskryminatsiyi u publitsystychnomu dyskursi [Linguopragmatics of Discrimination in Publicistic Discourse]", *FOP Lysenko IB Publ.*

Bhatia, Parminder, Robert Guthrie et Jacob Eisenstein (2016), "Morphological priors for probabilistic neural word embeddings", *arXiv preprint arXiv :1608.01056*.

Bhatia, Surbhi, Poonam Chaudhary et Nilanjan Dey (2020), *Opinion Mining in Information Retrieval*, Springer.

Bhoi, Amlaan et Sandeep Joshi (2018), "Various Approaches to Aspect-based Sentiment Analysis", *arXiv preprint arXiv :1805.01984*.

Biber, Douglas et Edward Finegan (1988), "Adverbial stance types in English", *Discourse processes*, 11, 1, p. 1-34.

– (1989), "Styles of stance in English : Lexical and grammatical marking of evidentiality and affect", *Text-interdisciplinary journal for the study of discourse*, 9, 1, p. 93-124.

Bitner, Mary Jo (1992), "Servicescapes : The impact of physical surroundings on customers and employees", *Journal of marketing*, 56, 2, p. 57-71.

Bock, Tim (2011), "Improving the Display of Correspondence Analysis Using Moon Plots", *International Journal of Market Research*, 53 (mai 2011), p. 307.

Body, Laurence et Christophe Tallec (2015), *L'expérience client : Le design pour innover, L'humain pour créer du lien, Le collaboratif pour accompagner le changement*, Eyrolles.

Bojanowski, Piotr, Edouard Grave, Armand Joulin et Tomas Mikolov (2017), "Enriching word vectors with subword information", *Transactions of the Association for Computational Linguistics*, 5, p. 135-146.

Bratman, Michael (1987), *Intention, Plans, and Practical Reason*, Harvard University Press.

– (1990), "Dretske's desires", *Philosophy and Phenomenological Research*, 50, 4, p. 795-800.

Bréchon, Pierre (2012), "Les valeurs des Français", in *Os novos comportamentos sociais e individuais*.

Brezina, Vaclav (2014), *Effect sizes in corpus linguistics : keywords, collocations and diachronic comparison*, présenté à la conférence International Computer Archive of Modern and Medieval English (ICAME) 2014.

Brezina, Vaclav, Matt Timperley et Tony McEnery (2018), *LancsBox v. 4. x [computer software]*.

Broad, Charlie Dunbar (1954), "Emotion and sentiment", *The Journal of Aesthetics and Art Criticism*, 13, 2, p. 203-214.

Brownlee, Jason (2017), *Deep Learning for Natural Language Processing : Develop Deep Learning Models for your Natural Language Problems*, Machine Learning Mastery.

Brun, Caroline et Caroline Hagège (2013), "Suggestion Mining : Detecting Suggestions for Improvement in Users' Comments", *Research in Computing Science*, 70, p. 199-209.

Buvet, Pierre-André (2010), "Description des prédicats du premier ordre pour la compréhension automatique des textes", *Les tables-La grammaire du français par le menu : Mélanges en hommage à Christian Leclère*, 6, p. 37-48.

Carlos, Cohan Sujay et Madhulika Yalamanchi (2012), "Intention Analysis for Sales, Marketing and Customer Service", in *Proceedings of the 2012 International Conference on Computational Linguistics : Demonstration Papers*, p. 33-40.

Carreiras, Manuel, Morton Ann Gernsbacher et Victor Villa (1995), "The advantage of first mention in Spanish", *Psychonomic Bulletin & Review*, 2, 1, p. 124-129.

Cattell, Raymond B. (1940), "Sentiment or attitude ? The core of a terminology problem in personality research", *Character & Personality ; A Quarterly for Psychodiagnostic & Allied Studies*, p. 6-17.

Cha, Seong-Soo, Cheol Park et Xiaowu Wang (2019), "A cross-national study on restaurant attributes between Korea and China", *International Journal of Culture, Tourism and Hospitality Research*.

Chafe, Wallace (1986), *Evidentiality : The Linguistic Coding of Epistemology (Advances in Discourse Processes)*, Ablex Publishing, t. 20.

Charaudeau, Patrick (1992), *Grammaire du sens et de l'expression*, Hachette.

Chatar-Moumni, Nizha (2013), "L'expression verbale des émotions : présentation", *Langue française*, 4, p. 3-11.

Cheminée, Pascale (2006), "« Vous avez dit clair » ? Le lexique des pianistes, entre sens commun et terminologie", *Cahiers du LCPE : Dénomination, désignation et catégories*, 7, p. 39-54.

Chen, Huimin, Maosong Sun, Cunchao Tu, Yankai Lin et Zhiyuan Liu (2016), "Neural Sentiment Classification with User and Product Attention", in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 1650-1659.

Chen, Tianqi et Carlos Guestrin (2016), "Xgboost : A scalable tree boosting system", in *Proceedings of the 22nd ACM SigKDD International Conference on Knowledge Discovery and Data Mining*, p. 785-794.

Chen, Zhiyuan, Bing Liu, Meichun Hsu, Malu Castellanos et Riddhiman Ghosh (2013), "Identifying Intention Posts in Discussion Forums", in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1041-1050.

Chesley, Paula, Bruce Vincent, Li Xu et Rohini Srihari (2006), "Using Verbs and Adjectives to Automatically Classify Blog Sentiment", in *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*.

Chollet, François *et al.* (2015), *Keras*, https://github.com/fchollet/keras.

Clark, Lee Anna et David Watson (2016), "Constructing validity : Basic issues in objective scale development", *Psychological Assessment*, 7, 3, p. 309-319.

Cohen, Ayala (1980), "On the graphical display of the significant components in two-way contingency tables", *Communications in statistics-Theory and methods*, 9, 10, p. 1025-1041.

Cohen, Jacob (1960), "A coefficient of agreement for nominal scales", *Educational and psychological measurement*, 20, 1, p. 37-46.

Condoravdi, Cleo et Sven Lauer (2012), "Imperatives : Meaning and illocutionary force", *Empirical issues in syntax and semantics*, 9, p. 37-58.

