République Algérienne Démocratique et Populaire
الجمهورية الجزائرية الديمقراطية الشعبية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التعليم العالي و البحث العلمي

ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE

المدرسة الوطنية العليا للإعلام الآلي
(المعهد الوطني للتكوين في الإعلام الآلي سابقا)
École nationale Supérieure d'Informatique
ex. INI (Institut National de formation en Informatique)

Mémoire de fin d'études

Pour l'obtention du diplôme Master en Informatique

Option : Systèmes Informatiques et Logiciels

Etat de l'art sur l'intégration de la crédibilité dans la tache de fouille d'opinion

Réalisé par :
Mme. SLIMANI Wassila Maria

Encadré par :
Dr. SAID EL HADJ Lynda (ESI)

Promotion : 2022/2023

Dédicace

“

Je dédie ce mémoire,

À ma chère mère, pour son soutien, ses encouragements et son amour inconditionnel,

À mon père, pour sa confiance en moi et ses encouragements,

À ma chère mamie Halima, pour ses sacrifices, pour les longues heures qu'elle a priées pour mon bien-être et ma réussite et pour tout l'amour qu'elle m'a donnée,

À ma petite sœur Cerine et mon petit frère Yanis pour leur amour, leur soutien et les moments de joie que nous avons partagés,

À la mémoire de ma tante Kenza, la personne la plus gentille au monde, que ton âme repose en paix,

À mon cher oncle Abderezak pour avoir toujours été là quand j'avais besoin de lui sans jamais se plaindre,

À ma chère tante Wahiba et son mari Sidamer, pour leur soutien, leur aide, leur confiance et leurs prières.

À Sarah, Ahlem, Roumaissa, Lynda et Dounia pour leur amitié, leur soutien et toutes nos belles aventures.

À Wisseme et Rania pour leurs conseils et leur aide,

À tous ceux qui ont cru en moi, À tous ceux qui me sont chers, À vous tous,

Merci.

”

- Maria

Remerciements

"No one who achieves success does so without the help of others. The wise and confident acknowledge this help with gratitude."

-Alfred North Whitehead

En préambule à ce rapport, je souhaite exprimer ma reconnaissance envers toute personne ayant contribué à la réussite de ce projet.

Avant tout, je remercie Allah tout-puissant de m'avoir mis sur ce chemin et de m'avoir donné la patience et la volonté nécessaire afin de réaliser ce travail.

Je tiens à remercier toutes les personnes qui ont été impliquées dans la rédaction de ce rapport de Master. Particulièrement, j'adresse mes gratitudes les plus sincères à mon encadrante, madame Lynda Said Lhadj (Maitre de conférence à l'ESI). Tout au long de la période de mon stage, j'ai pu compter sur son expertise, sa bienveillance et ses conseils. Je suis reconnaissante du temps qu'elle m'a consacré ainsi que des ressources qu'elle m'a fournies.

Je remercie également toute l'équipe pédagogique de l'École nationale supérieure d'informatique (ESI), non seulement pour avoir assuré la méthodologie de recherche et de rédaction tout en long de cette formation master, mais aussi pour avoir assuré une formation de qualité durant mon cursus dans cette école.

Et enfin, je ne serai remercier assez ma famille et mes amies pour leur soutien et encouragements au fil de mes études.

Résumé

De nos jours, les plateformes sociales apparaissent comme des sources de preuves essentielles pour les entreprises qui ont besoin de cibler des consommateurs potentiels souhaitant acheter leurs produits/services. Les réseaux sociaux connaissent un intérêt croissant dans divers milieux scientifiques et commerciaux. En effet, les entreprises sont souvent soucieuses de leur identité en ligne, car elles sont conscientes que leur réputation peut être affectée ou même détruite par un commentaire ou un avis, c'est pourquoi elles essaient de mieux comprendre les attentes et les critiques que les internautes leur adressent et elles tentent de les caractériser selon deux axes :

(i) Pour estimer leur réputation

(ii) Pour les utiliser dans l'estimation de la satisfaction client et l'amélioration de leur expérience

Toutefois, un nouveau problème se pose, les avis exprimés sur une marque, un produit ou un service donné ne sont pas forcément crédibles. Cela est dû aux nouveaux acteurs qui ont fait de ce risque un business. En effet, des spammeurs, qui peuvent être des particuliers ou des entreprises, rédigent expressément des avis négatifs/positifs pour nuire ou promouvoir une marque. ces avis sont trompeurs, car ils ne sont pas fondés. Les plateformes les plus susceptibles d'être touchées par ces tentatives frauduleuses sont les sites d'avis en ligne, car ces derniers sont considérés comme des sources populaires de bouche-à-oreille électronique qui aident les consommateurs dans leurs processus de prise de décision. Étant donné que les faux avis publiés sont souvent inexacts et incertains, il peut être difficile de distinguer les avis spam des vrais. Pour cela, depuis la formalisation de la détection des spams d'opinion en 2008, de nombreux chercheurs se sont intéressés à cette problématique.

Le présent rapport s'intéresse aux méthodes présentes dans la littérature afin de répondre aux problématiques soulevés précédemment : son objectif sera de définir la notion de crédibilité d'une opinion et d'identifier les travaux qui s'y sont intéressés en addition à la caractérisation des opinions et à l'identification des spammeurs. Il sera organisé principalement en deux parties, la première concernera l'analyse des sentiments des clients et la deuxième abordera le problème de la détection d'avis trompeurs tout en présentant les approches présentes, les défis majeurs et les axes de recherches futurs.

Mots clés : Fouille D'opinion, Détection De Spam d'opinion, Crédibilité, E-Réputation.

Abstract

Nowadays, social platforms appear as essential sources of evidence for companies that need to target potential consumers wishing to buy their products/services. Social networks are experiencing growing interest in various scientific and commercial circles. Indeed, companies are often concerned about their online identity, as they are aware that their reputation can be affected or even destroyed by a comment or a review, which is why they try to better understand the expectations and criticisms that Internet users address them, and they attempt to characterize them according to two axes :

(i) To assess their reputation

(ii) For use in estimating customer satisfaction and improving their experience

However, a new problem arises, the opinions expressed on a given brand, product or service are not necessarily credible. This is due to the new players who have made this risk a business. Indeed, spammers, either individuals or companies, expressly write negative/positive opinions to harm or promote a brand. these reviews are misleading because they are unfounded. The platforms most likely to be affected by these fraudulent attempts are online review sites, as these are considered popular sources of electronic word-of-mouth that help consumers in their decision-making processes. Since fake reviews posted are often inaccurate and uncertain, it can be difficult to distinguish spam reviews from real ones. For this reason, since the formalization of opinion spam detection in 2008, many researchers have been interested in this issue.

This report focuses on the methods present in the literature to address the issues raised above : its objective will be to define the notion of opinion credibility and to identify the works that tackled it in addition to the characterization of reviews and identification of spammers. It will be organized mainly in two parts, the first will concern the analysis of customer sentiments and the second will address the problem of detecting misleading reviews while presenting the current approaches, the major challenges and areas for future research.

Keywords : Opinion Mining, Opinion Spam Detection, Credibility, E-Reputation.

ملخص

أ

في الوقت الحاضر، تظهر المنصات الاجتماعية كمصادر أساسية للأدلة للشركات التي تحتاج إلى استهداف المستهلكين المحتملين الراغبين في شراء منتجاتهم/خدماتهم. تشهد الشبكات الاجتماعية اهتمامًا متزايدًا في مختلف الأوساط العلمية والتجارية. في الواقع، غالبًا ما تهتم الشركات بهويتها عبر الإنترنت، لأنها تدرك أن سمعتها يمكن أن تتأثر أو حتى تدمر من خلال تعليق أو رأي، وهذا هو السبب في أنها تحاول فهم التوقعات والانتقادات التي يواجهها مستخدمو الإنترنت بشكل أفضل بشكل أفضل، محاولة توصيفهم وفقًا لمحورين :

(i) لتقييم سمعتهم

(ii) لتقدير رضا العملاء وتحسين تجربتهم

علي الرغم من كل هذا، تظهر مشكلة جديدة تتمثل في أن الآراء المعبر عنها بشأن علامة تجارية معينة أو منتج أو خدمة ليست بالضرورة موثوقة. هذا يرجع إلى اللاعبين الجدد الذين جعلوا هذه المخاطرة عملاً تجارياً. وبالفعل، فإن كتاب الآراء الزائفة، الذين يمكن أن يكونوا أفراداً أو شركات ، يكتبون آراء سلبية/إيجابية لإلحاق الضرر بعلامة تجارية أو الترويج لها. هذه التعليقات مضللة لأنها لا أساس لها من الصحة. الأنظمة الأساسية التي من المرجح أن تتأثر بهذه المحاولات الاحتيالية هي مواقع المراجعة عبر الإنترنت، حيث تُعدّ مصادر شائعة للآراء الإلكترونية التي تساعد المستهلكين في صنع القرار الخاص بهم. نظرًا لأن التعليقات المزيفة التي يتم نشرها غالبًا ما تكون غير دقيقة وغير مؤكدة، فيكون من الصعب التمييز بين التعليقات المزيفة والتعليقات الحقيقية. لهذا السبب، منذ إضفاء الطابع الرسمي على الكشف عن الآراء الكاذبة في عام 2008، اهتم العديد من الباحثين بهذه المسألة.

يركز هذا التقرير على الأساليب الموجودة في الأدبيات لمعالجة القضايا التي أثيرت أعلاه : سيكون هدفه تحديد مفهوم مصداقية الرأي وتحديد الأعمال التي تهتم به إضافة إلى توصيف الآراء وتحديد كاتبي الآراء الكاذبة. سيتم تنظيمه بشكل أساسي في جزأين، الأول سيتعلق بتحليل مشاعر العملاء والثاني سيتناول مشكلة اكتشاف الآراء المضللة مع تقديم الأساليب المعتمدة والتحديات الرئيسية ومجالات البحث في المستقبل.

كلمات مفتاحية : التنقيب عن الآراء، الكشف عن الآراء الكاذبة، المصداقية، السمعة الإلكترونية

Table des matières

Table des figures

Liste des tableaux

Liste des sigles et acronymes

AI	*Artificial Intelligence*
AMT	*Amazon Mechanical Turk*
ANN	*Artificial Neural Network*
ANOVA	*Analysis of variance*
AUC	*Area Under the receiver operating characteristics Curve*
BERT	*Bidirectional Encoder Representations from Transformers*
BOW	*Bag Of Words*
CBOW	*Continuous Bag of Word*
CNN	*Convolutif Neural Network*
DR	*Dimensionality Reduction*
ELMo	Embeddings from Language Models
EUPHORIA	*nEural mUlti-view aPproacH fOr RevIew spAm*
FFNN	*Feedforward neural network*
FN	*False Negative (Faux négatif)*
FP	*False Positive (Faux positif)*
GAN	*Genrative Adversial Network*
GloVe	*Global Vectors for Word Representation*
GMM	*Gaussian Mixture Modelling*
GRU	*Gated recurrent units*
ICA	*Independent Component Analysis*
KLDA	*Kernel Linear Discriminant Analysis*
KPCA	*Kernel Principal Component Analysis*

LASSO *Least Absolute Shrinkage and Selection Operator*

LDA *Linear Discriminant Analysis*

LIWC *Linguistic Inquiry and Word Count*

LR *Logistic Regression*

LSI *Latent Semantic Indexing*

LSTM *Long-Short Term Memory*

MGSD *Multi-iterative Graph-based opinion Spam Detection*

MLM *Masked Language Model*

MLP *Multi-Layer Perceptron*

MMI *Modified Mutual Information*

NB *Naive Bayes*

NLP *Natural Language Processing*

OOV *Out Of Vocabulary*

P *Précision*

PCA *Principal Component Analysis*

PNN *Perceptron Neural Network*

POS *Part Of Speech*

PR *Precision-Recall Curve*

PU(learning) *Positive Unlabeled (Learning)*

R *Rappel*

ReLU *Rectified Linear Unit*

RI *Recherche d'information*

RLOSD *Representation Learning based Opinion Spam Detection*

RNN *Recurrent Neural Network*

ROC *Receiver Operating Characteristic curve*

SMOTE *Synthetic Minority Over-sampling TEchnique*

SOM *Self-Organizing Maps*

SVD *Singular Value Decomposition*

SVM *Support Vector Machine*

TALN *Traitement automatique du langage naturel*

TF *Term Frequency)*

TF-IDF *Term Frequency (TF)-inverse document frequency)*

TFP *Taux de faux positifs*

TN *True Negative (Vrai négatif)*

TP *True Positive (Vrai positif)*

TSVM *Trasnductive Support Vector Machine*

TVP *Taux de vrais positifs*

URL *Uniform Resource Locator*

WE *Word Embedding*

Word2vec *Word to Vector*

Introduction générale

Contexte

Grâce aux progrès de la technologie, en particulier du Web, on observe de nos jours une croissance sans cesse grandissante du commerce électronique en termes du marketing d'entreprise en ligne. Ce type de marketing est réussi non seulement grâce à la publicité, mais aussi grâce à la disponibilité d'opinions publiques sur les entreprises -qui adoptent cette approche- et leurs produits/services, ce qui fournit aux clients une source d'information fiable pour la prise de décision.

Fig. 1 : Résultats de l'enquête Ever-Growing Power of Reviews.

Les opinions en ligne affectent les décisions d'achat des clients. La Figure 1 ci-dessus représente les résultats d'une enquête menée par The PowerReviews sous le nom de Ever-Growing Power of Reviews [1], en 2021, qui a interrogé 6 538 clients à travers les États-Unis.

[1] https://www.powerreviews.com/insights/power-of-reviews-survey-2021/

Selon cette enquête, les notes et les avis sont devenus le facteur le plus important qui influe sur les décisions d'achat, se classant au-dessus du prix, de la livraison gratuite, de la marque et des recommandations de la famille et des amis. Et ceci contrairement à des enquêtes similaires qu'ils ont menées en 2014 et 2018, où le prix était le facteur le plus important.

Suite à l'impact significatif des avis sur les décisions des clients, les entreprises sont conscientes maintenant plus que jamais de l'importance de leur notoriété numérique et sont intéressées à connaître leur e-réputation afin de la soigner et de la maintenir pour pouvoir à la fois fidéliser leurs clients et en conquérir de nouveaux.

Problématique

Cette abondance d'information sur le web a fait émerger un nouveau problème : il a été démontré que dans des conditions concurrentielles et classées, les opinions publiques peuvent être utilisées avec des intentions malveillantes qui pourraient être de promouvoir ou de rétrograder un produit ou un service cible dans le but d'induire en erreur les consommateurs pour un certain gain (financier ...). C'est ce qu'on appelle le spam d'opinion.

Cette augmentation des avis trompeurs sur le web a réduit la crédibilité des opinions publiques et a affecté la confiance des clients qui ne savent plus quoi croire. Ces avis spam peuvent être soit générés par des personnes payées pour cela ou par des algorithmes, ce qui rend la tâche de détection plus difficile puisque ces algorithmes peuvent être ajustés pour battre les techniques de détection de l'état de l'art.

Le spam d'opinion est un sujet d'intérêt qui a conduit à l'émergence de nombreuses approches qui visent à le détecter et le classifier pour faciliter la prise de décision. Cette problématique a été abordée pour la 1 fois par (Jindal et B. Liu 2008). Nous nous intéressons aux travaux de la littérature qui ont été proposés, depuis, avec un intérêt particulier aux questions suivantes :

Ceci nous a amenés à poser les questions suivantes :

- Comment peut-on définir la polarité d'une opinion ? Et est-elle un indicateur déterminant pour la crédibilité des opinions ?

- Compte tenu du contexte actuel, quelle serait la définition de la crédibilité ?

- Que peut-on définir en tant que spam d'opinion ?

Objectifs

- Examiner les définitions et les concepts clés liés à l'analyse de sentiments et la détection de spam d'opinions.

- Présenter les différentes méthodes et techniques utilisées pour l'analyse de sentiments et la détection de spam d'opinions, en mettant l'accent sur les méthodes basées sur l'apprentissage automatique.

- Identifier les défis et les limites associés à l'analyse de sentiments et la détection de

spam d'opinions.

Structure du mémoire

Ce rapport de Master vise à synthétiser les travaux de l'état de l'art pertinents aux problématiques posées, et est constituée de deux chapitres, à savoir :

Chapitre 01 : L'analyse des sentiments

Dans ce premier chapitre couvre les concepts de base, les tâches, les sous-domaines et les défis rencontrés dans le domaine de l'analyse des sentiments.

Chapitre 02 : La détection de Spam

Dans ce chapitre, nous introduisons le domaine de la détection de spam d'opinions. Nous présentons les concepts de base de la détection de spam nécessaires pour la compréhension du domaine, son processus général, ainsi qu'une synthèse des travaux de la littérature.

Conclusion et perspectives

Enfin, dans cette partie, nous concluons notre mémoire par un bilan des résultats obtenus ainsi qu'une présentation des perspectives de recherche.

Annexes

L'annexe de ce rapport contient des ressources supplémentaires permettant une meilleure compréhension du sujet traité, comme des exemples de résultats d'application d'algorithme.

Chapitre 1
L'analyse des sentiments

1.1 Introduction

Web 2.0 est un terme utilisé pour décrire une nouvelle ère d'Internet, cette notion a été formalisée pour la première fois en 2005 par Tim O'Reilly dans son article intitulé « What is Web 2.0 ? » [1]. En effet, le web est devenu une plateforme de communication, de création, de collaboration et de sociabilité : tout utilisateur, qu'il soit un particulier, une société ou un gouvernemental, peut publier tout type de contenu sur différentes plateformes. Ces contenus peuvent raconter une expérience, un avis sur un produit acheté, un service ou simplement une personne ou une personnalité publique. Cette possibilité de pouvoir s'exprimer dans différents contextes a rendu les plateformes du web une mine d'information que les organisations gouvernementales, de services ou industrielles cherche à exploiter dans l'élaboration de leur stratégie.

Pour les entreprises commerciales en particulier, les avis des clients ou des opportunités pour lesquelles elles dépensaient des montants colossaux afin de les collecter sont désormais disponibles instantanément pour peu qu'elle identifie les plus pertinents ou les plus crédibles.

Plusieurs enquêtes (tel que celle de The PowerReviews[2] citée auparavant) ont prouvé l'effet majeur des opinions en ligne sur les décisions d'achat des consommateurs. Ces opinions, autrefois des opportunités pour lesquelles les entreprises, commerciales en particulier, dépensaient des montants colossaux afin de les collecter, sont désormais disponibles instantanément pour peu qu'elles identifient les plus pertinents ou les plus crédibles.

Avec l'émergence du web 2.0, l'accessibilité à ces informations n'est plus le problème, mais c'est plutôt leur collecte et leur traitement afin de les intégrer dans le processus de prise de décision qui sont un défi majeur. Hélas, une étude de ces données et une mise en place de modèles adéquats est nécessaire afin de surmonter certains obstacles causés par le volume de ces données, leur vélocité et leur hétérogénéité.