Cortis, Keith, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh et Brian Davis (2017), "SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News", in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 519-535.

Creswell, Cassandre, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi et Bonnie Webber (2003), *Penn discourse treebank : Building a large scale annotated corpus encoding dltag-based discourse structure and discourse relations*.

Culioli, Antoine (1980), "Valeurs aspectuelles et opérations énonciatives : l'aoristique", *Recherches linguistiques*, 5, p. 182-193.

– (1985), *Notes du séminaire de DEA, 1983-1984*, Université de Paris VII - Département de recherches linguistiques.

Dal, Georgette (1997), *Grammaire du suffixe -et(te)*, Didier érudition.

Das, Sanjiv et Mike Chen (2001), "Yahoo ! for Amazon : Extracting market sentiment from stock message boards", in *Proceedings of the Asia Pacific Finance Association Annual Conference*.

Dave, Kushal, Steve Lawrence et David M. Pennock (2003), "Mining the peanut gallery : Opinion extraction and semantic classification of product reviews", in *Proceedings of the 12th International Conference on World Wide Web*, p. 519-528.

Desagulier, Guillaume (2017), *Corpus Linguistics and Statistics with R*, Springer.

– (2020), "Multivariate Exploratory Approaches", To appear in In Magali Paquot & Stefan Th. Gries (eds.). Practical handbook of corpus linguistics. Berlin & New York : Springer.

Ding, Xiao, Ting Liu, Junwen Duan et Jian-Yun Nie (2015), "Mining User Consumption Intention from Social Media Using Domain Adaptive Convolutional Neural Network", in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, t. 15, p. 2389-2395.

Ding, Xiaowen, Bing Liu et Philip S. Yu (2008), "A Holistic Lexicon-Based Approach to Opinion Mining", in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, p. 231-240.

Domingos, Pedro (2012), "A Few Useful Things to Know about Machine Learning", *Communications of the ACM*, 55, 10, p. 78-87.

Dressler, Wolfgang U. et Lavinia M. Barbaresi (2011), *Morphopragmatics : Diminutives and intensifiers in Italian, German, and other languages*, Walter de Gruyter, t. 76.

Drouin, Patrick (2003), "Term extraction using non-technical corpora as a point of leverage", *Terminology*, 9, 1, p. 99-115.

Dryer, Matthew S. (2013), "Prefixing vs. Suffixing in Inflectional Morphology", in *The World Atlas of Language Structures Online*, sous la dir. de Matthew S. Dryer et Martin Haspelmath, Max Planck Institute for Evolutionary Anthropology.

Dunning, Ted E. (1993), "Accurate methods for the statistics of surprise and coincidence", *Computational linguistics*, 19, 1, p. 61-74.

Eensoo, Egle et Mathieu Valette (2015), "Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité", in *22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015)*.

Efthymiou, Angeliki (2017), "A morphosemantic investigation of diminutive verbs in French and Modern Greek", in *Mediterranean Morphology Meetings*, t. 11, p. 26-35.

Eshkol-Taravella, Iris (2015), *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral*, Habilitation à diriger des recherches, Université d'Orléans.

Fauconnier, Jean-Philippe (2015), *French Word Embeddings*, http://fauconnier.github.io.

Flamein, Hélène (2018), *Etude de la perception d'une ville. Repérage automatique, analyse et visualisation*, thèse de doct., Université d'Orléans.

Fleiss, Joseph L., Bruce Levin, Myunghee Cho Paik *et al.* (1981), "The measurement of interrater agreement", *Statistical methods for rates and proportions*, 2, 212-236, p. 22-23.

Fradin, Bernard (2003), "Le traitement de la suffixation en-et", *Langages*, p. 51-77.

Friendly, Michael (1992), "Graphical methods for categorical data", in *Proceedings of the 17th Annual SAS Users Group International (SUGI) Conference*, p. 1367-1373.

Gabrielatos, Costas (2018), "Keyness Analysis : nature, metrics and techniques", in *Corpus Approaches to Discourse*, Routledge, p. 225-258.

Gabrielatos, Costas et Anna Marchi (2012), *Keyness : Appropriate metrics and practical issues*, Paper presented at Corpus-assisted Discourse Studies International Conference.

Galatanu, Olga (2003), "La sémantique des possibles argumentatifs et ses enjeux pour l'analyse de discours", in *El texto como encrucijada : estudios franceses y francófonos*, Universidad de La Rioja, p. 213-226.

Gamon, Michael, Anthony Aue, Simon Corston-Oliver et Eric Ringger (2005), "Pulse : Mining Customer Opinions from Free Text", in *Proceedings of IDA-05, the 6th International Symposium on Intelligent Data Analysis*, t. 3646, p. 121-132.

Ganu, Gayatree, Noemie Elhadad et Amélie Marian (2009), "Beyond the Stars : Improving Rating Predictions using Review Text Content", in *12th International Workshop on the Web and Databases, WebDB 2009*.

Gardin, Pierre (2009), "Application de la théorie de l'Appraisal à l'analyse d'opinion", *MajecSTIC 2009*.

Georgakopoulou, Alexandra (2006), "Postscript : Computer-mediated communication in sociolinguistics", *Journal of sociolinguistics*, 10, 4, p. 548-557.

Gernsbacher, Morton Ann (1990), *Language Comprehension as Structure Building*, Psychology Press.

Ghosh, Aniruddha, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden et Antonio Reyes (2015), "Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter", in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 470-478.

Giboreau, Agnès et Laurence Body (2012), *Marketing sensoriel : Une démarche globale pour les produits et les services*, Vuibert.

Goldberg, Andrew B, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson et Xiaojin Zhu (2009), "May All Your Wishes Come True : A Study of Wishes and How to Recognize Them", in *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 263-271.

Gordon, Steven L. (1991), "The socialization of children's emotions : emotional culture, competence, and exposure", in *Children's understanding of emotion*, CUP Archive, p. 319.

Gosselin, Laurent (2010), *La validation des représentations. Les modalités en français*.

– (2017), "Les modalités appréciatives et axiologiques. Sémantique des jugements de valeur", *Cahiers de lexicologie : Revue internationale de lexicologie et lexicographie*, 111, p. 97-119.

– (2018), "Quand nommer, c'est juger. Les jugements de valeur internes aux noms d'humains", *Les noms d'humains–théorie, méthodologie, classification*, p. 44-101.

Grandi, Nicola (2002), *Morfologie in contatto*, Francoangelli.

Grandi, Nicola et Lívia Körtvélyessy (2015), "Introduction : Why evaluative morphology", in *Edinburgh handbook of evaluative morphology*, Edinburgh University Press Edinburgh, p. 3-20.

Grandi, Nicola et Fabio Montermini (2003), "Prefix-suffix neutrality in evaluative morphology", in *Mediterranean Morphology Meetings*, t. 4, p. 143-156.

Grefenstette, Gregory, Yan Qu, James G. Shanahan et David A. Evans (2004), "Coupling niche browsers and affect analysis for an opinion mining application", in *Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO)*, p. 186-194.

Grice, Herbert P. (1975), *Logic and conversation*.

Gries, Stefan Thomas (2009), *Quantitative Corpus Linguistics with R : A Practical Introduction (1st edition)*, Routledge.

Grisot, Cristina (2019), "Les domaines linguistiques de la subjectivité du locuteur : étude empirique avec données de corpus", *Travaux neuchâtelois de linguistique*, 71, p. 19-43.

Gross, Maurice (1995), "Une grammaire locale de l'expression des sentiments", *Langue française*, p. 70-87.

Grouin, Cyril, Béatrice Arnulphy, Jean-Baptiste Berthelin, Sarra El Ayari, Anne García-Fernandez, Arnaud Grappy, Martine Hurault-Plantet, Patrick Paroubek, Isabelle Robba et Pierre Zweigenbaum (2009), "Présentation de l'édition 2009 du DÉfi Fouille de Textes (DEFT'09)", in *Actes du cinquième DÉfi Fouille de Textes*, p. 35-50.

Grouin, Cyril, Jean-Baptiste Berthelin, Sarra El Ayari, Thomas Heitz, Martine Hurault-Plantet, Michele Jardino, Zohra Khalis et Michel Lastes

(2007), "Présentation de deft'07 (défi fouille de textes)", in *Actes du troisième DÉfi Fouille de Textes*, p. 3-10.

Ha, Jooyeon et SooCheong Shawn Jang (2010), "Perceived values, satisfaction, and behavioral intentions : The role of familiarity in Korean restaurants", *International Journal of Hospitality Management*, 29, 1, p. 2-13.

Halliday, Michael A. K. (1970), "Language structure and language function", *New horizons in linguistics*, 1, p. 140-165.

Hamilton, William L., Kevin Clark, Jure Leskovec et Dan Jurafsky (2016), "Inducing domain-specific sentiment lexicons from unlabeled corpora", in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, t. 2016, p. 595-605.

Hamon, Thierry, Amel Fraisse, Patrick Paroubek, Pierre Zweigenbaum et Cyril Grouin (2015), "Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (DEFT)", in *Actes de DÉfi Fouille de Textes*.

Hardie, Andrew (2014), "Log ratio : An informal introduction", *ESRC Centre for Corpus Approaches to Social Science (CASS)*.

Hartmann, Christina, Jing Shi, Alice Giusto et Michael Siegrist (2015), "The psychology of eating insects : A cross-cultural comparison between Germany and China", *Food quality and preference*, 44, p. 148-156.

Heinich, Nathalie (2017), *Des valeurs. Une approche sociologique*, Gallimard.

Hemmatian, Fatemeh et Mohammad Karim Sohrabi (2019), "A survey on classification techniques for opinion mining and sentiment analysis", *Artificial Intelligence Review*, p. 1-51.

Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh et Dwayne D. Gremler (2004), "Electronic word-of-mouth via consumer-opinion platforms : what motivates consumers to articulate themselves on the internet ?", *Journal of interactive marketing*, 18, 1, p. 38-52.

Herring, Susan (1996), "Two variants of an electronic message schema", *Pragmatics and Beyond New Series*, p. 81-108.

Hew, Khe Foon, Xiang Hu, Chen Qiao et Ying Tang (2020), "What predicts student satisfaction with MOOCs : A gradient boosting trees supervised

machine learning and sentiment analysis approach", *Computers & Education*, 145, p. 103724.

Holbrook, Morris B. et Elizabeth C. Hirschman (1982), "The experiential aspects of consumption : Consumer fantasies, feelings, and fun", *Journal of Consumer Research*, 9, 2, p. 132-140.

Hu, Minqing et Bing Liu (2004), "Mining and summarizing customer reviews", in *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 168-177.

Hunston, Susan (2010), *Corpus approaches to evaluation : Phraseology and evaluative language*, Routledge.

Hunston, Susan et Gill Francis (2000), *Pattern grammar : A corpus-driven approach to the lexical grammar of English*, John Benjamins Publishing, t. 4.

Hunston, Susan et Geoffrey Thompson (2000), *Evaluation in text : Authorial stance and the construction of discourse : Authorial stance and the construction of discourse*, Oxford University Press.

Jackiewicz, Agata (2016), *Etudes sur les discours évaluatifs et d'opinion*, L'Harmattan.

Jang, Hayeon, Munhyong Kim et Hyopil Shin (2013), "KOSAC : A full-fledged Korean sentiment analysis corpus", in *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, p. 366-373.

Jang, Hayeon et Hyopil Shin (2010a), "Effective use of linguistic features for sentiment analysis of korean", in *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, p. 173-182.

– (2010b), "Language-specific sentiment analysis in morphologically rich languages", in *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, p. 498-506.

Jarukan, Jitwongnan (2014), *L'analyse des adjectifs axiologiques dans les ouvrages touristiques sur la Thaïlande*, mém. de mast., Université Stendhal - Grenoble 3.

Jarvenpaa, Sirrka L. et Peter A. Todd (1996), "Consumer reactions to electronic shopping on the World Wide Web", *International Journal of electronic commerce*, 1, 2, p. 59-88.

Jeong, EunHa et SooCheong Shawn Jang (2011), "Restaurant experiences triggering positive electronic word-of-mouth (eWOM) motivations", *International Journal of Hospitality Management*, 30, 2, p. 356-366.