Cela a motivé les chercheurs du monde académique et industriel à s'intéresser à la fouille d'opinions. Aussi appelée analyse des sentiments, ce terme fait référence à l'extraction et la classification automatique d'opinions selon leur polarité.

Dans ce qui suit, nous allons tous d'abord définir les concepts clés de l'analyse des sentiments ainsi que les étapes de ce processus (Section 1.2 et 1.3). Nous poursuivrons dans la section 1.4 par la classification de la fouille d'opinion selon plusieurs critères dont

[1] https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html
[2] https://www.powerreviews.com/insights/power-of-reviews-survey-2021/

les niveaux de granularité et les objectifs de l'analyse. Nous nous intéresserons après par les approches adoptées (1.5), les jeux de données disponibles (1.6) ainsi que les métriques d'évaluations utilisées (1.7) dans ce domaine. Et pour finir, nous conclurons par une discussion des défis rencontrés ainsi que les axes de recherches futurs (Section 1.8).

1.2 Définitions

Dans ce qui suit, nous allons définir certains concepts de base nécessaires à la compréhension de notre domaine de recherche, principalement, l'opinion, l'E-réputation et la fouille d'opinion.

1.2.1 Opinion

Nous commençons par la définition de la notion qui sera le cœur de nos études, **l'opinion**. Selon le dictionnaire Larousse[3], Une opinion est un *« Jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense »* ; ou l'*« Ensemble des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc. »*. Selon le dictionnaire Le Robert[4] quant à lui, l'opinion, c'est la *« Manière de penser, de juger »*, c'est l'avis, la conviction, la croyance, l'idée, le jugement, la pensée et le point de vue.

En ce qui est du domaine informatique, une opinion peut être caractérisée comme une croyance sur des sujets généralement considérés comme subjectifs ; c'est le résultat d'une émotion ou d'une interprétation des faits, elle peut être positive, négative ou neutre (Ghelani et Bhalodia 2017). Similairement, l'étude de (Padmaja et Fatima 2013) définit une opinion comme un jugement de personnes formé sur une chose particulière et n'est pas nécessairement basé sur des faits ou des connaissances.

1.2.1.1 Composantes d'une opinion :

Une opinion fait généralement référence à la composition des éléments suivants :

- **Une entité cible**, ou *Opinion Target* en anglais, qui est la cible sur laquelle porte l'opinion, elle peut être un produit, un service, un sujet, un problème, une personne, une organisation ou un événement (B. Liu 2012). Dans l'exemple qui suit, nous prenons un "Laptop Asus" Comme entité cible de notre opinion.

- **Un aspect**, vu que l'opinion peut soit porter sur l'entité cible en général et l'ensemble de ces caractéristiques (Ce qu'on appelle aspect « GENERAL »), soit elle se concentre sur une de ses propriétés, un de ses attributs ou même un de ses composants. La capacité de stockage, celle de la batterie, la résolution de l'écran, etc sont toutes des aspects de l'entité "Laptop Asus".

Posons ces deux exemples d'opinion portant sur l'entité "Laptop Asus" :

Exemple 1 : *"I love everything about this laptop, from its screen's resolution to its speed to its sleek design. Marvellous machine 10/10"*

Exemple 2 : *"I'm not too fond of this laptop's graphical card. When I'm gaming on ultra it makes my games lag which cost me few victories here and there..."*

[3] https://www.larousse.fr/dictionnaires/francais/opinion/56197
[4] https://dictionnaire.lerobert.com/definition/opinion

Le premier exemple critique l'ensemble des caractéristiques de l'entité "Laptop Asus" en général, en revanche, le deuxième porte seulement sur l'aspect 'Carte graphique' de l'entité.

- **Un titulaire**, ou *Opinion Holder* en anglais, l'auteur, le porteur, le détenteur ou la source de l'opinion : c'est la personne/organisation ayant exprimé l'opinion.

- **Un instant**, c'est le moment où l'opinion a été exprimée par la source (l'auteur).

- **Une orientation**, également appelée orientation de sentiments, orientation sémantique, ou polarité, sur l'aspect de l'entité. La polarité peut être définie par des catégories telles que « positif », « neutre », et « négatif », ou encore par des nombres dénotant le degré de positivité ou de négativité.

L'exemple 1 présenté précédemment exprime une orientation positive envers l'aspect général de l'entité "Laptop Asus", par contre, l'exemple 2 est de polarité négative concernant l'aspect "Carte graphique" de l'entité "Laptop Asus".

1.2.1.2 Différents types d'opinions :

D'après (B. Liu 2020), les opinions peuvent être classées selon de nombreuses dimensions. Dans cette partie, nous discuterons de certaines des classifications principales :

- **Individuelle ou Collective :**

Une première dimension de classification des avis que nous pouvons conclure dès la définition de l'opinion dans les différentes encyclopédies (tel que celle de Larousse citée auparavant) est celle concernant le titulaire, ou l'auteur, de celle-ci.

Dans un premier temps, l'opinion est définie au niveau de l'individu comme une pensée d'une seule personne, c'est le cas des avis en ligne critiquant les différents produits ou services offerts sur des sites tel que Yelp, TripAdvisor, Amazon… La seconde définition fait référence à une opinion partagée entre un groupe de personnes ayant les mêmes idéologies sociales, politiques, économiques, etc.

- **Régulière ou Comparative :**

Dans la littérature, une opinion **régulière** est souvent désignée simplement comme une opinion. C'est une pensée exprimée envers une certaine entité, donc « I really enjoyed this experience, i'm defenitely coming again ! » est une opinion régulière. Comme nous le verrons dans le titre suivant, ce type d'opinion a deux sous-types principaux : Direct ou Indirect (B. Liu 2006, 2011, 2012, 2020).

L'opinion **comparative**, quant à elle, exprime une relation de similitudes ou de différences entre deux ou plusieurs entités et/ou une préférence du détenteur de l'opinion basée sur certains aspects communs des entités (Jindal et B. Liu 2006a,b ; B. Liu 2012, 2020). Par exemple, les phrases « Product A is much better than product B » et « Product A is the best » expriment deux avis comparatifs. Une opinion comparative est généralement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe, mais pas toujours (par exemple, « I prefer using product A instead of product B.»).

- **Directe ou Indirecte :**

Revenons aux opinions régulières, comme nous l'avons déjà mentionné, ces dernières peuvent être classifiées selon (B. Liu 2006, 2011, 2012, 2020) sous deux sous-types principaux : une opinion (régulière) directe ou indirecte. Une opinion **directe** est exprimée

directement sur une entité ou un aspect de celle-ci, contrairement à l'opinion **indirecte** qui se base plutôt sur l'effet positif ou négative de l'entité (ou d'un de ses aspects) sur d'autres entités.

Par exemple, l'opinion « This laptop's screen resolution is great. » est une opinion directe portant sur l'aspect de la résolution de l'écran de l'entité laptop. Tandis que l'opinion « ...The product gave me extreme rashes and my face was itchy for the next couple of days... » (Exemple extrait du dataset (Jessica Li 2018)) exprime l'effet indésirable du produit (Adapalene / benzoyl peroxide) sur une autre entité (la peau du titulaire) et donc elle exprime indirectement un avis négatif sur l'entité (le produit en question).

- **Implicite ou Explicite :**

Le dernier type d'opinion dont nous allons discuter classe les opinions (régulières ou comparatives) comme implicites ou explicites (B. Liu 2012).

Une opinion (régulière ou comparative) est dite **explicite** quand elle exprime une déclaration subjective en utilisant des mots d'opinion. Par exemple, « This product is **terrible** and it left a **horrendous** taste after I took it, it made me feel even more nauseous than before » exprime une opinion négative sur le produit à travers les mots d'opinion "Terrible", "Horrendous",...

Cependant, Une opinion **implicite** est une déclaration objective (régulière ou comparative) exprimant, généralement, des faits désirables ou indésirables, e.g. l'opinion "Bactrim cleared my skin up in less than a week." (extraite du dataset (Jessica Li 2018)) ne contient aucun mot d'opinion explicite, mais il est facile de déduire qu'elle exprime un effet désirable (élimination de l'acné) et donc qu'elle est positive.

Comme il est plus facile de détecter les opinions explicites que les opinions implicites, la plupart des recherches présentes dans la littérature se sont concentrées sur les avis explicites (B. Liu 2012).

Notons que dans la suite de ce mémoire, le terme « Opinion » ferra référence aux opinions régulières. En addition, le tableau 1.1 ci-dessous est un tableau récapitulatif contenant des exemples d'opinions de différentes catégories afin de mieux comprendre les dimensions de classification citées dans cette partie.

Tab. 1.1 : Exemples de types d'opinion

Document	**Type d'opinion**
I **love** this product ! it's quite **cheap** and its quality is **great**.	Positive, Directe, Explicite
This serum should be removed from the market before influencers start spreading it around !!!	Négative, Directe, Implicite
This product made my *skin* look **Amazing**, my *complexion* looks **better** than ever !	Positive, Indirecte, Explicite
After using this serum, my *skin* became oilier and I got my first breakout in years !	Négative, Indirecte, Implicite

1.2.2 L'analyse des sentiments

Il existe plusieurs termes qui renvoient à l'analyse des sentiments (Sentiment analysis), dont « Opinion mining », « Review mining », « Subjectivity analysis », « Sentiment AI », et d'autres termes qui sont maintenant tous connus sous l'appellation d'analyse des sentiments. Le terme « Sentiment Analysis » est apparu pour la première fois dans (Nasukawa et Yi 2003), alors que « Opinion mining » a été utilisée pour la première fois dans l'article de (Dave et al. 2003). Il existe certaines différences entre toutes ces dénominations, mais elles restent insignifiantes vu qu'elles réfèrent toutes à l'analyse des sentiments subjectifs dans un texte. Toutefois, le terme « Sentiment analysis » (Analyse des sentiments) est plus couramment utilisé dans l'industrie tandis que le terme « Opinion mining » (Fouille d'opinion) est fréquemment utilisé dans le milieu académique (B. Liu 2012).

L'analyse des sentiments, ou la fouille d'opinion, est un terme qui fait référence au domaine qui croise plusieurs disciplines telles que le traitement automatique du langage naturel (TALN ou NLP), la recherche d'information (RI), la fouille de texte ainsi que le web mining.

Le but principal de l'analyse des sentiments est de catégoriser les opinions exprimées au sein d'un texte selon le critère d'orientation (ou polarité) dans, généralement, trois classes distinctes Positive, Négative ou Neutre. (B. Liu 2012, 2020) définit l'analyse des sentiments, ou la fouille d'opinion, comme le domaine d'étude qui analyse les opinions, les sentiments, les appréciations, les attitudes et les émotions des personnes envers les entités et leurs attributs exprimés dans un texte écrit.

1.3 Processus de l'analyse des sentiments

L'analyse de sentiment peut aussi être définie comme un processus, qui prend en entrée une entité cible et donne en sortie un résumé des opinions vis-à-vis de l'entité ou de ses aspects (Dave et al. 2003).

Le processus de fouille d'opinion le plus cité dans la littérature est celui du professeur Bing Liu, de l'université de l'Illinois à Chicago (UIC) [5], (B. Liu 2012, 2020), il est considéré comme le noyau principal du pipeline de l'opinion mining. Il est composé de trois étapes principales (1.1) :

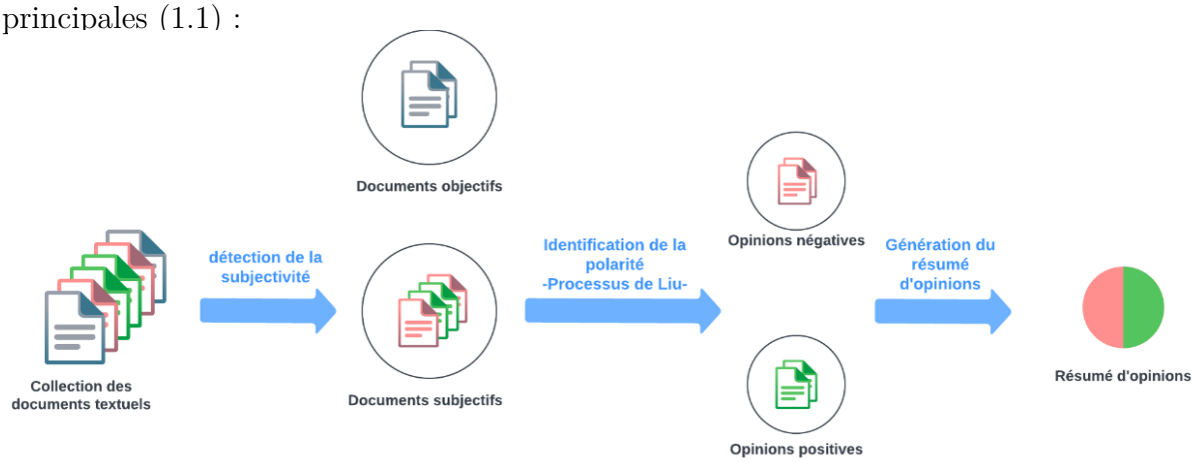


Fig. 1.1 : Le processus d'analyse des sentiments

[5] http://www.uic.edu/

- Dans un premier temps, après la collecte des documents textuels pertinents vis-à-vis de l'entité cible, ces derniers sont analysés afin de détecter leur **subjectivité**.

- Ensuite, les documents identifiés comme subjectifs sont analysés à leur tour afin d'identifier leur polarité en se basant sur un processus proposé par (B. Liu 2012, 2020).

- Pour pouvoir enfin générer un résumé, visuel ou textuel, des opinions exprimées envers la cible.

Nous allons maintenant décrire le processus d'identification de la polarité comme l'a formalisé le professeur Bing Liu (B. Liu 2012, 2020).

On commence par transformer le texte non structuré (l'opinion) en une forme plus structurée afin de pouvoir l'exploiter. Dans son livre « Sentiment Analysis and Opinion Mining » (B. Liu 2012), Liu définit une opinion par un quintuple (e_i, a_{ij}, s_{ijkl}, h_k, t_l), où :

- $\mathbf{e_i}$: Le nom de l'entité cible.

- $\mathbf{a_{ij}}$: Un aspect de l'entité cible e_i. Si l'opinion porte sur l'entité e_i elle-même dans son ensemble, l'aspect a_{ij} est « GENERAL ».

- $\mathbf{s_{ijkl}}$: Le sentiment, ou orientation, de l'opinion sur l'aspect a_{ij} de l'entité e_i.

- $\mathbf{h_k}$: Le titulaire ou la source de l'opinion. Il peut faire référence à un individu ou un groupe d'individus.

- $\mathbf{t_l}$: L'instant où l'opinion a été exprimée par la source d'opinion h_k.

Notons que dans cette définition, nous utilisons volontairement des indices pour souligner le fait que les cinq composants du quintuple doivent correspondre les unes aux autres. Donc un quintuple $\mathbf{s_{ijkl}}$ fait référence au sentiment de l'opinion exprime par h_k à l'instant t_l concernant l'aspect a_{ij} de l'entité e_i.

Considérons maintenant un document subjectif D, l'objectif est d'identifier la polarité de toutes les opinions qui y sont exprimées. Chacune de celles-ci sera représentée par le quintuple de Bing Liu décrit ci-dessus. À partir d'un document subjectif **D** :

1. Extraire les entités et grouper les synonymes éventuels dans des clusters d'entités. Chaque cluster représente une entité e_i.

2. Pour chaque entité e_i, extraire tous ses aspects associés. Chaque cluster représente un aspect a_{ij} pour une entité e_i.

3. Extraire le titulaire h_k de l'opinion ainsi que la date t_l de son émission.

4. Pour chaque aspect a_{ij} d'une entité e_i, déterminer les orientations des opinions. La polarité s_{ijkl} peut être positive, négative ou neutre.

5. Générer l'ensemble des tuples (e_i, a_{ij}, s_{ijkl}, h_k, t_l) qui représentent toutes les opinions exprimées dans le document D, sur la base des résultats des tâches précédentes.

1.4 Types d'analyse de sentiments

La classification automatique de la polarité peut être catégorisée en fonction de diverses perspectives (Yadollahi et al. 2017), cela peut expliquer le fait qu'il existe plusieurs termes faisant référence à la fouille d'opinion. Dans cette section (1.4), nous allons classifier l'analyse de sentiments selon deux critères principaux : d'abord la granularité, puis la tâche

effectuée.

1.4.1 Critère 01 : Les niveaux de granularité

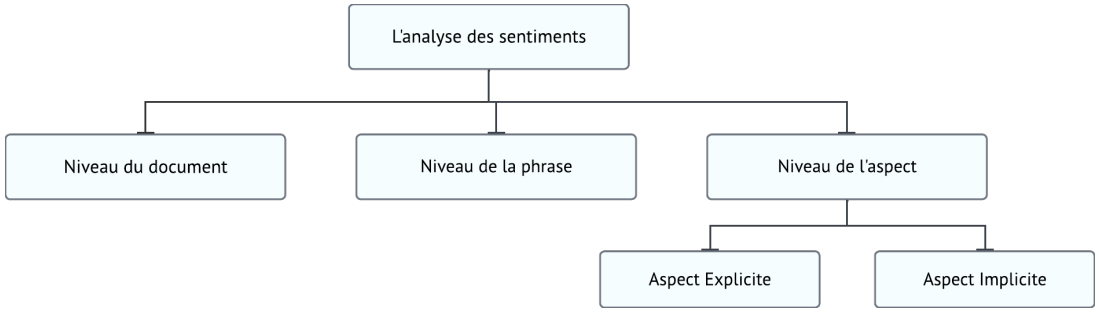


Fig. 1.2 : Niveaux de granularité de l'analyse des sentiments

La fouille d'opinion peut être effectuée à différents niveaux de granularité du texte de l'opinion (B. Liu 2012, 2020). Dans ce qui suit, nous présenterons les différents niveaux de l'analyse des sentiments comme illustré dans la figure 1.2 ci-dessus.

1.4.1.1 Niveau du document

L'objectif de la fouille d'opinion au niveau du document (macroscopique) est de déterminer la polarité d'un texte entier. L'hypothèse est que *le texte, qu'il soit court ou long, n'exprime qu'une opinion vis-à-vis d'une seule entité* (un seul produit, une seule marque…), par conséquent, ce type de fouille d'opinion ne s'applique pas aux documents qui évaluent ou comparent plusieurs entités, pour lesquelles une analyse plus fine est nécessaire.

L'analyse au niveau du document est importante, car elle donne une vue d'ensemble sur une entité cible, mais elle a ses lacunes (B. Liu 2012).

1.4.1.2 Niveau de la phrase

Au niveau mésoscopique (niveau de la phrase), la polarité est calculée individuellement pour chaque phrase subjective contenue dans un document. Et comme l'affirme (B. Liu 2012, 2020), ce type d'analyse est étroitement lié à ce que l'on appelle la « classification de subjectivité » dont le but est de classer si une phrase est subjective ou objective.