Ji, Shihao, Hyokun Yun, Pinar Yanardag, Shin Matsushima et SVN Vishwanathan (2015), "Wordrank : Learning word embeddings via robust ranking", *arXiv preprint arXiv :1506.02761*.

Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu et Tiejun Zhao (2011), "Target-dependent Twitter Sentiment Classification", in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 151-160.

Jones, Michael A., Kristy E. Reynolds et Mark J. Arnold (2006), "Hedonic and utilitarian shopping value : Investigating differential effects on retail outcomes", *Journal of business research*, 59, 9, p. 974-981.

Jurafsky, Dan, Victor Chahuneau, Bryan Routledge et Noah Smith (2014), "Narrative framing of consumer sentiment in online restaurant reviews", *First Monday*, 19.

Kalchbrenner, Nal, Edward Grefenstette et Phil Blunsom (2014), "A Convolutional Neural Network for Modelling Sentences", in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 655-665.

Kennedy, Alistair et Diana Inkpen (2006), "Sentiment classification of movie reviews using contextual valence shifters", *Computational intelligence*, 22, 2, p. 110-125.

Kerbrat-Orecchioni, Catherine (1980), *L'énonciation de la subjectivité dans le langage (4e édition) [version Kindle iOS]*, Armand Colin.

Khan, Aurangzeb, Baharum Baharudin, Lam Hong Lee et Khairullah Khan (2010), "A Review of Machine Learning Algorithms for Text-Documents Classification", *Journal of Advances in Information Technology*, 1, p. 4-20.

Kilgarriff, Adam (2009), "Simple maths for keywords", in *Proceedings of the Corpus Linguistics Conference*.

Kim, Hung-gyu, Beom-mo Kang et Jungha Hong (2007), "21st Century Sejong Corpora (to be) Completed", *The Korean Language in America*, 12, p. 31-42.

Kim, Jin-Ok, Sun-Sook Lee et Hwan-Seung Yong (2011), "Automatic classification scheme of opinions written in Korean", *Journal of Korean Institute of Information Scientists and Engineer : Database*, 38, 6, p. 423-428.

Kim, Soo-Min et Eduard Hovy (2004), "Determining the sentiment of opinions", in *Proceedings of the 20th International Conference on Computational Linguistics*, p. 1367-1373.

Kim, Sung-il, Jae-ho Lee et Morton Ann Gernsbacher (2004), "The advantage of first mention in Korean the temporal contributions of syntactic, semantic, and pragmatic factors", *Journal of psycholinguistic research*, 33, 6, p. 475-491.

Kim, Yoosin, Seung Ryul Jeong *et al.* (2015), "Comparing Machine Learning Classifiers for Movie WOM Opinion Mining." *KSII Transactions on Internet & Information Systems*, 9, 8.

Kim, Yoosin, Mingon Kang et Seung Ryul Jeong (2018), "Text Mining and Sentiment Analysis for Predicting Box Office Success", *KSII Transactions on Internet and Information Systems*, 12, 8, p. 4090-4102.

Kim, Youngsam et Hyopil Shin (2017), "Finding sentiment dimension in vector space of movie reviews : an unsupervised approach", *Journal of Cognitive Science*, 18, 1, p. 85-101.

Kiritchenko, Svetlana, Xiaodan Zhu et Saif M. Mohammad (2014), "Sentiment analysis of short informal texts", *Journal of Artificial Intelligence Research*, 50, p. 723-762.

Körtvélyessy, Livia et Pavel Stekauer (2011), *Lexis 6: Diminutives and Augmentatives in the Languages of the World.*

Landis, J. Richard et Gary G. Koch (1977), "The Measurement of Observer Agreement for Categorical Data", *Biometrics*, 33, 1, p. 159-174.

Lark, Joseph (2017), *Construction semi-automatique de ressources pour la fouille d'opinion*, thèse de doct., Université de Nantes.

Lark, Joseph, Emmanuel Morin et Sebastián Peña Saldarriaga (2015), "CANÉPHORE : un corpus français pour la fouille d'opinion ciblée", in *22e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, p. 418-424.

Lark, Joseph, Emmanuel Morin et Sebastián Peña Saldarriaga (2016), "Extraction d'opinions ambig'ues dans des corpus d'avis clients", in *23e*

conférence sur le Traitement Automatique des Langues Naturelles (TALN), p. 451-458.

Lê, Sébastien, Julie Josse, François Husson *et al.* (2008), "FactoMineR : an R package for multivariate analysis", *Journal of statistical software*, 25, 1, p. 1-18.

Lebart, Ludovic, Bénédicte Pincemin et Céline Poudat (2019), *Analyse des données textuelles*, Mesure et évaluation, Presses de l'Université du Québec, t. 11.

Lee, Jae Jun, Suhn Beom Kwon et Sung Mahn Ahn (2018), "Sentiment Analysis Using Deep Learning Model based on Phoneme-level Korean", *Journal of Information Technology Services*, 17, 1, p. 79-89.

Lefer, Marie-Aude et Natalia Grabar (2013), "French evaluative prefixes in translation : from automatic alignment to semantic categorization", in *'Theoretical and computational morphology : New trends and synergies' workshop, International Congress of Linguists*.

– (2014), "Evaluative prefixes in translation : From automatic alignment to semantic categorization", *Linguistic Issues in Language Technology*, 11, 6, p. 169-187.

Legallois, Dominique (2006), "Présentation générale. Le texte et le problème de son et ses unités : propositions pour une déclinaison", *Langages*, 3, p. 3-9.

– (2016), *La notion de construction*, http://encyclogram.fr/notx/012/012_Notice.php.

Legallois, Dominique et Stéphane Ferrari (2006), "Vers une grammaire de l'évaluation des objets culturels", in *Actes d'ISDDo6 : Colloque international Discours et Document, Schedæ*, p. 57-68.

Legallois, Dominique et Véronique Lenepveu (2014), "L'évaluation dans les textes : des relations inter-propositionnelles aux séquences discursives", *Langue française*, 4, p. 17-33.

Lemaître, Guillaume, Fernando Nogueira et Christos K. Aridas (2017), "Imbalanced-Learn : A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning", *Journal of Machine Learning Research*, 18, 1, p. 559-563.

Lewis, Robert C. et Richard Everett Chambers (2000), *Marketing leadership in hospitality. Foundations and practices (3rd edition)*, Wiley.

Lim, Joa-Sang et Jin-Man Kim (2014), "An empirical comparison of machine learning models for classifying emotions in Korean Twitter", *Journal of Korea Multimedia Society*, 17, 2, p. 232-239.

Liu, Bing (2012), *Sentiment analysis and opinion mining*, Morgan & Claypool Publishers.

– (2015), *Opinions, Sentiment, and Emotion in Text*, Sentiment Analysis : Mining Opinions, Sentiments, and Emotions, Cambridge University Press.

– (2017), "Many Facets of Sentiment Analysis", in *A practical guide to sentiment analysis*, Springer, p. 11-39.

Luo, Ying et YL Wang (2014), "Sen's critique of utilitarianism and its enlightenment on Chinese dream", *Contemporary Finance & Economics*, 6, 1.

Mann, William C. et Sandra A. Thompson (1988), "Rhetorical Structure Theory : Toward a functional theory of text organization", *Text*, 8, 3, p. 243-281.

Marchand, Morgane (2013), "Fouille d'opinion : ces mots qui changent de polarité selon le domaine", in *Actes de CORIA 2013 : Conférence en recherche d'information et applications*, p. 347-352.

Marcoccia, Michel (2016), *Analyser la communication numérique écrite [version Kindle iOS]*, Armand Colin.

Marcus, Adam, Michael S. Bernstein, Osama Badar, David R Karger, Samuel Madden et Robert C. Miller (2011), "Twitinfo : aggregating and visualizing microblogs for event exploration", in *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 227-236.

Martin, James R. et Peter R. White (2003), *The language of evaluation*, Springer.

Mathieu, Yvette Yannick (2005), "Annotation of emotions and feelings in texts", in *International Conference on Affective Computing and Intelligent Interaction*, Springer, p. 350-357.

– (2006), "A computational semantic lexicon of french verbs of emotion", in *Computing Attitude and Affect in text : Theory and Applications*, Springer, p. 109-124.

Maurel, Denis (1993), "Reconnaissance automatique d'un groupe nominal prépositionnel. Exemple des adverbes de date", *Lexique*, 11, p. 147-161.

McEnery, Tony (2016), "Keywords", in *Triangulating Methodological Approaches in Corpus Linguistic Research*, Routledge, p. 32-44.

Mehrabian, Albert et James A. Russell (1974), *An approach to environmental psychology*, MIT Press.

Mel′čuk, Igor′ Aleksandrovič (1994), *Cours de morphologie générale : Significations morphologiques*, Presses de l'Université de Montréal, t. 2.

Mikolov, Tomas, Kai Chen, Greg Corrado et Jeffrey Dean (2013a), "Efficient Estimation of Word Representations in Vector Space", in *1st International Conference on Learning Representations*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado et Jeffrey Dean (2013b), "Distributed representations of words and phrases and their compositionality", in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, t. 2, p. 3111-3119.

Moeschler, Jacques et Antoine Auchlin (2009), *Introduction à la linguistique contemporaine (3e édition) [version Kindle iOS]*, Armand Colin.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh et Svetlana Kiritchenko (2018), "SemEval-2018 Task 1: Affect in Tweets", in *Proceedings of The 12th International Workshop on Semantic Evaluation*, p. 1-17.

Mohammad, Saif M. et Peter D. Turney (2010), "Emotions evoked by common words and phrases : Using mechanical turk to create an emotion lexicon", in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, p. 26-34.

– (2013), "Crowdsourcing a word–emotion association lexicon", *Computational Intelligence*, 29, 3, p. 436-465.

Moraes, Rodrigo, JoãO Francisco Valiati et Wilson P GaviãO Neto (2013), "Document-level sentiment classification : An empirical comparison between SVM and ANN", *Expert Systems with Applications*, 40, 2, p. 621-633.

Morinaga, Satoshi, Kenji Yamanishi, Kenji Tateishi et Toshikazu Fukushima (2002), "Mining product reputations on the web", in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 341-349.

Mullen, Tony et Nigel Collier (2004), "Sentiment Analysis using Support Vector Machines with Diverse Information Sources", in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 412-418.

Müller, Charles (1979), *Langue française et linguistique quantitative*, Slatkine.

Munezero, Myriam, Calkin Suero Montero, Erkki Sutinen et John Pajunen (2014), "Are They Different ? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text", *IEEE Transactions on Affective Computing*, 5, 2, p. 101-111.

Murray, Henry Alexander et Christiana D. Morgan (1945), "A clinical study of sentiments (I & II)", *Genetic Psychology Monographs*.

Myung, Jae-Seok, Dong-Joo Lee et Sang-goo Lee (2008), "A Korean product review analysis system using a semi-automatically constructed semantic dictionary", *Journal of Korean Institute of Information Scientists and Engineer : Software and Applications*, 35, 6, p. 392-403.

Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani et Veselin Stoyanov (2016), "SemEval-2016 Task 4: Sentiment Analysis in Twitter", in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 1-18.

Nakov, Preslav, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter et Theresa Wilson (2013), "SemEval-2013 Task 2: Sentiment Analysis in Twitter", in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 312-320.

Nasukawa, Tetsuya et Jeonghee Yi (2003), "Sentiment analysis : Capturing favorability using natural language processing", in *Proceedings of the 2nd International Conference on Knowledge Capture*, p. 70-77.

Negi, Sapna, Kartik Asooja, Shubham Mehrotra et Paul Buitelaar (2016), "A Study of Suggestions in Opinionated Texts and their Automatic Detection", in *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, p. 170-178.

Negi, Sapna et P. Buitelaar (2017), "Suggestion mining from opinionated text", *Sentiment Analysis in Social Networks*, p. 129-139.

Negi, Sapna et Paul Buitelaar (2015), "Towards the Extraction of Customer-to-Customer Suggestions from Reviews", in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 2159-2167.

Negi, Sapna, Tobias Daudert et Paul Buitelaar (2019), "SemEval-2019 Task 9: Suggestion Mining from Online Reviews and Forums", in *Proceedings of the 13th International Workshop on Semantic Evaluation*, p. 877-887.