L'analyse au niveau de la phrase se base sur l'hypothèse que *chaque phrase dans un texte ne contient qu'un seul avis concernant une entité unique*. L'un de ses inconvénients est le fait que le contexte entourant la phrase peut être assez définitif sur son sens et donc sa polarité : " This experience has left me speechless" pourrait être une déclaration positive ou négative selon la contexte dans lequel il est utilisé.

1.4.1.3 Niveau de l'aspect

La fouille d'opinion au niveau de l'aspect est une analyse plus fine que les niveaux précédents, en se basant sur l'hypothèse qu'*une opinion possède une seule polarité et cible un aspect d'une entité (produit, service…).*

Ce type d'analyse d'opinion vise à identifier les différents aspects ciblés d'une entité et de définir le sentiment que la source de l'opinion porte à son égard. Par exemple une opinion qui déclare " I love this product ! " est positive certes, mais elle n'apporte pas de détails e.g. des points que le client préfère vis-a-vis du produit. Aussi, même si une opinion est positive, elle peut contenir des aspects négatifs sur l'entité et, par exemple,

pour un consommateur intéressé par la qualité de la caméra d'un smartphone, l'opinion " I love this phone, it's performant and great, even if the quality of its camera sucks" l'influencera probablement à ne pas acheter ce smartphone, même si la polarité globale de l'opinion est positive. Voici deux exemples d'opinions C_1 et C_2 :

- **C_1 :** « I really like this laptop ! I got it a few weeks ago and had an enjoyable experience using it. Even though I think it's a bit overpriced... The screen's resolution is great, the design is quite stylish and I really enjoyed that I can run my favorite animation programs without lag ! However I'll have to mention that it weights a ton, and it's even smaller than what I'm used to. »

- **C_2 :** « I've had this phone for a while and it's the worst investment I've ever made. The pictures' quality is low, the battery doesn't hold for long and it lags a lot ! It's too expensive for such low performances... I won't lie though, the design was sleek and quite cute. But overall I don't think that this product is worth all the hype it got on social media. »

Considérons les comme entrée de trois modèles de fouille d'opinion **M_1**, **M_2** et **M_3**. Chacun de ces modèles implémente un des niveaux de granularité décrits auparavant respectivement.

Le modèle **M_1** traite le document C_i dans son entièreté. Il détermine, en sortie, la polarité de C_1 comme positive et celle de C_2 comme négative.

Le modèle **M_2**, qui implémente une fouille d'opinion au niveau de la phrase, extrait d'abord les phrases subjectives du texte du commentaire C_i avant de déterminer sa polarité. Par exemple, la polarité de la phrase « However I'll have to mention that it weights a ton » du commentaire C_1 est identifiée comme étant négative, tandis que la phrase « I won't lie though, the design was sleek and quite cute. » du commentaire C_2 est de polarité positive.

Pour finir, le modèle **M_3**, quant à lui, implémente une fouille d'opinion basée sur l'aspect de l'entité cible, i.e. il détecte d'abord les aspects présents dans le texte du commentaire C_i avant d'identifier la polarité des opinions les concernant. En ce qui est de l'aspect "Weight" (Poids) de l'entité **Laptop** dans l'exemple C_1, qui est de type **implicite**, le modèle déduit que les opinions envers elle sont de polarité négative. Contrairement à l'aspect "pictures' quality" (la qualité des images) de l'entité **Phone** du commentaire C_2, qui est de nature explicite et dont l'orientation est positive.

1.4.2 Critère 02 : La tâche effectuée

Généralement, l'analyse des sentiments se focalise sur l'identification de l'orientation d'un texte, cependant, elle peut aller au-delà de cette polarité pour détecter les sentiments et les émotions spécifiques (la colère, la joie, la tristesse, etc.) et même des intentions (intéressé ou pas intéressé) de l'auteur du texte. Selon la manière dont nous voulons exploiter les résultats de l'analyse, nous pouvons l'ajuster afin de répondre à ces besoins. L'analyse des sentiments peut donc être classifiée selon les tâches qu'elle vise à accomplir, parmi ces tâches, nous pouvons citer :

1.4.2.1 Analyse affinée

L'analyse fine des sentiments fournit un niveau de polarité plus précis en le décomposant en d'autres catégories, généralement de **très positives** à **très négatives**. Cela

peut être considéré comme l'équivalent d'une opinion sur une échelle de 5 étoiles (Très positive 5 étoiles – Positive 4 étoiles – Neutre 3 étoiles – Négative 2 étoiles - Très négative 1 étoile).

1.4.2.2 Analyse des sentiments multilingue

L'analyse des sentiments multilingue est souvent très difficile, car elle implique beaucoup de prétraitement et de ressources. C'est considéré comme un défi courant dans le domaine. Selon les langues, les ressources disponibles peuvent être limitées et nécessiter la création de corpus personnalisés. Par exemple, la langue **anglaise** est une langue dont les ressources sont largement disponibles, tandis que l'**arabe**, quant à elle, a des corpus limités et présente d'autres difficultés telles que les multiples dialectes disponibles et la romanisation de l'écriture.

1.4.2.3 Détection des émotions

La détection des émotions vise à identifier des émotions spécifiques plutôt que la positivité et la négativité. Les exemples pourraient inclure le bonheur, la frustration, la colère et la tristesse (Nandwani et Verma 2021).

Un des défis rencontrés dans ce type d'analyse est le fait que certains mots et expressions peuvent être utilisés afin d'exprimer différentes émotions. Par exemple, l'expression " This experience had me shaking in my seat" renvoie, à première vue, à la **peur** cependant, dans un autre contexte, elle peut faire référence à la **surprise** et l'**enthousiasme**. Cette dépendance au contexte rend l'utilisation de simples lexiques (i.e. Liste de mots avec les émotions qu'ils véhiculent) insuffisante.

1.5 Approches adoptées dans l'analyse des sentiments

L'analyse des sentiments est un domaine multidisciplinaire qui comprend l'étude de divers domaines. Les différentes études de l'art (Ligthart et al. 2021b ; Ramesh et Weber s. d. ; K. Ravi et V. Ravi 2015) classifient les approches adoptées dans l'analyse des sentiments en trois catégories : Basées sur l'apprentissage automatique, Basées sur le lexique et des méthodes Hybrides comme l'illustre la Figure 1.3.

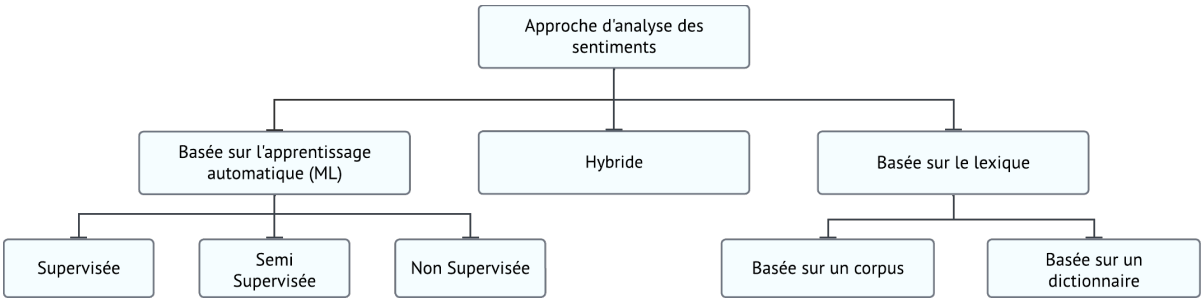


Fig. 1.3 : Taxonomie des approches adoptées dans l'analyse des sentiments

1.5.1 Approches basées sur l'apprentissage automatique

La classe des approches basées sur l'apprentissage automatique (ML) se divise à son tour en trois catégories : l'apprentissage supervisé, semi-supervisé et non supervisé.

1.5.1.1 Approches supervisées

La classification supervisée est l'approche la plus connue et la plus utilisée, elle consiste à entraîner un classificateur à partir d'un ensemble de données étiquetées permettant, par

la suite, la classification de toute nouvelle entrée selon son orientation (Prédire la polarité d'une nouvelle donnée). Selon (Mehta et Pandya 2020), parmi les approches appartenant à cette catégorie on peut citer :

a) Les classificateurs probabilistes : qui visent à prédire ou anticiper, étant donné une observation en entrée, une distribution de probabilité sur un ensemble de classes (plutôt que de ne fournir que la classe la plus probable à laquelle l'observation devrait appartenir). Parmi ces classificateurs, on trouve : Naïve Bayes, les réseaux bayésiens, l'entropie maximale, etc.

Par exemple, **Naïve Bayes** (Joyce 2003) est l'algorithme le plus simple à appliquer sur un ensemble de données. Comme son nom l'indique, cet algorithme suppose que les caractéristiques sont indépendantes les unes des autres et qu'il n'y a pas de corrélation entre elles. Cependant, ce n'est pas le cas dans la vraie vie. Cette hypothèse naïve selon laquelle les caractéristiques ne sont pas corrélées est la raison pour laquelle cet algorithme est appelé "naïf". En tant que tel, cela peut s'avérer être un inconvénient à l'utilisation de cet algorithme, car cette hypothèse rend l'algorithme de bayes naïf moins précis que les algorithmes plus compliqués.

En tant qu'outil de classification, Naïve Bayes est considéré comme efficace et simple, et sensible à la sélection des caractéristiques (Tiffani 2020). Dans (Tiffani 2020), les auteurs ont appliqué le classificateur Naïve Bayes sur des données prétraitées. D'abord les données ont été nettoyées (transformation des mots en minuscule, suppression des mots vides et radicalisation) puis parcourues par un N-gramme (Ils ont testé l'Unigramme, le Bigramme et le Trigramme) avant de les classer à l'aide du classificateur. Les résultats étaient plutôt prometteurs, car l'exactitude (accuracy) du modèle où Naïve Bayes était appliqué à l'unigramme était de 81,30% et celles des deux autres n'étaient pas inférieures à 71%. Les auteurs de (Shah 2021) ont appliqué ce classificateur avec un modèle bag-of-word sur un dataset publié par Amazon[6] sur kaggle[7] contenant des avis sur des produits smartphones (Vimal et Tarun 2019) résultant en une exactitude de 91,76% et un F1-score de 94%.

b) Les classificateurs linéaires : qui utilisent une combinaison linéaire des caractéristiques de l'opinion pour prendre une décision de classification et identifier la classe à laquelle elle appartient. On peut y trouver : les classificateurs SVM et ceux basés sur les réseaux de neurones.

En ce qui est des classificateurs **SVM**, en anglais Support Vector Machine, ce sont des modèles linéaires utilisés pour la classification et la régression. L'idée de SVM est simple : l'algorithme crée une ligne ou un hyperplan qui sépare les données en N classes, en l'appliquant pour un problème d'analyse des sentiments N=2 (ou trois si on prend en considération la classe **Neutre**). La figure 1.4 illustre un exemple de classification d'un ensemble de données en deux classes (Rouge et Bleu) en appliquant l'algorithme SVM dans le but de tracer une frontière de décision (Hyperplan de marge maximale) de telle manière que la séparation entre les deux classes soit aussi large que possible.

En appliquant un classificateur SVM sur des tweets du dataset " Twitter US airline sentiment", les auteurs de (Ravi Kumar et al. 2021) ont abouti à une exactitude de

[6] https://www.amazon.com/
[7] https://www.kaggle.com/

74.24% et un F1-score de 74%. D'autre part, les auteurs de (Shah 2021) ont appliqué SVM avec un modèle bag-of-word (comme pour Naïve Bayes) et ont obtenu une exactitude de 93,58% et un F1-score de 96%, un meilleur résultat que celui obtenu avec Naïve Bayes.

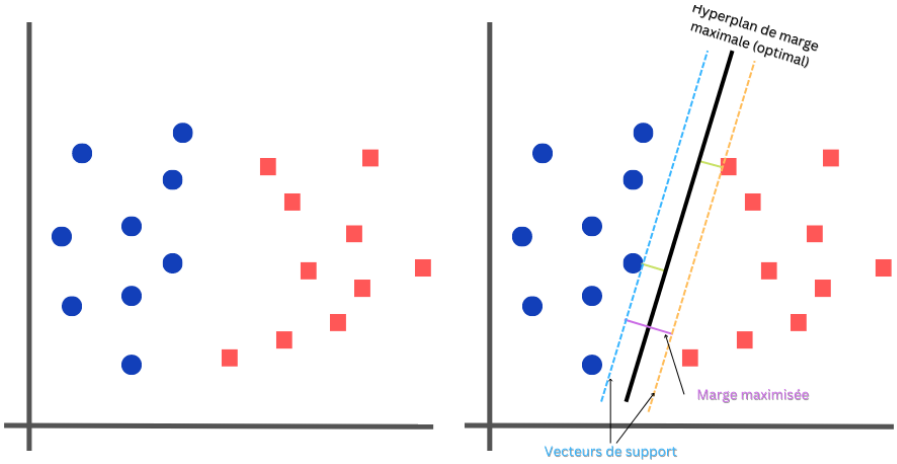


Fig. 1.4 : Exemple de classification en utilisant l'algorithme SVM (Support Vector Machine)

Une autre méthode populaire est celle des **réseaux de neurones**, qu'ils soient artificiels (ANN), convolutifs (CNN), récurrents (RNN)... un réseau de neurones est constitué de couches interconnectées de nœuds (petites unités) qui effectuent des opérations mathématiques afin de détecter des modèles dans les données. Les réseaux de neurones sont construits de manière à imiter le fonctionnement des neurones humains. De nombreuses architectures basées sur cette méthode ont obtenu d'excellentes performances dans le domaine de l'analyse des sentiments (Ravi Kumar et al. 2021 ; Shah 2021).

Par exemple, dans (Ravi Kumar et al. 2021), les résultats obtenus en implémentant les réseaux de neurones artificiels (ANNs) étaient plus performants que les autres modèles proposés (SVM, Arbres de décision), avec une exactitude de 75,99% et un F1-score de 74%. Pour (Shah 2021), la meilleure performance a été obtenue en utilisant un réseau de neurones convolutif (CNN) avec FastText (Bojanowski et al. 2017) (une extension du word-embedding Word2vec (Mikolov et al. 2013)), avec une exactitude de 94,62% et un F1-score de 96%.

c) Les classificateurs d'arbre de décision : où les données sont continuellement divisées en fonction d'un certain paramètre, résultant, au final, en un arbre composé d'un nœud et de feuilles. Le principal avantage de cette classe est sa capacité à utiliser différents sous-ensembles de fonctionnalités et règles de décision à différentes étapes de la classification. Sa simplicité est aussi une des raisons de sa popularité.

Dans (Ravi Kumar et al. 2021), l'utilisation des arbres de décision a donné une exactitude de 67.38% et un F1-score de 63%, c'était le modèle le moins performant comparé aux deux autres classificateurs utilisés (SVM et ANN comme cité précédemment).

d) Les classificateurs basés sur des règles : qui consistent en un ensemble de règles « SI-ALORS » (IF-THEN) obtenues en appréhendant statistiquement les données d'apprentissage. Nous pouvons produire les règles en fonction de nos besoins lors de la phase d'apprentissage (Mehta et Pandya 2020). Les approches supervisées surpassent parfois les performances d'autres approches d'apprentissage automatique et sont plus

faciles généralement à implémenter. Cependant, il faut noter que l'inconvénient majeur de ces approches est la dépendance aux ensembles de données annotées et la difficulté de se procurer ces ensembles. De plus, vu cette dépendance, la qualité des données annotées impacte considérablement la performance du modèle de classification.

1.5.1.2 Approches semi-supervisées

Pour la classification semi-supervisée, les modèles sont entraînés avec des ensembles de données non étiquetées augmentés par un ensemble, souvent limité, d'exemples de données étiquetées (Van Engelen et Hoos 2020). Ce type d'approches délivre des résultats assez satisfaisants et nécessite moins d'effort humain par rapport aux méthodes d'apprentissage supervisées.

Les auteurs de (N. Li et al. 2020) proposent un cadre d'apprentissage multitâche semi-supervisé (appelé **SEML**) pour l'analyse des sentiments basée sur les aspects. Comme cette analyse implique à la fois l'extraction d'aspects et la classification des sentiments d'aspect, les modèles supervisés utilisés pour la résoudre nécessitent un grand nombre d'avis étiquetés qui sont très coûteux ou indisponibles. Pour cela, les auteurs ont opté pour une technique semi-supervisée (SEML) qui a 03 caractéristiques clés : une formation à vues croisées (Cross-View Training) pour permettre un apprentissage de séquence semi-supervisé sur un petit ensemble d'avis étiquetés et un grand ensemble d'avis non étiquetés du même domaine, une exécution en deux sous-tâches (l'extraction d'aspects et la classification des sentiments d'aspect) simultanément en utilisant 03 couches neuronales récurrentes bidirectionnelles empilées pour apprendre les représentations des opinions et enfin une unité récurrente attentive à fenêtre mobile pour ces couches afin d'améliorer l'apprentissage de la représentation et la précision de la prédiction.

Les résultats expérimentaux montrent que SEML surpasse de manière significative les modèles de l'état de l'art sur les mêmes jeux de données ; pour la tâche d'extraction d'aspects, ils ont réussi à atteindre un F1-score de 83,37% sur D_L (des opinion sur des laptops), et 78,24% sur D_R (des opinion sur des restaurants), tandis que pour la tâche de classification des sentiments, ils ont atteint un F1-score de 59,85% sur D_L et de 69,17% sur D_R.

1.5.1.3 Approches non supervisées

La classification non supervisée vise à tirer des conclusions à partir d'un ensemble de données en entrée où ces données ne sont pas étiquetées, le résultat de cette classification serait le regroupement de ces données en des clusters selon leurs ressemblances. Ces méthodes aident à découvrir les modèles cachées dans les données.

Rich-resource data-based sentiment analysis has been well-developed, whereas the more challenging and practical multi-source unsupervised oneis seldom studied. This is because the latter encounter multiple problems which are mainly locate in the lack of supervision information, the semantic gaps among domains, and the loss of knowledge.

L'analyse de sentiment basée sur des données riches et ressources bien développées a été bien étudiée, alors que l'analyse non supervisée multi-source, plus difficile et pratique, est rarement étudiée. Cela est dû à de nombreux problèmes rencontrés par cette dernière, principalement dus au manque d'informations de supervision, aux écarts sémantiques entre les domaines et à la perte de connaissances.

L'analyse de sentiment basée sur des données riches et ressources bien développées a été bien étudiée, alors que l'analyse non supervisée multi-source, plus difficile et pratique, est rarement étudiée. Cela est dû à de nombreux problèmes rencontrés par cette dernière, principalement dus au manque d'informations de supervision, aux écarts sémantiques entre les domaines et à la perte de connaissances.