Negi, Sapna, Maarten De Rijke et Paul Buitelaar (2018), "Open domain suggestion mining : Problem definition and datasets", *arXiv preprint arXiv :1806.02179*.

Nickel, Maximilian et Douwe Kiela (2017), "Poincaré embeddings for learning hierarchical representations", *arXiv preprint arXiv :1705.08039*.

Oliver, Richard L. (2014), *Satisfaction : A behavioral perspective on the consumer : A behavioral perspective on the consumer*, Routledge.

Pang, Bo et Lillian Lee (2005), "Seeing Stars : Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales", in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, p. 115-124.

Pang, Bo, Lillian Lee et Shivakumar Vaithyanathan (2002), "Thumbs up ? Sentiment Classification using Machine Learning Technique", in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, p. 79-86.

Park, Cheol (2004), "Efficient or enjoyable ? Consumer values of eating-out and fast food restaurant consumption in Korea", *International Journal of Hospitality Management*, 23, 1, p. 87-94.

Park, Hyun-jung, Min-chae Song et Kyung-shik Shin (2018a), "Sentiment Analysis of Korean Reviews Using CNN : Focusing on Morpheme Embedding", *Journal of Intelligence and Information Systems*, 24, 2, p. 59-83.

Park, Sungjoon, Jeongmin Byun, Sion Baek, Yongseok Cho et Alice Oh (2018b), "Subword-level word vector representations for Korean", in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 2429-2438.

Paroubek, Patrick, Cyril Grouin, Patrice Bellot, Vincent Claveau, Iris Eshkol-Taravella, Amel Fraisse, Agata Jackiewicz, Jihen Karoui, Laura Monceaux et Juan-Manuel Torres-Moreno (2018), "DEFT2018: recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France", in *Actes de la conférence TALN 2018*, t. 2, p. 219-229.

Patwa, Parth, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio et Amitava Das (2020), "SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets", *arXiv preprint arXiv :2008.04277*.

Paumier, Sébastien et Jee-Sun Nam (2014), "Un système de dictionnaire de mots simples du coréen", in *Penser le Lexique-Grammaire. Perspectives actuelles*, sous la dir. de Fryni Kakoyianni-Doa, Collection Colloques, congrès et conférences. Sciences du Langage, histoire de la langue et des dictionnaires. 30th International Conference on Lexis and Grammar (Nicosia, Cyprus, 2011), Honoré Champion, p. 481-490.

Pecore, Stefania et Jeanne Villaneau (2018), "Complex and Precise Movie and Book Annotations in French Language for Aspect Based Sentiment Analysis", in *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg *et al.* (2011), "Scikit-learn : Machine learning in Python", *Journal of Machine Learning Research*, 12, p. 2825-2830.

Pennington, Jeffrey, Richard Socher et Christopher Manning (2014), "GloVe : Global Vectors for Word Representation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, p. 1532-1543.

Pine, B. Joseph et James H. Gilmore (2011), *The Experience Economy (Updated Edition)*, Harvard Business Review Press.

Poirier, Damien, Françoise Fessant, Cécile Bothorel, Emilie Guimier De Neef et Marc Boullé (2009), "Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films", *Revue des Nouvelles Technologies de l'Information*, RNTI-E-17, Pages 147-169.

Polanyi, Livia et Annie Zaenen (2006), "Contextual valence shifters", in *Computing attitude and affect in text : Theory and applications*, Springer, p. 1-10.

Pontiki, Maria, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar et Ion Androutsopoulos (2015), "Semeval-2015 task 12: Aspect based sentiment analysis", in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 486-495.

Pontiki, Maria, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra et Gülşen Eryiğit (2016), "SemEval-2016 Task 5: Aspect Based Sentiment Analysis", in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 19-30.

Pontiki, Maria, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos et Suresh Manandhar (2014), "SemEval-2014 Task 4: Aspect Based Sentiment Analysis", in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 27-35.

Poole, Robert (2016), "Good times, bad times : A keyword analysis of letters to shareholders of two Fortune 500 banking institutions", *International Journal of Business Communication*, 53, 1, p. 55-73.

Poria, Soujanya, Erik Cambria et Alexander Gelbukh (2016), "Aspect extraction for opinion mining with a deep convolutional neural network", *Knowledge-Based Systems*, 108, p. 42-49.

Potts, Christopher (2010), "On the negativity of negation", in *Semantics and Linguistic Theory*, t. 20, p. 636-659.

Poudat, Céline et Frédéric Landragin (2017), *Explorer un corpus textuel : Méthodes-pratiques-outils*, De Boeck Superieur.

Poudat, Céline, Ciara R. Wigham et Loïc Liégeois (2020), "Les corpus de la communication médiée par les réseaux : une introduction", *Corpus*, Corpus complexes. Traitements, standardisation et analyse des corpus de communication médiée par les réseaux, 20.

Prihodko, Ganna (2018), "Specific Nature of Evaluative Speech Acts", *Advanced Education*, 5, p. 201-205.

Prihodko, Ganna, Oleksandra Prykhodchenko et Andrii Galaidin (2019), "Realization of Evaluative Utterance in Different Types of Context", *Scientific Journal of Polonia University*, 32, 1, p. 68-76.

Pupier, Paul (1998), "Une première systématique des évaluatifs en français", *Revue québécoise de linguistique*, 26, 1, p. 51-78.

Qian, Qiao, Minlie Huang, Jinhao Lei et Xiaoyan Zhu (2017), "Linguistically Regularized LSTM for Sentiment Classification", in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1679-1689.

Ramanand, J., Krishna Bhavsar et Niranjan Pedanekar (2010), "Wishful Thinking - Finding suggestions and 'buy' wishes from product reviews", in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, p. 54-61.

Rayson, Paul (2008), "From key words to key semantic domains", *International journal of corpus linguistics*, 13, 4, p. 519-549.

Řehůřek, Radim et Petr Sojka (2010), "Software Framework for Topic Modelling with Large Corpora", in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45-50.

Rosenthal, Sara, Noura Farra et Preslav Nakov (2017), "SemEval-2017 Task 4: Sentiment Analysis in Twitter", in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 502-518.

Rosenthal, Sara, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter et Veselin Stoyanov (2015), "SemEval-2015 Task 10: Sentiment Analysis in Twitter", in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 451-463.

Rosenthal, Sara, Preslav Nakov, Alan Ritter et Veselin Stoyanov (2014), "SemEval-2014 Task 9: Sentiment Analysis in Twitter", in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 73-80.

Russell, James A. (2003), "Core affect and the psychological construction of emotion", *Psychological Review*, 110, 1, p. 145-172.

Salton, Gerard et Christopher Buckley (1988), "Term-weighting approaches in automatic text retrieval", *Information Processing & Management*, 24, 5, p. 513-523.

Schmid, Helmut (1994), "Probabilistic Part-of-Speech Tagging Using Decision Trees", in *Proceedings of International Conference on New Methods in Language Processing*.

Scott, Mike (1998), *WordSmith tools manual version 3.0*, Oxford University Press.

– (2020), *WordSmith Tools version 8 [computer software], Stroud : Lexical Analysis Software*.

Scott, Mike et Christopher Tribble (2006), *Textual patterns : Key words and corpus analysis in language education*, John Benjamins Publishing, t. 22.

Seale, Clive, Sue Ziebland et Jonathan Charteris-Black (2006), "Gender, cancer experience and internet use : a comparative keyword analysis of interviews and online cancer support groups", *Social science & medicine*, 62, 10, p. 2577-2590.

Searle, John R. (1972), *Speech Acts : An Essay in the Philosophy of Languag*, Cambridge University Press.

Severyn, Aliaksei et Alessandro Moschitti (2015), "UNITN : Training Deep Convolutional Neural Network for Twitter Sentiment Classification", in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 464-469.

Shin, Hyopil (2013), "Specifications and Analysis of the Korean Sentiment Analysis Corpus", *Language Research*.

Shin, Hyopil, Munhyong Kim et Suzi Park (2016), "Modality-based Sentiment Analysis through the Utilization of the Korean Sentiment Analysis Corpus", *EONEOHAG : Journal of the Linguistic Society of Korea*, 74, p. 93-114.

Silge, Julia et David Robinson (2016), "tidytext : Text mining and analysis using tidy data principles in R", *Journal of Open Source Software*, 1, 3, p. 37.

Singhal, Amit et I. Google (2001), "Modern information retrieval : A brief overview", *IEEE Data Engineering Bulletin*, 24, 4, p. 35-43.

Sokolova, Marina et Guy Lapalme (2011), "Learning opinions in user-generated web content", *Natural Language Engineering*, 17, 4, p. 541-567.

Song, Minchae, Hyunjung Park et Kyung-shik Shin (2019), "Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean", *Information Processing & Management*, 56, 3, p. 637-653.

Spangenberg, Eric R., Kevin E. Voss et Ayn E. Crowley (1997), "Measuring the hedonic and utilitarian dimensions of attitude : A generally applicable scale", *Advances in Consumer Research North American Advances*.

Stets, Jan E. (2006), "Emotions and sentiments", in *Handbook of social psychology*, Springer, p. 309-335.

Stone, Philip J., Dexter C. Dunphy et Marshall S. Smith (1966), *The general inquirer : A computer approach to content analysis*, MIT press.

Stosic, Dejan et Dany Amiot (2011), *Quand la morphologie fait des manières : les verbes évaluatifs et l'expression de la manière en français*.

Stratos, Karl (2017), "A Sub-Character Architecture for Korean Language Processing", in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Stump, Gregory T. (1993), "How peculiar is evaluative morphology ?", *Journal of linguistics*, 29, 1, p. 1-36.

Sung, Wonkyung, Jaeyoung An et Choong C. Lee (2020), "Development and Validation of the Letter-unit based Korean Sentimental Analysis Model Using Convolution Neural Network", *The Journal of Society for e-Business Studies*, 25, 1, p. 13-33.

Szarvas, György, Veronika Vincze, Richárd Farkas, György Móra et Iryna Gurevych (2012), "Cross-genre and cross-domain detection of semantic uncertainty", *Computational Linguistics*, 38, 2, p. 335-367.

Taboada, Maite (2011), "Stages in an online review genre", *Text & Talk*, 31, 2, p. 247-269.

Taboada, Maite, Caroline Anthony et Kimberly Voll (2006), "Methods for Creating Semantic Orientation Dictionaries", in *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll et Manfred Stede (2011), "Lexicon-based methods for sentiment analysis", *Computational linguistics*, 37, 2, p. 267-307.

Tang, Duyu, Bing Qin et Ting Liu (2015), "Learning Semantic Representations of Users and Products for Document Level Sentiment Classification", in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 1014-1023.

Teo, Thompson S.H. (2001), "Demographic and motivation variables associated with Internet usage activities", *Internet Research : Electronic Networking Applications and Policy*, 11, 2, p. 125-137.

Tong, Simon et Daphne Koller (2001), "Support vector machine active learning with applications to text classification", *Journal of Machine Learning Research*, 2, p. 45-66.

Turney, Peter D. (2002), "Thumbs up or thumbs down ? Semantic orientation applied to unsupervised classification of reviews", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 417-424.

Tutin, Agnès (2007), "Autour du lexique et de la phraséologie des écrits scientifiques", *Revue française de linguistique appliquée*, p. 5-14.

– (2010), "Dans cet article, nous souhaitons montrer que··· Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines." *LIDIL - Revue de linguistique et de didactique des langues*, 41, p. 15-40.

Tutin, Agnès, Francis Grossmann et Presses Universitaires de Rennes (2014), *L'écrit scientifique : du lexique au discours*, Presses universitaires de Rennes.

Vernier, Matthieu (2011), *Analyse à granularité fine de la subjectivité*, thèse de doct., Université de Nantes.

Vernier, Matthieu, Yann Mathet, François Rioult, Thierry Charnois, Stéphane Ferrari et Dominique Legallois (2007), "Classification de textes d'opinions : une approche mixte n-grammes et sémantique", in *Atelier Défi Fouille de Textes (DEFT'07) dans le cadre de la plate-forme*

AFIA 2007 (Association Française pour l'Intelligence Artificielle), p. 93-108.