L'analyse de sentiment basée sur des données est riche en ressources et a été bien développée, cependant l'analyse non supervisée, particulièrement multi-source, est rarement étudiée. Cela est dû à de nombreux problèmes rencontrés par cette dernière, principalement dus au manque d'informations de supervision, aux écarts sémantiques entre les domaines et à la perte de connaissances. Pour atténuer à ces problèmes, (Dai et al. 2021) ont proposé un cadre d'adaptation de domaine en deux étapes. La première est une méthode d'adaptation de domaine sélective, qui transfère les connaissances du domaine source le plus proche. La deuxième, quant à elle, est une méthode d'ensemble axée sur la cible, dans laquelle les connaissances sont transférées via une méthode d'ensemble bien conçue. Les expériences approfondies montrent des résultats prometteurs surpassant les méthodes concurrentes non supervisés.

1.5.2 Approches basées sur le lexique

Ce sont les approches traditionnelles qui consistent à calculer le sentiment d'un document à partir de l'orientation sémantique des mots ou des phrases qui le composent, ces orientations sont documentées dans des lexiques de sentiments. Nous pouvons en distinguer deux catégories selon la méthode de création des lexiques de sentiments :

1.5.2.1 Approches basées sur un dictionnaire

Les approches basées sur un dictionnaire (par exemple WordNet[8]) commencent par la collecte manuelle d'un ensemble de mots de sentiments, appelés graines, puis par le développement itératif de cet ensemble à partir des synonymes et des antonymes énumérés dans le dictionnaire, et se termine quand plus aucun nouveau mot n'est trouvé par une inspection manuelle si voulue.

Ces techniques donnent de meilleurs résultats pour l'usage général et non pas pour les cas spécifiques à un domaine précis (B. Liu 2020). Ceci puisque les orientations des mots collectés de cette manière sont générales ou indépendantes du domaine et du contexte. De nombreux mots de sentiment ont, en fait, des orientations dépendant du contexte. Par exemple, dans l'opinion " These speakers are so quiet " le mot 'quiet' (silencieux) est négatif tandis que dans " This neighborhood is so quiet at night " le mot 'quiet' exprime une polarité positive. Ainsi, l'orientation du mot "quiet' dépend du contexte.

À titre d'exemple, nous pouvons citer (Hossen et Dev 2021), où ils ont utilisé une approche basée sur un dictionnaire sur un ensemble de données de critiques de films, ils ont utilisé les bibliothèques micro-WordNet-Opinion 3.0 [9]. Ils ont eu de bons résultats : une exactitude de 72% dépassant les approches traditionnelles du lexique (qui atteignaient 51%). On peut également citer (Özçelik et al. 2021) où ils visaient à créer un nouveau lexique de polarité 'HisNet'. Cette recherche montre à quel point les ressources dans un tel domaine sont rares lorsqu'il s'agit de langues autres que l'anglais et, comme ils l'ont

[8] https://wordnet.princeton.edu

[9] https://github.com/aesuli/SentiWordNet/blob/master/papers/Micro-WNOp.pdf

mentionné, il est vain d'essayer de traduire des lexiques anglais existants vers une langue cible.

1.5.2.2 Approches basées sur un corpus

Les approches basées sur un corpus, quant à elles, peuvent être adaptées à des domaines spécifiques. Ces méthodes commencent par une liste de mots de sentiments, souvent à usage général, puis en découvre d'autres en addition à leurs orientations à partir d'un corpus de domaine, en utilisant une approche **statistique** ou **sémantique** (Aminuddin et al. 2021 ; B. Liu 2020). Elles sont cependant plus difficiles à mettre en place et même dans un même domaine, le problème de dépendance au contexte persiste, i.e. un mot peut avoir une orientation différente selon le contexte d'où il est extrait. Les auteurs de (Aminuddin et al. 2021) ont utilisé une approche sémantique basée sur un corpus afin d'identifier des mots-clés dans les avis collectés, les catégoriser en fonction des sentiments et identifier la polarité des lexiques (positive, neutre, négative).

1.5.3 Approches hybrides

Les approches hybrides sont celles qui visent à combiner les techniques d'apprentissage automatique et celles basées sur le lexique afin de produire des résultats optimaux en utilisant un ensemble de caractéristiques tirées de ces deux méthodes. Ainsi, les lacunes et les limites des deux approches peuvent être surmontées. Par exemple, (Putra et al. 2021) ont utilisé une méthode hybride combinant une méthode basée sur un dictionnaire (Sentiwordnet 3.0 (Baccianella et al. 2010)) et une méthode basée sur l'apprentissage automatique (SVM), leurs résultats montrent que leur méthode hybride surpasse l'approche basée sur le lexique (dictionnaire) et l'approche SVM.

1.6 Jeux de données disponibles

Afin de pouvoir entraîner et tester les approches adoptées pour résoudre le problème de classification de la polarité d'opinion, nous avons besoin de jeux de données. Certains de ces ensembles de données ne sont pas accessibles, car ils sont constitués par les entreprises qui collectent les avis de leurs clients, d'autres (les corpus académiques) sont mis à la disposition de la communauté en vue d'une évaluation comparative des modèles.

Tab. 1.2 : Quelques corpus utilisés dans les travaux d'analyse de sentiment

Dataset	Source	Volume	Domaine	Référence
Twitter US Airline Sentiment[10]	Twitter	Environ 14,600 avis	Compagnies aériennes	(Kaggle 2015)
Stanford Sentiment Treebank[11]	Rotten Tomatoes	Plus de 10,000 avis	Films	(Socher et al. 2013)
Amazon Product Reviews[12]	Amazon	233.1 million avis	Produits	(Ni et al. 2019)
Internet Movie Database (IMDB)[13]	IMDB	50.000 avis	Films	(Maas et al. 2011)
LABR[14]	GoodReads	63,257 avis	Livres	(Aly et Atiya 2013)
Sentiment140[15]	Twitter	1,600,000 avis	Produits	(Go et al. 2009)

Dans cette section, nous allons présenter des exemples de corpus académiques utilisés dans l'analyse des sentiments. Le tableau 1.2 ci-dessus présente un ensemble de jeux de données disponibles publiquement et utilisés dans les différents travaux de l'analyse des sentiments.

1.7 Métriques et méthodes d'évaluation

Afin de déterminer l'efficacité d'une méthode et de la comparer avec d'autres méthodes, nous devons l'évaluer. Pour cela, plusieurs métriques d'évaluations ont été mises en place, et quel que soit le type du corpus, nous pouvons distinguer deux types de méthodes d'évaluation : les méthodes de tests statistiques ainsi que les méthodes de retour de pertinence.

1.7.1 Méthodes de tests statistiques

Les méthodes de tests statistiques supposent que les documents du corpus ont été classés manuellement et que l'annotation a été validée par des experts. L'annotation peut aboutir à un corpus équilibré, ou un corpus déséquilibré. À cet effet, nous pouvons diviser l'évaluation des performances des modèles en deux catégories selon la distribution des instances positives et négatives :

- Pour les jeux de données équilibrés, la métrique du F-score est souvent utilisée.

- Pour les jeux de données déséquilibrés, la courbe ROC et AUC sont plus adaptées.

À noter que les métriques de test sont différentes des fonctions de perte (Loss functions). Tandis que, les fonctions de perte montrent une mesure des performances du modèle *pendant* l'entraînement de ce dernier, les métriques sont utilisées pour juger et mesurer les performances du modèle *après* l'entraînement.

1.7.1.1 Le F-score

Le problème d'analyse des sentiments revient à un problème de classification (B. Liu 2020). Pour mesurer l'efficacité d'un classificateur dans ce type de problème, trois mesures sont souvent utilisées : la précision **P**, le rappel **R** et le **F-score** (F-measure, F1 measure). Ces métriques doivent être contrôlées pendant toutes les phases de la chaîne de traitement et pas seulement dans la phase d'analyse de la polarité.

Tab. 1.3 : Table de confusion - Analyse des sentiments

Classe Réelle / Classe Prédite	Opinion positive	Opinion négative
Opinion positive	TP (Vrai Positif)	FN (Faux Négatif)
Opinion négative	FP (Faux Positif)	TN (Vrai Négatif)

Définissons d'abord les matrices de confusion. Une **matrice de confusion** est une

[10] https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment
[11] https://nlp.stanford.edu/sentiment/treebank.html
[12] https://nijianmo.github.io/amazon/index.html
[13] https://ai.stanford.edu/~amaas/data/sentiment/
[14] https://github.com/mohamedadaly/LABR
[15] https://www.kaggle.com/datasets/kazanova/sentiment140

matrice 2 x 2 montrant la distribution des performances d'un modèle de classification sur des données. Elle présente à quel point le modèle fonctionne, ce qui doit être amélioré et quelle erreur il commet. Le tableau 1.3 ci-dessus illustre la table de confusion, où :

- **TP** - Vrai positif (le résultat de classe positive (Opinion positive) correctement prédit du modèle)
- **TN** - Vrai négatif (le résultat de classe négative (Opinion négative) correctement prédit du modèle)
- **FP** - Faux positif (le résultat de classe positive (Opinion positive) mal prédit du modèle)
- **FN** - Faux négatif (le résultat de classe négative (Opinion négative) prédit de manière incorrecte du modèle)

Revenons au **F-score**. Ce terme a été introduit par Van Rijsbergen[16], c'est une mesure statistique qui prend en compte à la fois le rappel R et la précision P. La **précision** (le degré de solidité) indique le degré de vérité des résultats obtenus, tandis que le rappel R (le degré d'exhaustivité) indique leur pertinence. La mesure de F-score est la moyenne harmonique entre la précision P et le rappel R[17].

Un modèle parfait a un F-score de 1. D'une manière générale, le calcul du F-score se fait selon la formule 1.1 suivante :

$$F - score = 2 \times \frac{P \times R}{P + R} \tag{1.1}$$

La précision P et le rappel R sont calculés selon les formules suivantes :

$$P_i = \frac{nombre\ de\ documents\ correctement\ attribues\ a\ la\ classe\ i}{nombre\ de\ documents\ attribues\ a\ la\ classe\ i} = \frac{TP}{TP + FP} \tag{1.2}$$

$$R_i = \frac{nombre\ de\ documents\ correctement\ attribues\ a\ la\ classe\ i}{nombre\ de\ documents\ appartenant\ a\ la\ classe\ i} = \frac{TP}{TP + FN} \tag{1.3}$$

1.7.1.2 Courbe ROC et AUC

La « **courbe ROC** » (de l'anglais Receiver Operating Characteristic curve) est une mesure de performance d'un classificateur binaire, souvent présentée sous la forme d'un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. La courbe ROC trace le taux de vrais positifs TVP (fraction des positifs qui sont correctement détectés) en fonction du taux de faux positifs TFP (fraction des négatifs qui sont incorrectement détectés) comme représentée dans la Figure 1.5. Ces valeurs sont calculées selon les équations 1.4.

$$TVP = \frac{Vrai\ Positif}{Vrai\ Positif + Faux\ Negatif} \qquad TFP = \frac{Faux\ Positif}{Faux\ Positif + Vrai\ Negatif} \tag{1.4}$$

[16] https://en.wikipedia.org/wiki/C._J._van_Rijsbergen
[17] https://www.expert.ai/glossary-of-ai-terms/f-score-f-measure-f1-measure/

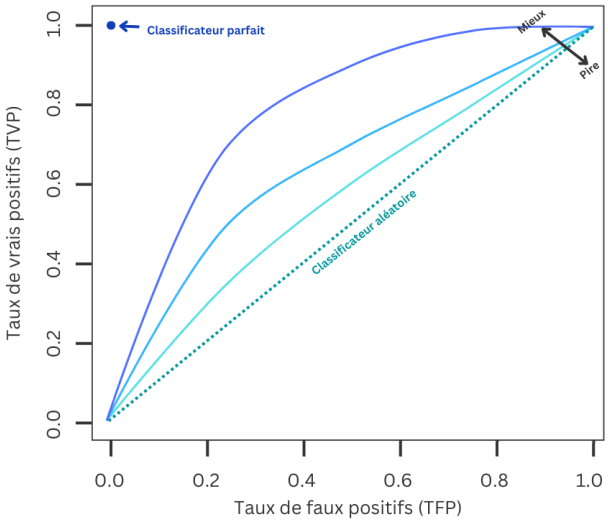


Fig. 1.5 : La courbe ROC (Receiver Operating Characteristic curve)

En ce qui est de l'**AUC** (de l'anglais Area Under the receiver operating characteristics Curve), ce terme fait référence à l'aire sous la courbe ROC. La mesure AUC fournit une mesure agrégée des performances pour tous les seuils de classification possibles. Les valeurs d'AUC sont comprises dans une plage de 0 (toutes les prédictions sont erronées) à 1 (toutes les prédictions sont correctes).

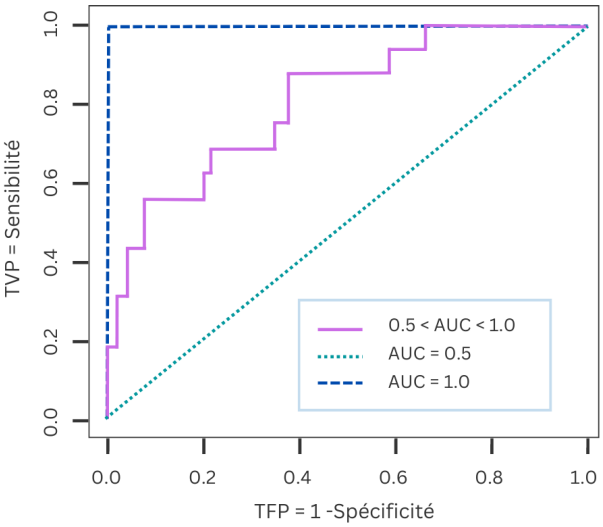


Fig. 1.6 : La métrique AUC (Area Under the receiver operating characteristics Curve)

L'aire sous la courbe ROC (AUC) peut être interprétée comme la probabilité que, parmi deux opinions choisies aléatoirement, une positive et une non-positive (négative), la valeur du marqueur soit plus élevée pour l'opinion positive que pour l'opinion négative. Nous pouvons donc conclure que, comme illustré sur la figure 1.6[18] ci-dessus, une AUC de 0,5 (50%) indique que le marqueur est non informatif, et son augmentation indique une amélioration des capacités discriminatoires, (avec un maximum de 1,0 i.e. 100%).

[18] https://www.idbc.fr/tutoriel-comment-lire-une-courbe-roc-et-interpreter-son-auc/

1.7.2 Méthodes de retour de pertinence

Le retour de pertinence (Relevance feedback) est un domaine particulier de la recherche d'information, proposé en 1971 dans les travaux de Rocchio relevance-Feedback, c'est un moyen de prendre en compte le ressenti des utilisateurs vis-à-vis des informations renvoyées après une première recherche. Ceci est basé sur l'idée que l'utilisateur est la seule personne qui sait exactement ce qu'il recherche, et qu'il est donc le mieux placé pour juger de la pertinence des informations renvoyées par un système de recherche.

Ce principe de validation a été adopté dans le domaine de la fouille d'opinions. Il est utilisé en validant les premiers résultats par les experts, puis les réinjecter dans les systèmes eux-mêmes pour qu'ils soient réévalués par les utilisateurs.

1.8 Défis et discussions

L'analyse des sentiments est un domaine émergeant et dynamique, un tel dynamisme entraîne l'apparition de défis. Parmi les défis majeurs dans le domaine de l'analyse des sentiments, on peut citer la dépendance de la langue, la dépendance du contexte, la dépendance du domaine, les fautes d'orthographe ou de frappe, le langage figuratif, le dynamisme temporel des opinions et les fausses opinions (W. Zhang et al. 2018). Dans ce qui suit, nous en détaillerons quelques-uns.

1.8.1 Dépendance du contexte

En général, tout propos subjectif est tenu dans un certain contexte, par exemple un mot donné peut être un indicateur de positivité dans un contexte et de négativité dans un autre (Chaturvedi et al. 2018 ; Ding et B. Liu s. d. ; Hussein 2018 ; W. Zhang et al. 2018). Ainsi, la prise en considération du contexte lors du pré-traitement ou le post-traitement des données est parfois cruciale pour l'obtention de résultats plus précis, surtout avec l'existence de mots polysémiques et même énantiosémique. Prenons, par exemple, ces deux expressions : « The instuctor was so enthusiastic you can tell that he is an **amateur** about this ! » et « This handmade candle is so poorly made, you can tell it was the work of an **amateur**... ». Dans la première expression, le mot 'amateur' fait référence à une personne qui aime un sujet et y est compétente, le deuxième par contre fait référence à une personne incompétente et exprime donc que l'entité cible n'est pas de bonne qualité.

1.8.2 Dépendance du domaine

La dépendance du domaine est partiellement une conséquence des changements de vocabulaire (Hussein 2018). Ainsi, la même expression peut indiquer un différent sentiment en fonction du domaine. Certaines approches, et c'est le cas de la plupart, donnent de bonnes performances dans un domaine, mais fonctionnent atrocement dans un autre. Cela reste un problème difficile à contourner pour plusieurs.

1.8.3 Fautes d'orthographe

Un autre défi qui peut apparaître durant l'analyse des sentiments est celui des fautes d'orthographe. Il en existe deux types, celles qui sont volontaires et celles qui ne le sont pas. Les erreurs délibérément commises peuvent avoir un impact sur la polarité d'un

commentaire, en fait une augmentation des voyelles et de ponctuations peut indiquer un sentiment fortement positif (par exemple « I'm soooooooooo HAPPPPY with this new product !!!! ») ou fortement négatif (par exemple « This has been the WOOOOORST EXPERIENCE EVEEERRR »). En revanche, les erreurs causées par les méconnaissances ou les fautes de frappes n'ont aucun effet sur la polarité si on arrive à distinguer le mot d'origine (Exemple : « Skol » pour « School »). (Oueslati et al. 2018) ont considéré les erreurs d'orthographe comme une caractéristique durant leur analyse des sentiments.

1.8.4 Langage figuratif

Le langage figuratif peut aussi poser problème lors de l'analyse des sentiments (Hercig et Lenc 2017), or, il est difficile pour une machine de comprendre des choses que nous, les humains, peuvent distinguer facilement. Nous pouvons citer : les métaphores (« He is a shining star »), la métonymie (« La salle a applaudi l'orchestre ») et même l'ironie et le sarcasme (Maynard et Greenwood 2014) où les gens expriment leurs sentiments en utilisant le vocabulaire de polarité opposée, par exemple la phrase « L'expérience sur votre site web est MAGNIFIQUE ! Aucun bug en vue… » serait détectée comme une phrase positive par une machine, cependant un humain pourrait bien y soupçonner de l'ironie.

1.8.5 Crédibilité de l'opinion

Un autre défi majeur rencontré dans le domaine de la fouille d'opinions aujourd'hui est la détection de la crédibilité des opinions publiées et l'authenticité de leurs propriétaires (Jindal et B. Liu 2008 ; Oueslati et al. 2018). L'analyse de la crédibilité de l'opinion est cruciale lors de l'analyse sémantique d'un ensemble de documents, elle peut se montrer compliquée avec les différences de comportement entre les spammeurs et l'apparition des algorithmes de génération de spam qui peuvent être facilement paramétrables pour battre les méthodes de l'état de l'art. Ce problème est devenu, ces dernières années, un sujet d'intérêt intéressant les chercheurs académiques comme ceux de l'industrie, c'est ce qui nous a poussés à en faire l'objet du prochain chapitre de cette partie du mémoire.