Vernier, Matthieu et Laura Monceaux (2010), "Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques", *Traitement Automatique des Langues*, 51, 1, p. 125-149.

Villaneau, Jeanne, Stefania Pecore, Farida Saïd et Pierre-François Marteau (2019), "Combiner analyse syntaxique de surface et apprentissage supervisé pour la fouille d'opinion ciblée : expérimentations sur des données d'opinion concernant les livres", in *Actes de EGC'2019 : 19e conférence sur l'extraction et la gestion des connaissances*.

Vincent, Marc et Grégoire Winterstein (2013), "Construction et exploitation d'un corpus français pour l'analyse de sentiment", in *Actes de TALN'2013 : 20e conférence sur le Traitement Automatique des Langues Naturelles*, p. 764-771.

Wachsmuth, Henning, Martin Trenkmann, Benno Stein, Gregor Engels et Tsvetomira Palakarska (2014), "A review corpus for argumentation analysis", in *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, p. 115-127.

Wang, Yequan, Minlie Huang, Xiaoyan Zhu et Li Zhao (2016), "Attention-based LSTM for aspect-level sentiment classification", in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 606-615.

Webber, Bonnie (2004), "D-LTAG : extending lexicalized TAG to discourse", *Cognitive Science*, 28, 5, p. 751-779.

Wicaksono, Alfan Farizki et Sung-Hyon Myaeng (2012), "Mining Advices from Weblogs", in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, p. 2347-2350.

– (2013), "Automatic Extraction of Advice-Revealing Sentences Foradvice Mining from Online Forums", in *Proceedings of the 7th International Conference on Knowledge Capture*, p. 97-104.

Wiebe, Janyce (2000), "Learning Subjective Adjectives from Corpora", in *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, p. 735-740.

Wiebe, Janyce et Ellen Riloff (2005), "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts", in *Computational Linguistics and Intelligent Text Processing*, Springer, p. 486-497.

Wiebe, Janyce M., Rebecca F. Bruce et Thomas P. O'Hara (1999), "Development and Use of a Gold-Standard Data Set for Subjectivity Classifications", in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, p. 246-253.

Wilson, Theresa, Janyce Wiebe et Paul Hoffmann (2005), "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", in *Proceedings of the 2005 Conference on Human Language Technology and Conference on Empirical Methods in Natural Language Processing*, p. 347-354.

Wilson, Theresa, Janyce Wiebe et Rebecca Hwa (2004), "Just how mad are you ? Finding strong and weak opinion clauses", in *Proceedings of the National Conference on Artificial Intelligence*.

Wu, Yingcai, Shixia Liu, Kai Yan, Mengchen Liu et Fangzhao Wu (2014), "OpinionFlow : Visual Analysis of Opinion Diffusion on Social Media", *IEEE transactions on visualization and computer graphics*, 20, 12, p. 1763-1772.

Yessenalina, Ainur, Yisong Yue et Claire Cardie (2010), "Multi-Level Structured Models for Document-Level Sentiment Classification", in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, p. 1046-1056.

Yu, Hong et Vasileios Hatzivassiloglou (2003), "Towards answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 129-136.

Zhang, Lei (2012), *Analyse automatique d'opinion : problématique de l'intensité et de la négation pour l'application à un corpus journalistique*, thèse de doct., Université de Caen.

Zhu, Xiaodan, Svetlana Kiritchenko et Saif Mohammad (2014), "NRC-Canada-2014: Recent improvements in the sentiment analysis of Tweets", in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 443-447.

Zhuang, Li, Feng Jing et Xiao-Yan Zhu (2006), "Movie review mining and summarization", in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, p. 43-50.

Hyun Jung KANG

Regards croisés sur les avis en ligne :
approches du TAL et de la linguistique de corpus

Résumé

La thèse se situe dans la lignée des recherches en Traitement Automatique des Langues (TAL) sur la fouille d'opinions et propose la modélisation, l'analyse outillée et le traitement automatique des évaluations en ligne des restaurants. L'évaluation des restaurants visités ne se limite pas aux opinions positives ou négatives données par les clients, mais peut avoir d'autres fonctions : le visiteur laisse son avis pour donner son opinion (opinions), faire des suggestions (suggestion), exprimer ses intentions (intention) ou décrire son expérience (description). Chaque fonction est analysée du point de vue quantitatif à travers l'extraction de mots-clés, la répartition des catégories morphosyntaxiques et sa position au sein de l'évaluation. Un module de détection automatique de chacune de ces fonctions est développé. La méthode utilisée est fondée sur l'apprentissage de surface et l'apprentissage profond. Les performances obtenues pour chaque fonction sont interprétées en tenant compte des spécificités du corpus traité. La généralisabilité du modèle développé a été testée et validée sur d'autres données : un corpus relevant du domaine de l'hôtellerie et un corpus écrit dans une autre langue – le coréen.

Mots-clés fouille d'opinion, avis en ligne, évaluation, apprentissage de surface, apprentissage profond, traitement automatique des langues

Hyun Jung KANG

Crossed Perspectives on Online Reviews :
NLP and Corpus Linguistics Approaches

Abstract

This dissertation examines, from various perspectives, evaluations of restaurants written by customers and published online. The study highlights why customers write reviews on the internet, analyzes the linguistic markers used to express and share their experiences, and identifies the unique features of online reviews. From manual observation of the corpus, I propose a model to study evaluations of experiences in restaurants and show that evaluations can have several functions and contain different information: reviewers write reviews to give opinions, make suggestions, express intentions, and describe their experiences. The typology enabled the investigation of the nature of evaluations, the way reviewers evaluate restaurants, and their motivations for expressing evaluations on the internet. I conducted various quantitative analyses using corpus linguistics methods, including lexicometry and determining the distribution of parts of speech and the positions of evaluation categories within reviews. I performed the automatic detection of the evaluation categories using traditional machine learning and deep learning models. Finally, I assessed the generalizability of the model on additional corpora of reviews from another domain (hotels) and in another language (Korean).

Keywords opinion mining, online reviews, evaluation, machine learning, deep learning, computational linguistics