1.9 Conclusion

Dans ce chapitre, nous nous sommes intéressées au domaine de l'analyse des sentiments. Nous avons d'abord défini quelques concepts de base, puis nous avons modélisé la fouille d'opinion sous forme de processus, pour ensuite la classifier selon deux critères. Par la suite, nous avons présenté les jeux de données disponibles, quelques approches adoptées dans ce domaine et les métriques et méthodes d'évaluation. Pour finir, nous avons conclu ce chapitre en discutant des différents défis rencontrés dans ce domaine, parmi eux, on trouve la détection de spam d'opinion qui sera le sujet du deuxième chapitre de ce rapport vu qu'elle reste un défi ouvert dans le domaine, et comprendre la crédibilité d'une opinion est essentiel pour utiliser efficacement l'analyse de sentiment dans des applications réelles vu qu'elle détermine la fiabilité des résultats.

Chapitre 2
La détection de Spam d'opinion

2.1 Introduction

Il est vrai que la tache d'analyse de sentiment est une tache prépondérante pour la prise de décision que ce soit pour l'amélioration de la relation client ou de l'e-réputation. Cependant, avec l'ouverture du web, disponibilité… énormément d'avis on a constaté qu'il y a pb de crédibilité. Les opinions non crédibles constituent alors un biais dans la tache de prise de décision. Il est par conséquent important de les distinguer pour les éliminer ou les traiter afin qu'elles ne nuisent.

Désormais, les internautes peuvent partager leurs expériences et leurs opinions autour des divers produits, services, marques ou personnes sur les différentes plateformes (les réseaux sociaux, les blogs…).

Cette abondance informative représente une source importante pour accéder aux retours d'expérience client, à leurs suggestions et leurs idées. Quand ces avis sont crédibles, ils constituent un support d'aide à la décision pour les clients, d'un côté, et d'un autre, ils permettent aux entreprises d'améliorer la qualité de leurs services et de mieux satisfaire leurs clients.

Ces avis, disponibles en masse et facilement accessibles, peuvent influencer les décisions d'achat des clients et même totalement changer leurs points de vue envers une cible (une marque, un produit, un service…) ; les entreprises sont, de nos jours, plus conscientes que jamais de cette influence et sur l'effet qu'elle peut avoir sur leur réputation, voire leurs profits. D'une part, une réputation en ligne généralement positive (i.e. les avis disponibles sont majoritairement positifs) est considérée comme un point de force par rapport aux compétiteurs pouvant impacter positivement les ventes. D'autre part, une quantité de critiques négatives (même inférieure au taux d'opinions positives) peut nuire, voire détruire, la réputation d'une cible menaçant son existence en diminuant ses profits.

Dans ce chapitre, nous définirons les concepts principaux du problème de détection de spam d'opinion dans la section 2.3. Nous présenterons ensuite le processus de détection de spam d'opinion dans la section 2.4, les sections qui suivent détaillent les étapes de ce processus, à savoir la collecte et les jeux de données 2.5, le prétraitement 2.6, la représentation 2.7, la classification 2.8 et enfin l'évaluation 2.9. Pour conclure, nous discuterons des défis rencontrés dans ce domaine et des orientations pour les recherches futures dans la section 2.10.

2.2 Motivation

Conscients de cet effet, certains marchands créent des fausses critiques dans le but de gagner du profit, soit en créant des avis positifs afin de promouvoir leurs produits et services ou des avis négatifs pour nuire aux images de leurs compétiteurs. Certains de ces marchands optent pour le recrutement de rédacteurs professionnels et les chargent de rédiger des critiques sur leurs produits et services, plusieurs plateformes sont devenues source d'offres de rédaction d'avis pareilles tel que les groupes Facebook, les sites de travail libre comme Fiverr[1],... la figure 2.1 illustre 05 offres de services de rédaction de critiques par des travailleurs indépendants sur la plateforme Fiverr, il est évident que ces rédacteurs n'ont aucune expérience reflétant la réalité des produits/services qu'ils critiquent et que leur seule motivation est le profit financier et non le partage d'informations utiles pour les clients.

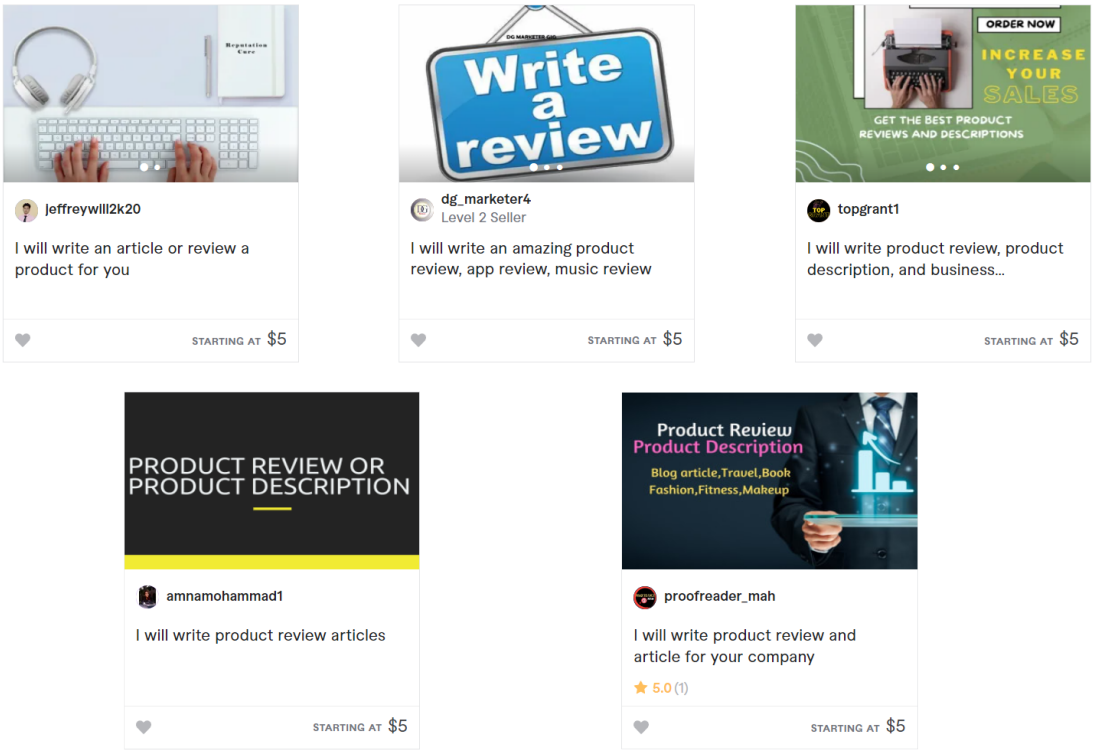


Fig. 2.1 : Offres de rédaction d'avis trompeurs sur le site Fiverr

D'autres vendeurs utilisent des algorithmes qui génèrent ces avis automatiquement, la figure 2.2 illustre un exemple : l'outil AISEO[2] qui génère des fausses critiques sur un produit, des algorithmes de la sorte sont facilement ajustables pour battre les techniques de détection de l'état de l'art. Ces opinions qui véhiculent de fausses informations sont appelées « spam d'opinion ».

[1] https://www.fiverr.com/
[2] https://aiseo.ai/templates/review-generator.html

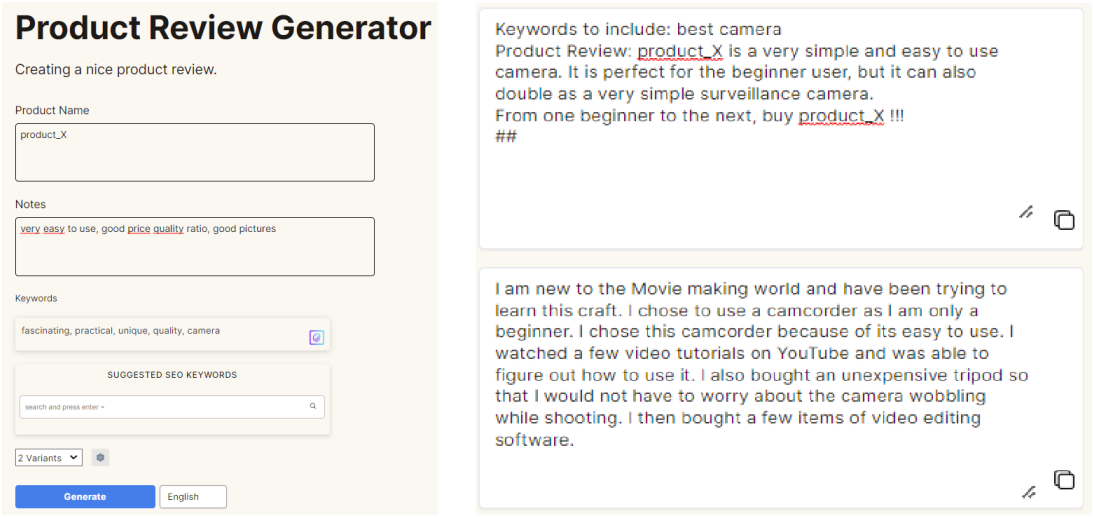


Fig. 2.2 : AISEO : Outil de génération d'avis

Le problème de spam d'opinion a été formalisé en 2008 par (Jindal et B. Liu 2008), il est généralement abordé comme un problème de classification. Mais avec les comportements et les styles différents, même imprévisibles, des rédacteurs, ainsi que la puissance des algorithmes de génération de spam d'opinion et la quantité limitée de données annotées, le problème de spam d'opinion s'est avéré difficile et complexe, ce qui a suscité beaucoup d'intérêt et est devenu un sérieux problème étudié tant au niveau des entreprises qu'au niveau académique. C'est ainsi qu'est né le besoin de disposer d'un outil offrant la possibilité de distinguer les opinions qui sont crédibles de ceux qui ne le sont pas.

Divers travaux sont proposés pour résoudre cette problématique ; devant ce nombre d'avis disponibles aujourd'hui, quelle est la meilleure approche afin de distinguer les avis crédibles des avis spam ?

2.3 Définitions

Dans ce qui suit, nous allons définir deux concepts essentiels à notre compréhension de ce domaine de recherche.

2.3.1 La crédibilité

Le terme « Crédibilité » remonte au milieu du 16 siècle, selon le dictionnaire Larousse[3], il est issu du mot latin médiéval « credibilitas » et signifie « Caractère de quelque chose qui peut être cru » ou « Caractère de quelqu'un qui est digne de confiance ». Une information, un propos ou un document peut être considérée crédible selon le niveau de fiabilité et de confiance que le récepteur accorde à sa source et à son canal de transmission.

Dans (Castillo et al. 2011), les auteurs ont employé la crédibilité dans le sens de la fiabilité, comme une qualité perçue composée de multiples dimensions offrant des motifs raisonnables d'être cru tout en n'étant évaluée qu'avec les informations disponibles. D'autre part, la crédibilité de l'information a été définie dans (R. Li et Suh 2015) comme étant la mesure dans laquelle on perçoit l'information comme étant crédible, et est un prédicteur puissant de l'action ultérieure d'un consommateur d'information, comme la

[3] https://www.larousse.fr/dictionnaires/francais/cr%C3%A9dibilit%C3%A9/20311

recommandation ou la volonté d'adopter le point de vue de l'information reçue.

Dans le domaine de fouille d'opinion, la crédibilité est considérée comme une qualité importante qu'une opinion doit satisfaire. Une opinion crédible est une opinion qui reflète une expérience réellement vécue par le client qui l'exprime, elle ne doit pas contenir de propos mensongers concernant l'entité qu'elle cible, que ce soit un produit, une marque, un service…

2.3.2 Le spam

En consultant les différents dictionnaires et encyclopédies, tels que Larousse[4], nous trouvons que la majorité de ces derniers définissent le terme 'spam' comme du courrier indésirable (Becchetti et al. 2008) : « Courrier électronique non sollicité envoyé en grand nombre à des boîtes aux lettres électroniques ou à des forums, dans un but publicitaire ou commercial ». Or ceci n'est qu'un seul type de spam et il en existe plusieurs autres.

Parmi les autres types de spam, nous pouvons trouver : Le spam des sites web, le spam social ainsi que le spam d'opinion. La figure 2.3 ci-dessous résume la taxonomie du spam dans la littérature comme nous allons l'aborder dans cette section (2.3.2).

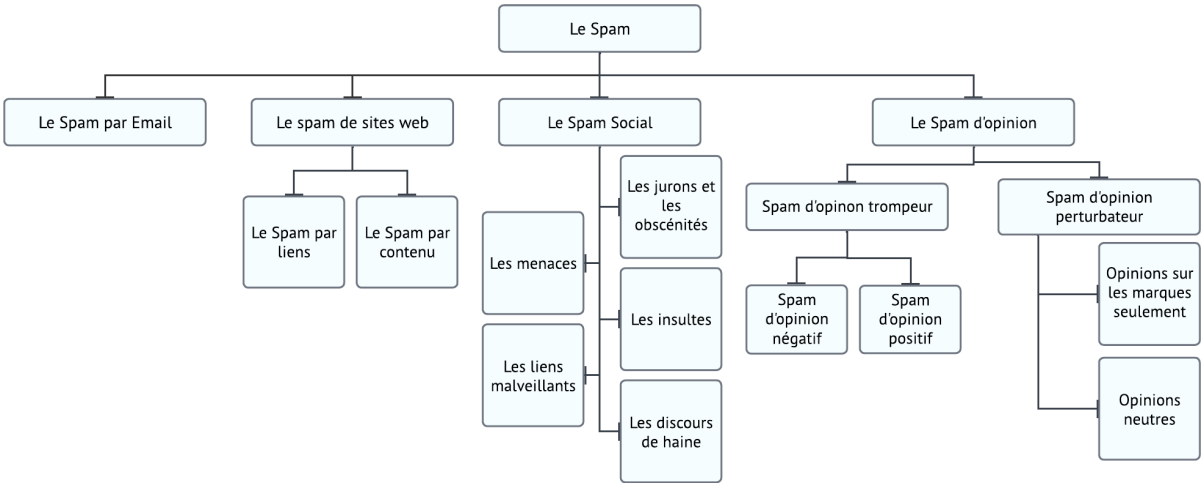


Fig. 2.3 : La taxonomie du spam dans la littérature

2.3.2.1 Le spam de sites web

Le spam de sites web est le deuxième type de spam le plus répandu après le spam par e-mail, il représente une tentative de manipulation du classement des moteurs de recherche. Selon l'article de (Becchetti et al. 2008) Il en existe deux types principaux, à savoir le spam de lien et le spam de contenu.

Tandis qu'ils visent tous deux à atteindre le sommet des résultats des moteurs de recherche sans tenir compte de toute valeur à offrir à l'utilisateur, **le spam de contenu** utilise des mots-clés populaires (mais non pertinents) ou des balises méta répétitives faisant ainsi semblant d'offrir une aide et des informations qui finissent par être à la fois superficielles et inutiles afin de tromper les moteurs de recherche en rendant ces sites pertinents pour de nombreuses requêtes de recherche, en revanche **le spam de lien** y parvient en créant des liens hypertextes vers de nombreux autres sites web ou en payant d'autres sites pour créer un lien vers le leur.

[4] https://www.larousse.fr/dictionnaires/francais/spam/10910104

Comme nous pouvons le constater, le spam de sites web vise à tromper les moteurs de recherche afin d'obtenir une meilleure visibilité pour le site concerné, ainsi sa cible est la machine et non l'utilisateur puisque ce dernier peut facilement repérer ce type de spam.

2.3.2.2 Le spam social

Le spam social est un contenu indésirable apparaissant sur les services de réseaux sociaux, et tout site web incluant du contenu généré par l'utilisateur. Selon (Washha et al. 2017), il peut se manifester sous une énorme variété de formes, parmi elles nous pouvons citer :

Les jurons et les obscénités contenus dans les commentaires soumis par les utilisateurs. Ils sont répandus malgré les efforts des algorithmes de filtrage et de censure visant à masquer ou bannir ces types de contenu. Et ceci, à cause des techniques de détournement orthographique qui rendent ces mots non détectables par ces algorithmes, mais toujours reconnaissables à l'œil humain. Parmi ces techniques, on trouve le "LeetSpeaking" et le "cloaking" où certaines lettres sont remplacées par des chiffres ("c0rpus" au lieu de "corpus"), des symboles ("l!ste" au lieu de "liste") ou en insérant de la ponctuation entre les lettres ("l.i.v.r.e" au lieu de "livre").

Les insultes soumises par l'utilisateur sont des commentaires offensants envers une ou plusieurs cibles. Les intimidateurs peuvent profiter de l'anonymat pour harceler leurs cibles sans en assumer la responsabilité.

Les discours de haine sont des propos ou écrits offensants exprimant des préjugés fondés sur l'origine ethnique, le sexe, la religion ou tout autre motif similaire. Ce type de propagande est souvent utilisé pour créer des conflits et inciter de la haine entre des groupes de personnes.

Les menaces sont des déclarations écrites avec l'intention d'infliger du mal ou des dommages à une personne ou un groupe de personnes. Elles peuvent être une réponse à un acte ou une déclaration, ou sous forme de chantage pour obtenir de l'argent, des informations ou d'autres avantages en échange de ne pas donner suite à la menace.

Les liens malveillants sont des hyperliens insérés par certains utilisateurs dans les commentaires dans le but de causer des dommages aux internautes ou à leurs appareils électroniques. Un simple clic sur ces liens peut entraîner le téléchargement de logiciels malveillants, la redirection vers des sites inappropriés ou la perte d'informations.

2.3.2.3 Le spam d'opinion

Le spam d'opinion est un type de spam subtil, bien conçu et délibérément rédigé pour paraître authentique (Ott et al. 2011).Il vise à influencer les décisions d'achat des consommateurs et à changer la réputation d'une entité cible pour le meilleur ou pour le pire. C'est à ce type de spam que nous nous intéressons dans notre recherche.

Le spam d'opinion est différent des types de spams mentionnés précédemment. Il se distingue du spam web du fait qu'il cible les clients et non pas les moteurs de recherche, et donc l'utilisation de liens hyperlinks peut être facilement détectée, et la répétition de mots clés non pertinents n'a aucun sens si on veut que le contenu du spam soit convaincant. D'autre part, il se démarque du spam social du fait qu'il vise à influencer l'opinion individuelle ou publique en faveur ou contre une entité.

En formalisant le problème de détection de spam, (Jindal et B. Liu 2008) ont identifié trois types de spam d'opinion :

- Les opinions mensongères ou trompeuses : ceux qui trompent délibérément les lecteurs et ne sont pas fondées sur de vraies expériences et visent soit à nuire à l'image d'une entité (spam négatif) ou à l'améliorer en donnant des critiques positives non méritées (spam positif).

- Les opinions qui portent sur les marques seulement : ceux qui commentent sur la marque et non pas sur l'entité à évaluer. Ils peuvent être considérés comme spam, car ils n'apportent pas d'informations de valeur aux clients.

- Les opinions neutres (non-avis) : ce ne sont pas des avis, et il en existe 02 types : les annonces et d'autres textes aléatoires non pertinents. Ils n'ont aucun impact sur la réputation du produit/service.

Le deuxième et le troisième type de spam d'opinion sont appelés **spam d'opinion perturbateur**, car ils ne représentent pas de menace pour les clients, mais en masse, ils perturbent leur expérience.

En revanche, le premier type de spam d'opinion cité, à savoir celui des opinions mensongères, est-ce qu'on appelle **le spam d'opinion trompeur**, car il est difficile à identifier manuellement, et de nos jours, il est devenu de plus en plus répandu sur les réseaux sociaux, de plus en plus sophistiqué et subtil, sa détection est devenue un défi majeur. En effet, (Ott et al. 2011) ont confirmé que ce type de spam est si subtile que le consommateur a tendance à confondre une opinion spam pour une opinion authentique. Montrons cette subtilité à travers un exemple, voici deux opinions extraites du premier jeu de données public de référence (gold standard dataset) construit dans le domaine de la détection de spam d'opinion[5] (Ott et al. 2011) :

Opinion 01 : « *Last week I stayed at the Hilton Chicago for 4 days and 3 nights and I was very pleased with the experience. As soon as I approached the front desk, I knew right away the staff was friendly and courteous. They had given me a list of local attractions such as the Museum of Science and Industry, the Broadcast Museum, the Ford Center For The Performing Arts and Willis Tower. I stayed in a junior suite which included a king size bed, a 27 inch television, non-allergenic feather pillows and a variety of other amenities. The indoor pool was rather large as well as the gym which includes countless treadmills and even a jogging track. Eating in the area is not a problem as the Hilton Chicago is home to Kitty O'Sheas which is a restaurant offering authentic Irish fare. In conclusion, I had a wonderful time staying here and I can not wait to plan my return trip to the Hilton Chicago.* ».

Opinion 02 : « *My family and I have just had a two week holiday in Chicago, and we stayed for a week at the Hilton Towers. We had a fantastic time and enjoyed every minute of our stay at this incredible hotel. The sheer size of the place is breathtaking and the atmosphere is very friendly and hospitable. If you prefer discreet modern boutiquey-type hotels, maybe this wouldn't be your thing. However if you enjoy old style glamour and glitz, you will be bowled over by this hotel. The place literally sparkles thanks to the enormous chandeliers that are everywhere in the lobby. Tons of celebs have stayed here*

[5] https://myleott.com/op-spam.html

over the decades and the history of the hotel and past guests etc is fascinating. We got a very reasonable rate for our stay and could not fault anything about this very special place ! ».

Maintenant, si nous devions essayer de déterminer avec juste le contenu lequel des avis présentés est le faux et lequel est le vrai, quelle serait notre réponse ? Très probablement, il serait difficile, voire impossible, d'identifier lequel des deux est un spam. Dans l'ensemble de données construit dans (Ott et al. 2011), 'Opinion 1' a été annotée comme une opinion trompeuse, tandis que 'Opinion 2' a été jugée authentique.

Les deux exemples présentés étaient des opinions positives (qu'elles soient trompeuses ou non), mais il est important de noter que des opinions trompeuses de différentes polarités ont un impact différent, en effet, un nombre plus élevé d'avis positifs incite un client à acheter un produit et renforce les gains financiers des fabricants, tandis que les avis négatifs incitent les consommateurs à rechercher des alternatives, entraînant des pertes financières. De plus, la relation entre la polarité de l'opinion trompeuse et la qualité réelle du produit ciblé a un impact sur la gravité de la situation. la figure 2.4 illustre une explication de cela telle que présentée dans l'étude (Jindal et B. Liu 2008).

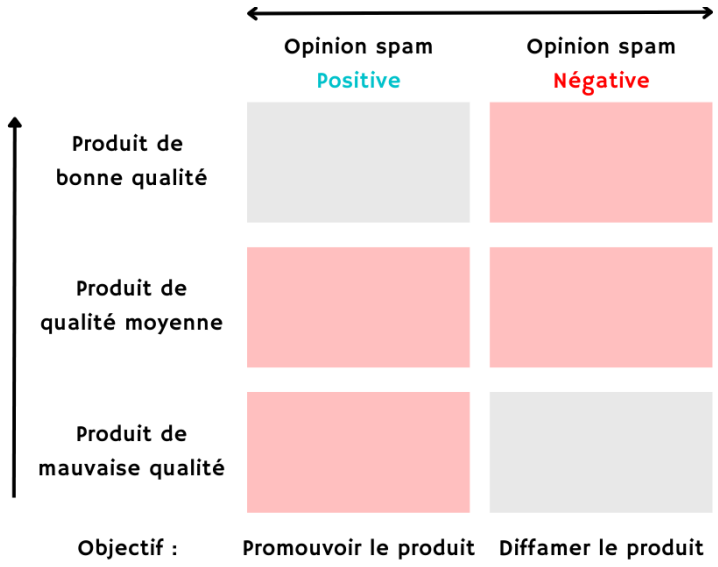


Fig. 2.4 : La relation entre la polarité des opinions spam et la qualité des produits selon (Jindal et B. Liu 2008)

Les opinions appartenant à la première colonne (i.e. Les opinions trompeuses positives) ont pour objectif de promouvoir l'entité cible (le produit), et sont généralement écrites par les fabricants du produit ou par des personnes ayant un intérêt quelconque dans le produit (économiques ou autre). Notons que même si les avis appartenant à la première case (les avis spam positifs pour un produit de bonne qualité) puissent être corrects, cependant ils ne sont pas basés sur une expérience réelle et donc ils sont considérés trompeurs malgré tout.

En revanche, les opinions de la deuxième colonne (i.e. Les opinions trompeuses négatives) sont très probablement écrites par des concurrents et même si la critique peut être vraie (cas de la dernière case ou l'opinion spam est négative et le produit est de mauvaise qualité) les intentions des rédacteurs restent malveillantes.

Il est évident que les opinions des cas particuliers cités auparavant (les cases colorées

en gris) ne sont pas si nocives, car même si elles ne sont pas fondées sur des expériences réelles, elles correspondent à la véritable qualité du produit ciblé. Ceci contrairement aux opinions spam correspondant aux cases en rouges qui peuvent avoir un impact néfaste sur non seulement les vendeurs, mais aussi les clients. C'est pour cela que (Jindal et B. Liu 2008) affirme la nécessité de s'accentuer sur la détection du spam d'opinions dans ces cas précis.

2.3.3 Le spammeur

Étant donné que le terme « spam » désigne couramment le spam par e-mail, il est évident que les spammeurs seraient souvent définis comme une personne ou une organisation qui envoie des messages non pertinents ou non sollicités par émail, généralement à un grand nombre d'utilisateurs, à des fins de publicité, d'hameçonnage, de diffusion de logiciels malveillants, etc.

Cependant, dans notre domaine de recherche concernant le spam d'opinion, nous définissons un spammeur comme une entité (personne ou organisation) qui publie des avis spam. D'après la définition de (Jindal et B. Liu 2008), un spammeur est un rédacteur qui a publié au moins un avis trompeur.

Selon (Andresini et al. 2022), l'étude du comportement d'un rédacteur d'avis pour la détection du spam provient de l'hypothèse qu'un spammeur écrit constamment du spam. Cela nous fournit une autre perspective pour détecter le spam d'avis : « Reconnaître qu'un rédacteur est un spammeur permet de marquer toute opinion dont il est l'auteur, qu'il ait déjà publié ou qu'il publiera à l'avenir, comme étant spam. »

Un spammeur peut être classifié, suivant le critère de professionnalisme ou de motivation, en deux catégories principales (B. Liu 2020) :

- Rédacteurs de spam professionnels : qui sont motivés par un gain financier. Ils sont généralement des indépendants (e.g. la figure 2.1 précédente qui présente des exemples d'offres de rédaction d'avis trompeurs réellement publiées sur Fiverr) ou des employés d'entreprises chargés d'écrire des fausses critiques dans le cadre de leurs activités. Les opinions spam professionnelles sont généralement plus faciles à détecter, puisque leurs auteurs en écrivent en grand nombre, ce qui peut laisser des modèles linguistiques et comportementaux facilement détectables par les algorithmes d'exploration de données. Cependant, le problème est qu'au moment où leur style d'écriture et leur comportement anormal sont reconnus et ces spammeurs sont finalement identifiés, leurs mensonges pourraient déjà être répandus et leur but pourraient être atteint. Donc, il est primordial de détecter ces modèles le plus tôt possible, mais cela reste difficile et pour empirer les choses, un spammeur identifié peut juste abandonner son compte et en créer un nouveau pour propager ses fausses critiques encore une fois.

- Rédacteurs de spam non professionnels : qui ne génèrent pas beaucoup de spam et qui ne sont généralement pas motivés par l'argent. Ces personnes écrivent principalement afin d'aider leurs connaissances ou leurs entreprises. Leur but est soit de promouvoir leurs produits ou ceux de leurs amis, de détruire l'image de leurs concurrents, ou même de nuire à leurs employeurs (anciens ou actuels) et entreprises. Il existe aussi des spammeurs qui rédigent de fausses critiques juste pour le plaisir

Comme ils n'écrivent pas beaucoup de critiques, ils peuvent ne pas avoir les mêmes habitudes que les professionnels, cependant, cela ne veut pas dire que leurs activités ne laisseront pas de trace. Par exemple, si un utilisateur consulte fréquemment une page d'avis sans en laisser un, puis suite à une critique négative, laisse une critique fortement positive, cet avis est considéré suspect car cette personne peut être le propriétaire du produit/service cible ou quelqu'un d'associé.

De plus, un autre groupe de personnes se situe à la frontière entre les vrais et les faux critiques, un autre groupe qui brise l'hypothèse faite ci-dessus, les Influenceurs et autres, ceux qui ont contribué à de nombreuses critiques authentiques et, ce faisant, ont construit leur réputation mais ont ensuite été contactés par des marques pour présenter leurs produits comme contenu sponsorisé et le critiquer positivement. Ce type de 'spammeurs' offrent un ensemble d'avis qui contient un mélange d'avis authentiques et trompeurs.

D'autre part, nous pouvons catégoriser les spammeurs selon deux types d'entité, les spammeurs individuels et les spammeurs de groupe (Mukherjee et al. 2011, 2012, 2013a ; B. Liu 2020). Ces deux types ont des caractéristiques différentes qui peuvent être exploitées afin de faciliter leur détection.

- Les spammeurs individuels : sont des personnes qui se chargent de rédiger des opinions spam eux même, sans travailler avec d'autres personnes, avec un seul compte (ou userId).

- Les groupes de spammeurs, ou de comptes spam, travaillent ensemble, consciemment ou pas, dans le but de promouvoir ou rabaisser certains produits ou services. Généralement professionnels et parfois non, les spammeurs de groupe suivent principalement un des deux modèles suivants :

1) Le premier modèle est celui d'un groupe de spammeurs (personnes) qui travaillent en collusion pour promouvoir une entité cible et/ou nuire à la réputation d'une autre. Professionnels ou pas, les membres de ce groupe pourraient ne pas connaître l'existence les uns des autres, ou pourraient ne pas connaître leurs activités. Par exemple, un auteur de livre peut demander à un groupe d'amis d'écrire des critiques positives pour l'un de ses nouveaux ouvrages. Ces amis peuvent ne pas connaître l'activité de l'autre, et ils ne sont normalement pas des spammeurs professionnels.

2) Le deuxième modèle est la pratique qu'on appelle "sock puppeting" : Une seule personne ou organisation enregistre plusieurs comptes (chacun avec un ID utilisateur différent) et commence à rédiger des spams en utilisant ces comptes. Ces multiples comptes se comportent exactement comme un groupe travaillant en collusion.

Hélas cette classification n'est pas binaire, il y a d'autres scénarios qui pourraient venir compliquer la détection des spammeurs. Par exemple, une personne pourrait parfois travailler individuellement et puis en groupes pour certaines cibles, il pourrait aussi arriver qu'elle achète un produit et le critique honnêtement. Avec toutes ces situations qui peuvent arriver, le problème de détection de spammeurs se montre complexe à résoudre.

2.4 Processus de la détection de spam d'opinion

Dans la littérature, la détection du spam d'opinion était autrefois divisée en deux taches, la première étant la classification d'un avis comme véridique ou spam, la seconde

étant l'identification des spammeurs. Ce n'est que ces dernières années que les chercheurs ont commencé à envisager le problème sous un autre angle, celui des produits ciblés .i.e découvrir quels produits sont ciblés par les spammeurs. De nombreuses études montrent l'intérêt de prendre en considération les trois points de vue (Rastogi et al. 2020). Cette information supplémentaire s'avérera utile lorsqu'il s'agira d'extraire des caractéristiques plus tard.

Nous décrivons dans cette section les étapes du processus de la fouille de spam d'opinion. Enfaîte, ce processus suit les mêmes étapes que le processus générique de fouille d'opinions. Les étapes de ce processus de détection de fraude sont comme suit :

2.4.1 Collecte des données

Cette première étape consiste à se procurer des jeux de données nécessaires pour l'entraînement et l'évaluation des modèles de classification. Selon les caractéristiques et les marqueurs que l'on souhaite exploiter, il existe différents types de données que les jeux de données choisis doivent fournir. Ces ensembles de données pourraient être des datasets disponibles au public pour des fins académiques, ou des données que les chercheurs collectent eux-mêmes à travers les réseaux sociaux et les différents sites permettant aux internautes de partager leurs avis envers des entités cibles (Évènements, produits, sujets,...). Nous en discuterons avec plus de détails dans la section 2.5.

2.4.2 Préparation et prétraitement des données

Après la collecte, les données passent par un processus qui garantit qu'elles sont suffisamment nettoyées, structurées et optimisées pour la suite. Dans cette étape, les données brutes sont nettoyées du bruit, les données manquantes sont remplacées, les données incorrectes sont supprimées ou remplacées,... Ensuite, afin de réduire le problème de dimensionnalité, ou comme on l'appelle souvent la malédiction de la dimensionnalité (Andresini et al. 2022), il est souvent conseillé d'effacer les variables fortement corrélées, d'ignorer les individus hors population, etc.

De plus, dans cette étape, toutes les données textuelles (ici les critiques) doivent être prétraitées, généralement à l'aide de techniques NLP, telles que la tokenisation, la lemmatisation, la racinisation, la normalisation de la casse, etc. sans oublier la suppression ou le remplacement de tout ce qui est jugé non pertinent, comme les URLs, les mentions, les hashtags, émojis, etc. Notant qu'un nettoyage excessif d'un texte brut peut nous faire éliminer les modèles pertinents qu'il contient, il peut donc valoir la peine d'expérimenter différents paramètres, également puisque l'analyse des sentiments passe également par une telle étape, et puisque la polarité pourrait être une caractéristique utile à exploiter lors de la détection de spam, il est important de noter que des éléments tels que les hashtags et les émojis peuvent avoir un impact sur la polarité de l'examen, il peut donc s'avérer utile d'étudier leur effet sur les performances d'un modèle.

2.4.3 Extraction et représentation des caractéristiques

Cette étape est la dernière étape avant celle de la classification, son rôle est donc d'optimiser les données prétraitées et de mettre en valeur les indicateurs pertinents pour fournir un bon ensemble de données pour le modèle de classification. En effet, les performances d'un modèle dépendent fortement de la qualité des données d'entrée. Pour cela, cette étape

sera consacrée à l'extraction de caractéristiques pertinentes depuis les données prétraitées puis la présentation de ce jeu de données sous forme de vecteurs de dimension réduite et uniforme que le modèle de classification pourra traiter. Nous distinguons deux types de familles d'approches dans cette étape : l'ingénierie des caractéristiques et l'apprentissage des représentations, elles seront présentées dans la section 2.7 de ce chapitre.

2.4.4 Classification

L'ensemble de données étant enfin complet, il est maintenant temps de le diviser en un ensemble d'apprentissage et un ensemble de test. L'ensemble d'apprentissage sera l'entrée du modèle de classification, ce sont les données sur lesquelles le modèle s'entraînera, sa taille est généralement supérieure à celle de l'ensemble de test (généralement 80% de l'ensemble de données). L'ensemble de test, quant à lui, sera utilisé pour tester les performances du modèle entraîné, c'est-à-dire s'il a suffisamment bien appris pour pouvoir classer des données autres que celles sur lesquelles il a été entraîné. Habituellement, cette division se fait de manière aléatoire, mais certaines études ont montré d'autres méthodes. Par exemple, les auteurs de (Andresini et al. 2022) ont divisé leurs données en s'assurant que les données présentes dans l'ensemble d'entraînement sont plus anciennes que celles de l'ensemble de test, afin d'imiter une situation réelle, comme vous le savez, un modèle entraîné devra prédire la crédibilité d'un avis sans avoir aucune information sur ceux qui seront postés après.

Par la suite, le modèle sera entraîné sur l'ensemble d'entraînement. La détection de spam est considérée comme un problème de classification binaire, elle consiste à associer une opinion en entrée à une classe (Spam/Non spam) en fonction de sa crédibilité. Il existe de multiples approches adoptées pour résoudre ce problème de classification, et divers critères selon lesquels ces méthodes sont catégorisées, nous en reparlerons dans la section 2.8.

2.4.5 Évaluation

Pour finir, afin de s'assurer des bonnes performances du modèle implémenté, ce dernier est testé à l'aide de l'ensemble de test construit précédemment, et évalué avec certaines métriques qui seront le sujet de la section 2.9.

Dans ce qui suit, nous discuterons de ces étapes fondamentales du processus de détection de spam avec plus de détails en présentant certains travaux de l'état de l'art les concernant.

2.5 Jeux de données

Les ensembles de données peuvent être catégorisés selon les différentes méthodes de construction. Les chercheurs et entreprises peuvent construire leur propre datasets selon ces méthodes ou ils peuvent exploiter les jeux de données disponibles publiquement. Dans le tableau 2.1, nous présentons certains ensembles de données de référence (Gold-standard) utilisés dans les travaux de la littérature pour la détection de spam d'opinion, ces datasets sont catégorisées selon les méthodes de construction existantes dans la littérature, à savoir : les méthodes manuelles, les méthodes basées sur des règles, les méthodes basées sur des algorithmes de filtrage et les méthodes basées sur le crowdsourcing (Ren et Ji 2019). Ces corpus ne sont pas vraiment récents, mais ils sont des références sur lesquelles

de nombreux chercheurs ont basé leurs travaux.

Tab. 2.1 : Ensembles de données de référence (Gold-standard) utilisés dans les travaux de la littérature pour la détection de spam d'opinion.

Méthode de construction	Source	Volume	Domaine	Référence
Basée sur des règles	Amazon[6]	5,8M d'avis, 2,14M d'utilisateurs	Livres, Musique, DVD, Produits industriels	(Jindal et B. Liu 2008)
Manuelle	Epinions	6000 avis	Produits	(F. H. Li et al. 2011)
	TripAdvisor[7]	6000 avis	Hôtels	(Ren et al. 2014)
Basée sur des algorithmes de filtrage	Yelp	67395 avis, 38063 utilisateurs	Hôtels, Restaurants	YelpCHI (Mukherjee et al. 2013b)
	Yelp	9 765 avis	Hôtels	(H. Li et al. 2014)
	Yelp	359 052 avis, 160225 utilisateurs	Hôtels	YelpNYC (Rayana et Akoglu 2015)
	Dianping	608 598 avis, 260277 utilisateurs	Hôtels	YelpZip (Rayana et Akoglu 2015)
Basée sur le crowdsourcing	TripAdvisor	800 avis	Hôtels	(Ott et al. 2011)
	TripAdvisor	1600 avis	Hôtels	(Ott et al. 2013)
	TripAdvisor	3 032 avis	Hôtels, Médecins, Restaurants	(Jiwei Li et al. 2014)

2.6 Approches de prétraitement

En raison de la grande taille des données collectées, en addition à leurs origines multiples et hétérogènes dans la plupart des cas, ces datasets sont très susceptibles au bruit, aux données manquantes et incohérentes. La qualité des données affecte les performances des classificateurs, pour cela une phase de prétraitement est nécessaire pour assurer une exploration de données efficace. Un processus de prétraitement peut inclure plusieurs étapes :

D'abord, il faut **évaluer la qualité des données collectées**, en les analysant dans le but de trouver des données manquantes, incohérentes ou incompatibles. Une fois ces problèmes détectés, nous devons **nettoyer les données** soit en ignorant les tuples défectueux ou en remplissant manuellement les données erronées avec des valeurs nulles ou par défaut. Nous devons également mettre toutes les données, telles que les dates, sous

[6] Amazon.com

[7] https://www.tripadvisor.com/

un format uniforme, notamment pour les données collectées à partir de sources multiples. Une autre étape couramment utilisée est la suppression des doublons, cependant, cela ne nous convient pas puisque les commentaires en double peuvent être un indicateur de spam d'opinion qui mérite d'être vérifié.

En ce qui concerne les données textuelles, ici généralement le corps de l'opinion, nous pouvons nettoyer ce que nous jugeons inutile, par exemple les URLs, les mentions, les balises HTML,... ne sont pas pertinents pour notre analyse. Cependant, les émojis, les hashtags et les symboles peuvent contenir des indicateurs utiles pour la détection de spam ou pour l'analyse des sentiments, si nous prenons la polarité des opinions en considération. On peut ajouter à cela la suppression de la ponctuation, l'élimination des mots vides,... Cela est ensuite suivi par la transformation des données à travers plusieurs méthodes telles que la tokénisation, la racinisation, la lemmatisation,...

Comme l'a souligné (Sun et al. 2017), le prétraitement des données a atteint des performances prometteuses pour les documents canoniques (actualités, ...) cependant, il est insatisfaisant lorsqu'il est appliqué à des critiques qui ne sont généralement pas grammaticales. Le problème ne fait qu'empirer lors de l'application de tâches telles que la tokenisation sur des langues plus flexibles comme le chinois.

2.7 Approches de représentation

Comme nous l'avons déjà mentionné, la qualité et la représentation des données alimentant les modèles de classification ont un impact important sur les performances de ces derniers. La phase de représentation concerne la définition des caractéristiques à exploiter durant la classification. Dans la littérature, il arrive que certaines étapes de prétraitement et de représentation se chevauchent, ou qu'elles se confondent complètement, il est donc souvent difficile de trouver une ligne claire qui les sépare. Il existe deux familles d'approches utilisées dans cette étape : l'ingénierie des caractéristiques et l'apprentissage des représentations. Nous allons présenter dans cette section la taxonomie de ces approches comme l'illustre la figure 2.5. Cependant, nous devons garder à l'esprit que certaines catégories présentées pourraient se chevaucher, et que le domaine de l'apprentissage automatique est en constante évolution, de nouvelles méthodes et techniques pouvant brouiller les frontières entre ces catégories.

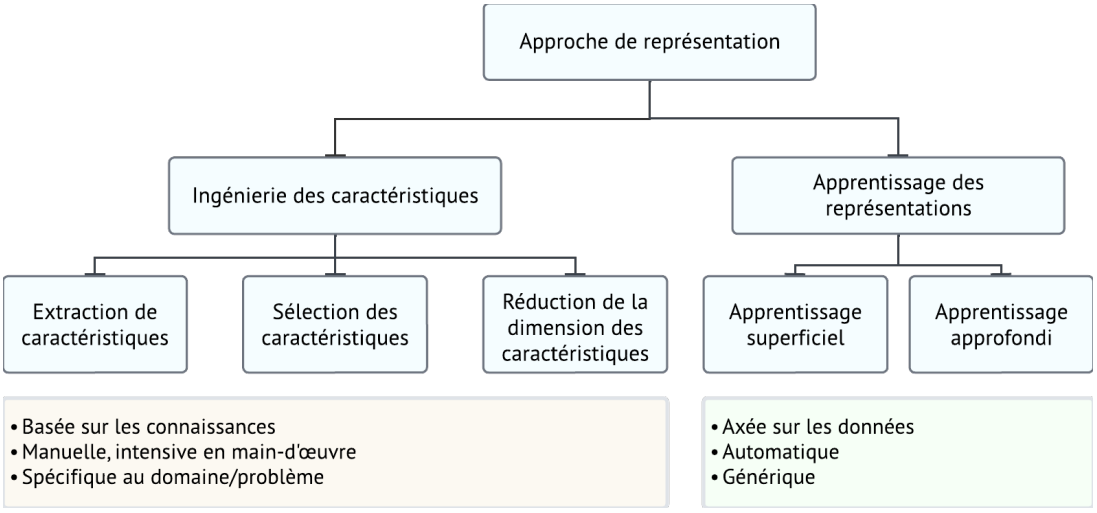


Fig. 2.5 : Taxonomie des approches de représentation

2.7.1 Ingénierie des caractéristiques

En parcourant la littérature, il ne semble pas y avoir beaucoup d'attention concentrée sur l'ingénierie des caractéristiques (Brownlee 2014 ; Yan et Yu 2019), nous trouvons différentes définitions et classifications, mais tout se résume à ceci : l'ingénierie des caractéristiques est le processus de conversion manuelle des données brutes en informations significatives (caractéristiques) en utilisant des connaissances spécifiques au domaine. Ce processus vise à aboutir à un nombre raisonnablement limité de caractéristiques utiles à faible dimensionnalité afin d'améliorer les performances du classifieur par rapport à ne lui fournir que des données brutes. De cette définition, nous pouvons déduire trois sous-classes de cette catégorie : la sélection de caractéristiques, la réduction de la dimensionnalité des caractéristiques et l'extraction de caractéristiques.

2.7.1.1 Sélection des caractéristiques

La sélection des caractéristiques est le processus d'identification des caractéristiques les plus cohérentes, non redondantes et pertinentes à utiliser dans la construction de modèles. La réduction méthodique de la taille des ensembles de données est importante, car la taille et la variété des ensembles de données continuent de croître. Souvent, il est considéré comme faisant partie de la phase de prétraitement, car il n'introduit aucune nouvelle caractéristique, mais sélectionne plutôt un sous-ensemble des caractéristiques brutes qui sont plus interprétables. Son but est de trouver les caractéristiques qui ont un impact déterminant sur les résultats du modèle prédictif afin de réduire le nombre de caractéristiques et les risques de surajustement sans sacrifier le pouvoir prédictif Ceci revient à supprimer les données superflues et pour permettre au modèle de se concentrer uniquement sur les caractéristiques importantes. Dans la littérature, la sélection de caractéristiques est classifiée en deux catégories, celle des méthodes supervisées et non supervisées. Pour les approches non supervisées, on détermine la corrélation entre les caractéristiques sans prendre en considération la variable cible (ici l'étiquette de crédibilité). En revanche, les techniques supervisées examinent la relation entre chacune des caractéristiques et l'étiquette cible et incluent trois sous-catégories : les Méthodes d'enveloppement, les méthodes de filtrage et les méthodes embarquées (Zheng et Y. Zhang 2008).

a) Les Méthodes d'enveloppement, en anglais Wrapper, utilisent la recherche itérative. Elles calculent des modèles avec un sous-ensemble de caractéristiques de départ et évaluent l'importance de chaque caractéristique. Ensuite, elles itèrent et essaient un sous-ensemble différent de caractéristiques jusqu'à ce que le sous-ensemble optimal soit atteint. Les méthodes d'enveloppement les plus populaires incluent la sélection directe (Forward Selection), la sélection inverse (Backward Selection) et la sélection pas à pas (Stepwise selection). La **"sélection directe"**, ou l'élimination directe, est une méthode gourmande itérative qui ajoute au modèle une caractéristique à la fois et l'évalue, elle continue d'itérer jusqu'à ce qu'aucune amélioration ne soit constatée. La **"sélection inverse"**, d'autre part, est l'inverse de la sélection directe, elle commence par toutes les caractéristiques puis supprime une caractéristique à la fois et continue jusqu'à ce que toutes les caractéristiques avec des p-valeurs insignifiantes soient supprimées du modèle. Et enfin, la **"sélection pas à pas"** qui est un hybride des deux autres et commence avec 0 caractéristiques, puis ajoute de manière itérative la caractéristique ayant la p-valeur significative la plus faible, tout en vérifiant à nouveau celles ajoutées précédemment et

en supprimant celles qui sont devenues insignifiantes, de cette façon toutes les caractéristiques incluses dans le modèle final seront significatives. Les inconvénients majeurs de ces méthodes sont le temps de calcul important pour les données avec de nombreuses caractéristiques, le fait que la solution peut ne pas être la méthode optimale vu la susceptibilité de faux départs puisqu'une recherche exhaustive est impossible, et le fait qu'elle a tendance à surajuster le modèle lorsqu'il n'y a pas une grande quantité de points de données. (Zheng et Y. Zhang 2008)

b) Les méthodes de filtrage, quant à elles, sont les plus simples et les plus utilisées dans la littérature. Elles examinent les caractéristiques corrélées (interactions caractéristiques-cible) et sélectionnent le meilleur sous-ensemble à donner au classificateur. Elles commencent par toutes les caractéristiques puis sélectionnent le meilleur sous-ensemble en les classant à l'aide d'une mesure descriptive utile (autre que le taux d'erreur). Parmi les plus populaires, on trouve l'ANOVA, la corrélation de Pearson et le facteur d'inflation de la variance. Ces méthodes sont rapides et faciles à interpréter, cependant elles sont moins précises et le risque d'inclure des caractéristiques redondantes est présent puisqu'elles sont aveugles aux interactions entre les caractéristiques. (Zheng et Y. Zhang 2008)

Et enfin **c) les méthodes embarquées** qui tentent de sélectionner simultanément un sous-ensemble de caractéristiques tout en ajustant le modèle, elles sont intégrées dans le cadre de l'algorithme d'apprentissage. Elles optimisent souvent une fonction objective qui récompense conjointement l'exactitude de la classification et pénalise l'utilisation de plus de caractéristiques. Dans cette méthode, nous trouvons des approches de régularisation (qui incluent Lasso, Ridge et Elastic Nets) et celles basées sur des algorithmes (qui peuvent être utilisées en utilisant n'importe quel type d'algorithme basé sur des arbres tels que des arbres de décision, RandomForest, etc.). Ces méthodes combinent les qualités de celles présentées précédemment, elles sont rapides comme les méthodes de filtrage, mais plus précises qu'elles, elles prennent en considération l'interaction des caractéristiques comme le font les méthodes Wrapper tout en étant moins sujettes au sur-ajustement. (Zheng et Y. Zhang 2008)

2.7.1.2 Réduction de la dimension des caractéristiques

Parfois, nous ne pouvons pas obtenir une bonne dimensionnalité juste en sélectionnant les caractéristiques. Pour combler cette lacune, nous utilisons des techniques de réduction de dimensionnalité (DR). Bien que les deux soient utilisés pour réduire le nombre d'entités dans un jeu de données, il existe une différence importante : la sélection de caractéristiques consiste simplement à sélectionner et à exclure des caractéristiques données sans les modifier, par contre, la réduction de dimensionnalité transforme les entités en une dimension inférieure. Le but des techniques DR est de réduire les caractéristiques sans perdre beaucoup d'informations authentiques et d'améliorer les performances. Il en existe deux principaux types : les **DR linéaires**, qui incluent l'analyse en composantes principales (ACP)(Pearson 1901), l'analyse en composantes indépendantes (ICA)(Comon 1994), l'analyse discriminante linéaire (LDA)(MI et al. 2003), la décomposition en valeurs singulières (SVD)(Golub et Van Loan 1996), l'indexation sémantique latente (LSI)(Deerwester et al. 1990), etc, ainsi que les **DR non linéaires** dont l'analyse en composantes principales du noyau (KPCA)(Schölkopf et al. 1998), l'incorporation de

voisins stochastiques distribués en T, l'analyse discriminante linéaire du noyau (KLDA), la mise à l'échelle multidimensionnelle, cartographie isométrique (Isomap)(Tenenbaum et al. 2000), etc. Le tableau 2.2 résume les méthodes de réduction de dimensionnalité les plus utilisées dans le domaine de détection de spam.

Tab. 2.2 : Les méthodes de réduction de dimensionnalité les plus utilisées dans le domaine de détection de spam

Méthodes DR linéaires		
Nom	Caractéristiques	Travaux
Analyse en composantes principales (PCA)	- Technique statistique non supervisée. - Transforme un ensemble de variables corrélées en un ensemble plus petit de variables non corrélées appelées « composantes principales ». - Construit une représentation à faible dimension des données multivariées en préservant la variation. - Utilise des données denses.	(Alhaj et al. 2022), (Sedighi et al. 2017), (Sakurada et Yairi 2014)
Analyse Discriminante Linéaire (LDA)	- Méthode supervisée. - Trouve une combinaison linéaire de caractéristiques d'entrée qui optimise la séparabilité des classes. - Les données doivent être distribuées normalement.	(Goel et al. 2021)
Décomposition en valeurs singulières (SVD)	- Connue dans le domaine de l'étude de la RI pour réduire la dimensionnalité dans des applications telles que la classification de texte, où les documents sont présentés avec des vecteurs. - Fonctionne bien avec des données rares.	(Alhaj et al. 2022)
Indexation sémantique latente (LSI)	- C'est l'analyse de la sémantique latente c'est-à-dire cachée dans un corpus de texte. - Elle transforme les données brutes dans un espace différent afin que deux documents/mots sur le même concept soient mappés à proximité (afin qu'ils aient une similitude de cosinus plus élevée). - Elle utilise SVD de la matrice terme-document.	(T. Wang et Zhu 2014)
Méthodes DR non linéaires		
Nom	Caractéristiques	Travaux
Analyse en Composantes Principales du noyau (KPCA)	- Efficace pour les datasets linéairement séparables. - Effectue un mappage non linéaire sur l'espace des caractéristiques de grande dimension par une fonction noyau, puis emploie une PCA linéaire dans l'espace des caractéristiques	(Harris 2022), (Sakurada et Yairi 2014)

Auto-encodeur	- C'est une forme de RN non supervisée qui est formée pour compresser et encoder des données puis de les reconstruire à partir de la représentation codée réduite avec une erreur minimale. - Elle ne peut encoder que des données similaires à celles sur lesquelles il a été formé. - Les données reconstruites sont une version dégradée de la version originale. - Elle est auto-supervisé, en utilisant l'entrée elle-même comme cible. - L'une de ses principales applications est la réduction de dimensionnalité à des fins de représentation	(Shaalan et al. 2021), (Sakurada et Yairi 2014)

2.7.1.3 Extraction de caractéristiques

L'extraction de caractéristiques est le processus de création de nouveaux indicateurs qui pourraient ne pas être vus dans les données brutes, et cela, en utilisant des connaissances d'experts. Contrairement à la sélection de caractéristiques qui conserve un sous-ensemble des caractéristiques d'origine, les algorithmes d'extraction de caractéristiques transforment les données en un nouvel espace de caractéristiques. Parfois, il arrive que l'extraction de caractéristiques et l'ingénierie de caractéristiques soient utilisées de manière interchangeable alors que les deux sous-catégories précédentes sont considérées comme faisant partie de la phase de prétraitement.

En matière de détection de spam d'opinion, la littérature classe les caractéristiques principalement en deux grandes catégories : les caractéristiques textuelles qui concernent le contenu de l'opinion, et les caractéristiques comportementales qui concernent le détenteur de l'opinion. Souvent des caractéristiques concernant le produit ciblé et sa popularité sont ajoutées à la liste.

- **Les caractéristiques liées au texte de l'avis :**

Les caractéristiques basées sur le contenu ont été les plus utilisées dans les méthodes de l'état de l'art, ces dernières sont principalement basées sur l'hypothèse qu'un spammeur a peu ou pas de connaissances sur le produit qu'il examine et, à ce titre, les avis authentiques et spam devraient présenter différents styles d'écriture, un indice pour les distinguer. Ces caractéristiques incluent des caractéristiques linguistiques telles que le lexique, la grammaire et les caractéristiques sémantiques du texte de l'opinion ainsi que des métadonnées. (Mohawesh et al. 2021b ; Ren et Ji 2019). Nous présentons maintenant certaines caractéristiques textuelles utilisées dans les travaux de recherche, nous posons d'abord les deux documents D_1 et D_2, extraits de TripAdvisor, comme exemples sur lesquels nous expliquerons certains algorithmes présentés :

- $\mathbf{D_1}$ **:** *« Had a meal here, food was cold and tasted bad. The smell was weird and nauseating. Our waitress was friendly but the service was very slow. »*

- $\mathbf{D_2}$ **:** *« We had an excellent meal, food was freshly cooked and delicious. Our waitress was very friendly. It was enjoyable. »*

1. **Les sacs de mots**, Bag of Words (BOW) en anglais, sont l'une des techniques de vectorisation les plus utilisées. L'algorithme BOW résulte en une représentation du texte

qui décrit l'occurrence des mots dans un document. La figure A.1, que vous trouverez dans l'annexe, présente le résultat de l'application de BOW sur les deux documents D_1 et D_2. Cet algorithme est simple et intuitif, et il ne rencontre pas d'erreurs "En dehors du vocabulaire". Cependant, il ne prend pas en compte les nouveaux mots ni l'ordre de la phrase. Parmi les travaux ayant exploité la technique BOW, on peut citer : (Hajek et al. 2020 ; Kennedy et al. 2020 ; Patel et Thakkar 2014).

2. **Les n-grammes** sont une chaîne connectée de N éléments apparaissant consécutivement dans un document. Cette méthode est basée sur l'idée que ce type de représentation pourrait capturer la structure d'une langue du point de vue statistique tout en maintenant l'ordre des mots, contrairement à BOW. La figure A.2 de l'annexe est un exemple simple de n-gramme appliqué sur une phrase extraite de D_1, elle montre le résultat de l'utilisation d'un unigramme (lorsque N=1), d'un bigramme (N= 2) et d'un trigramme (N=3). Même si cet algorithme est également simple et capable de capturer le sens sémantique de la phrase, il ignore toujours les nouveaux mots et à mesure que N augmente, la dimension de la formation de vecteurs augmente et le ralentit. Cet algorithme a été largement utilisé dans de nombreuses recherches (Ahmed 2017 ; Fusilier et al. 2015 ; Goel et al. 2021 ; Hajek et al. 2020 ; Jiwei Li et al. 2014 ; Mukherjee et al. 2012 ; Ott et al. 2013, 2011 ; Sedighi et al. 2017 ; T. Wang et Zhu 2014).

3. **POS,** Part-of-speech tagging, et également appelé étiquetage grammatical ou désambiguïsation des catégories de mots, est le processus consistant à attacher chaque mot dans le texte (le corps de l'opinion) avec une étiquette grammaticale, appelée étiquette POS, en fonction de son emplacement. La figure A.3 de l'annexe représente un exemple simple où nous avons appliqué la technique POS sur une D_1 et nous avons représenté les étiquettes par différentes couleurs. Parmi les études ayant exploité POS on peut citer : (Alsubari et al. 2020 ; Goel et al. 2021 ; Kennedy et al. 2020 ; Jiwei Li et al. 2014 ; Mukherjee et al. 2012 ; Ott et al. 2011 ; Sedighi et al. 2017 ; T. Wang et Zhu 2014). Selon (Alsubari et al. 2020), il semblerait que les avis véridiques contiennent plus de noms et d'adjectifs alors que les avis frauduleux ont plus de verbes et d'adverbes.

4. **LIWC**, ou en anglais Linguistic Inquiry and Word Count qui veut dire Recherche linguistique et comptage de mots, est un outil d'analyse de texte populaire[8] qui peut être utilisé pour analyser, extraire et calculer des caractéristiques linguistiques significatives sous différents aspects. Il s'appuie sur plus de 100 dictionnaires intégrés pour classer les mots cibles en fonction de différentes catégories. Sa dernière version est LIWC-22, c'est un logiciel assez puissant mais il est cher. (Alsubari et al. 2020 ; Jiwei Li et al. 2014 ; Mukherjee et al. 2013b ; Ott et al. 2011) ont utilisé cet outil dans leur recherche.

5. **Tf-Idf**, de l'anglais Term Frequency-Inverse Document Frequency, est une métrique de pondération souvent utilisée dans la recherche d'informations et le traitement du langage naturel, c'est une version modifiée de l'approche Term-Frequency (TF). TF est une technique qui utilise le nombre de mots apparaissant dans des documents pour déterminer la similitude entre eux. Il représente chaque document **d** par un vecteur (de même longueur) présentant chaque terme **t**, appartenant à l'ensemble des documents, avec sa probabilité d'exister dans d. La valeur TF concernant un terme t dans un document d est calculée selon l'équation 2.1. Comme TF pondère l'importance d'un terme et que

[8] https://www.liwc.app/

cela augmente avec le nombre de fois qu'il apparaît, cela pourrait faire en sorte que des termes généralement plus courants que d'autres tels que les mots vides ("The", "Then",...) dominent le nombre de fréquences sans accorder assez d'importance aux mots plus discriminants, c'est là que TF-IDF est une amélioration : l'utilisation d'IDF (calculé à l'aide de l'équation 2.2, où **N** est le nombre total de documents et $|d\epsilon D : t\epsilon d|$ est le nombre de documents contenant le terme t) réduit l'impact de tels termes. TF-IDF est calculé selon l'équation 2.3 ci-dessous. La figure A.4 de l'annexe illustre un exemple d'application de TF et de TF-IDF sur D_1 et D_2. TF-IDF a été utilisé dans (Ahmed 2017 ; Hassan et Islam 2021 ; H. Li et al. 2014 ; Patel et Thakkar 2014 ; Sedighi et al. 2017 ; T. Wang et Zhu 2014).

$$\mathrm{TF}(t,d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \tag{2.1}$$

$$\mathrm{IDF}(t,D) = \log \frac{N}{|d\epsilon D : t\epsilon d|} \tag{2.2}$$

$$TF-IDF(t,d) = tf \cdot idf_{t,d} = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \cdot \log \frac{N}{|d\epsilon D : t\epsilon d|} \tag{2.3}$$

6. **La stylométrie** est le domaine de la linguistique qui exploite des méthodes statistiques afin de définir les propriétés du style d'un texte. Les caractéristiques stylométriques peuvent être divisées en caractéristiques lexicales (telles que la longueur moyenne d'une phrase, le nombre total de termes,...) et syntaxiques (qui incluent la fréquence de ponctuation, etc.) (Ahmed 2017). Les auteurs de (Shojaee et al. 2013), qui ont extrait 234 caractéristiques stylométriques du dataset de (Ott et al. 2011), ont obtenu un F-score de 84% en utilisant à la fois des caractéristiques lexicales et syntaxiques surpassant les résultats qu'ils ont obtenus sans les combiner. Cette méthode a été utilisée dans de nombreux travaux tels que (Harris 2022 ; Jindal et B. Liu 2008 ; Kennedy et al. 2020 ; Rastogi et Mehrotra 2018 ; Rastogi et al. 2020 ; Shan et al. 2021 ; Shojaee et al. 2013). Parmi les caractéristiques utilisées on trouve la longueur du corps de l'opinion (RL), la diversité lexicale (le pourcentage de termes uniques), le ratio de mots négatifs (RNW) et positifs (RPW), le ratio de mots imaginatifs (RImW) et informatifs (RInW),le nombre de mots spatio-temporels, etc.

7. **Les caractéristiques sémantiques** traitent de la signification cachée des mots et de la relation entre eux. Cela aide à identifier les similitudes entre les avis, car même si le spammeur modifie certains mots dans l'un des avis afin d'induire les lecteurs en erreur, les caractéristiques sémantiques arrivent à détecter les avis générés à partir d'autres. Ainsi, par exemple, si l'auteur remplace le mot "Delicious" par "Tasty" ou "Horrible" par "Awful", un modèle basé sur des caractéristiques sémantiques pourrait toujours être en mesure de reconnaître la similarité entre eux. Les expériences de (L. Li et al. 2015) et de (Kim et al. 2015) montrent que l'utilisation de caractéristiques sémantiques a effectivement amélioré les performances de leurs modèles, dans (L. Li et al. 2015) ces caractéristiques ont surpassé LIWC, POS et les n-grammes dans les textes inter-domaines. Dans (Ziani et al. 2021), les auteurs ont travaillé sur les caractéristiques sémantiques de la langue arabe en exploitant les relations de cohérence entre les syntagmes d'une phrase, car deux syntagmes peuvent soit se contredire, soit se renforcer. Parmi les multiples connecteurs présents dans la langue arabe, ils ont en choisi 6 : les connecteurs d'explication, de cause, de condition, de ressemblance, d'opposition et de différence.

Comme nous l'avons montré dans cette partie, les caractéristiques textuelles sont

importantes dans la détection des spams d'opinion, elles ont été largement utilisées et donnent de bonnes performances, cependant en raison de l'émergence d'algorithmes générateurs de spam et du fait que les rédacteurs de spam sont aujourd'hui si professionnels que leur style est suffisamment subtile pour échapper même aux yeux des experts en détection de spam d'opinion, ces caractéristiques sont devenues insuffisantes pour différencier efficacement les opinions véridiques des spams.

- **Caractéristiques liées au comportement de l'auteur :**

Dans le but de pallier ces lacunes, les chercheurs se sont tournés vers une autre entité importante en matière de spam d'opinion, le spammeur, c'est-à-dire qu'ils ont commencé à utiliser des caractéristiques mettant en valeur le comportement de l'examinateur. Ces caractéristiques ont tendance à concerner l'opinion actuelle (les informations qui l'entourent) et l'histoire de son auteur. Nous allons maintenant présenter certaines des caractéristiques les plus utilisées dans la littérature, nous les avons extraites des travaux suivants (Andresini et al. 2022 ; Harris 2022 ; Hussain et al. 2019 ; Jindal et B. Liu 2008 ; Mukherjee et al. 2013a ; Ott et al. 2011 ; Rastogi et Mehrotra 2018). Dans ce qui suit, **r** fait référence à un avis, **a** à un auteur et **p** à un produit.

1) **ARD : Average rating deviation** : L'écart de notation moyen peut être utile pour identifier les spammeurs, vu qu'un examinateur impartial devrait évaluer un produit en fonction de la note moyenne attribuée à celui-ci. Cependant, les spammeurs tentant de promouvoir ou de rétrograder un produit lui attribuent généralement une note élevée ou extrêmement basse, ce qui peut nous aider à les identifier. Cette caractéristique peut être calculée selon la formule 2.4.

$$ARD\,(a) = avg\frac{|r_{ap} - \bar{r}_p|}{4} \tag{2.4}$$

2) ***Maximum Review Similarity (MRS) :*** Le fait qu'un seul évaluateur publie des avis similaires sur un produit distinct pourrait être une forte indication qu'il est un spammeur, vu que la rédaction de nouvelles critiques prend du temps, ces derniers ont tendance à copier leurs avis pour différents produits. C'est pourquoi il peut être utile de calculer la similarité maximale des avis à l'aide de la similarité cosinus afin de détecter cette anomalie. On peut la calculer selon la formule 2.5 tandis que le cosine entre ces deux opinions s'obtient par la formule 2.6.

$$MRS\,(a) = max_{r_i, r_j \in R_a, i<j} cos(r_i, r_j) \tag{2.5}$$

$$cos(r_i, r_j) = \frac{r_i r_j}{\|r_i\|\|r_j\|} \tag{2.6}$$

3) ***Average Review Length (ARL) :*** Plusieurs recherches ont montré que la plupart des spammeurs ne détaillent pas leurs critiques, ce qui semble logique car la plupart d'entre eux ne connaissent pas grand-chose aux produits qu'ils examinent et ne passent pas beaucoup de temps à rechercher leur cible et à rédiger leurs avis. Ainsi, la longueur moyenne de l'opinion d'un évaluateur a été considérée par beaucoup comme une caractéristique qui pourrait aider à distinguer les spammeurs des examinateurs honnêtes. Pour calculer cette caractéristique, il suffit de rassembler tous les avis écrits par l'examinateur et de calculer la moyenne de leur longueur, certaines études transforment ensuite cette valeur calculée en un booléen égalant vrai s'il était inférieur à une certaine valeur, ou faux sinon.

4) ***Burstiness (BST) :*** Des études rapportent que les examinateurs authentiques publient des critiques de temps en temps, tandis que les spammeurs d'opinion qui ne sont généralement pas des membres de longue durée d'un site publient plus fréquemment. Cette pratique irrégulière consistant à publier plusieurs avis dans un court laps de temps pourrait être un indicateur qu'un utilisateur est un spammeur. L'éclatement des avis est défini à l'aide de la fenêtre d'activité d'un utilisateur (la différence entre les dates de publication du premier et du dernier avis), si les avis sont publiés sur une période raisonnablement longue, cela indique probablement une activité normale, cependant, lorsque la plupart des avis sont publiés dans un délai très court cela pourrait indiquer une infliction de spam. En fixant une valeur seuil τ (28 jours dans certaines études que nous avons examinées), nous définissons la formule de cette caractéristique comme expliqué dans la formule 2.7, où *F(a)* et *L(a)* désignent respectivement le premier et le dernier commentaire publié par un auteur.

$$BST(a) = \begin{cases} 0, & \text{Si } L(a) - F(a) > \tau \\ 1 - \dfrac{L(a) - F(a)}{\tau}, & \text{Sinon} \end{cases} \tag{2.7}$$

5) ***Early review ratio (ERR) :*** Étant donné que les premiers avis ont un impact significatif sur les ventes de produits et services, la plupart des spammeurs essaient d'être parmi les premiers examinateurs afin de contrôler le sentiment et d'induire les acheteurs en erreur. Afin de calculer le ratio des premières critiques pour chaque auteur, nous appliquons la formule 2.8.

$$ERR\,(a) = \frac{|\{r \in R_a : r\ isfirst\ review\}|}{|R_a|} \tag{2.8}$$

6) ***Maximum Number of Reviews per Day (MRD) :***

Dans de nombreuses études, il a été observé que 90% des utilisateurs n'écrivent jamais plus d'un avis en une journée, tandis qu'environ 75% des spammeurs publient plus de 5 avis certains jours. Ainsi, ce comportement anormal pourrait être un indicateur utile. Nous définissons cette caractéristique en calculant le nombre maximum d'avis par jour pour un auteur et en le normalisant par la valeur maximale de nos données en suivant la formule 2.9.

$$MRD(a) = \frac{MaxRev(a)}{\max_{a \in A}(MaxRev(a))} \tag{2.9}$$

7) ***Positive Ratio (PR) et Negative Ratio (NR) :***

La polarité des critiques rédigées par un utilisateur peut être un indicateur pour savoir s'il est un spammeur ou non. Par exemple, si un auteur a un pourcentage élevé d'avis positifs sur certains produits, il pourrait ne pas être digne de confiance, et comme environ 85% des spammeurs écrivent plus de 80% de leurs avis positifs, cela pourrait valoir la peine d'être étudié. Le ratio d'avis positifs ou négatifs pour un auteur est calculé en divisant le nombre d'avis de la polarité choisie par le nombre total d'avis qu'il a publié.

8) ***Early rating deviation (ERD) :***

Comme indiqué précédemment, les spammeurs essaient de maximiser leur impact en publiant leurs avis le plus tôt possible. Nous utilisons l'écart de notation précoce pour capturer la nature frauduleuse des avis rédigés au début du lancement du produit en utilisant la formule 2.10 où $w_{rp} = \frac{1}{(t_{rp})^{\alpha}}$ est le poids d'un avis sur un produit et t_{rp} est l'heure à laquelle il a été publié.

$$ERD\,(r) = \frac{\left|r_{_p} - \overline{r}_{_p}\right|}{4} \cdot_{W_{r_p}} \tag{2.10}$$

9) ***Les métadonnées :***

Ces caractéristiques donnent des informations sur l'avis et son auteur plutôt que sur le contenu, des caractéristiques telles que l'adresse IP d'un utilisateur, son comportement sur le site Web, etc. lorsqu'elles sont combinées avec d'autres caractéristiques peuvent s'avérer bénéfiques pour détecter des avis et des comportements anormaux ou étranges. Cependant, ces données ne sont pas disponibles sur la plupart des sources de données, ce qui limite ainsi leur utilité pour détecter les spams.

Puisque les caractéristiques générées selon cette famille d'approches sont spécifiques et dépendent du contexte et de la tâche à accomplir, il n'existe pas d'ensemble universel de caractéristiques qui fonctionne bien pour toutes les tâches. De plus, l'ingénierie des caractéristiques est manuelle et dépend fortement de la contribution d'experts, ce qui demande beaucoup de travail et de temps.

2.7.2 Apprentissage des représentations

D'autre part, l'apprentissage des représentations ou apprentissage des caractéristiques est un ensemble de techniques permettant la transformation automatique des données d'entrée. Son objectif principal est de fournir des représentations abstraites et utiles pour les tâches de ML telles que la classification, la prédiction, la segmentation d'image, la reconnaissance de la parole et bien d'autres encore (Bengio et al. 2013 ; Yan et Yu 2019).

Par rapport aux caractéristiques conçues à la main décrites ci-dessus, l'apprentissage des caractéristiques est automatique, prend moins de temps et nécessite une connaissance minimale du domaine humain pour produire de meilleurs résultats. De plus, contrairement à l'ingénierie des caractéristiques, ce type d'apprentissage ne nécessite pas d'efforts supplémentaires pour concevoir des caractéristiques pour une nouvelle tâche et a une plus grande capacité de généralisation. Selon la plupart des études (Zhong et al. 2019), l'apprentissage des caractéristiques peut être divisé en deux catégories selon le niveau de leur abstraction hiérarchique des données : superficiel ou profond.

En raison de l'indisponibilité de données à grande échelle, de l'existence de nombreux paramètres et de la non-convexité élevée des réseaux, les architectures profondes complexes n'ont pas pu être facilement entraînées (Zhong et al. 2019), ce qui a fait que les méthodes d'**apprentissage superficiel des représentations** ont dominé ce domaine auparavant. Ce type de techniques ML consiste à extraire des caractéristiques des données sans transformations profondes. Ces techniques impliquent généralement des modèles linéaires et non linéaires ou de simples réseaux de neurones à anticipation avec un petit nombre de couches, et elles ne nécessitent pas de ressources de calcul étendues ni de grandes quantités de données d'entraînement. Ces techniques sont souvent utilisées dans des cas où les données sont relativement simples et peuvent inclure des méthodes supervisées telles que LDA, et non supervisées comme le clustering (K-means, Modélisation des mélanges gaussiens (GMM), etc), la factorisation matricielle (PCA, ICA,...) et certains réseaux de neurones simples tels que les cartes auto-organisées (SOM) et les auto-encodeurs simples. Certains de ces concepts ont été présentés précédemment pour d'autres tâches, cela est

dû au fait que plusieurs techniques ML peuvent servir à diverses fins.

Plus récemment, grâce à la disponibilité des données, a l'augmentation de la puissance de calcul et des machines, l'apprentissage profond est devenu plus renommé et les chercheurs ont commencé à l'utiliser davantage pour les tâches de représentation de caractéristiques. Il a été introduit pour la première fois dans le domaine de NLP en 2012 après le succès majeur qu'il a démontré dans la reconnaissance d'objets et de la parole (Zhou et al. 2020). Contrairement aux algorithmes d'apprentissage superficiel, **les méthodes d'apprentissage profond** des représentations peuvent généralement entraîner de meilleures abstractions des données d'origine pour les tâches de classification et de détection. Dans ce qui suit, nous discuterons de certaines des méthodes les plus intéressantes dans ce type d'apprentissage de représentation :

2.7.2.1 Auto-encodeurs :

Les encodeurs automatiques sont une technique de ML que nous avons vue précédemment, c'est un outil polyvalent qui peut être utilisé pour une variété de tâches. Dans le contexte de la réduction de dimensionnalité, comme nous l'avons vu dans la section 2.7.1.2, les auto-encodeurs sont entraînés sur des données d'entrée de grande dimension et apprennent une représentation compressée des données dans un espace de dimension inférieure. Alors qu'en termes d'apprentissage de caractéristiques, un auto-encodeur peut être formé sur un ensemble de données pour apprendre une représentation compacte et significative des données d'entrée. Les modèles basés sur l'auto-encodeur sont des réseaux de neurones qui tentent de copier leurs entrées vers leurs sorties, ils sont considérés comme faisant partie des modèles d'apprentissage non supervisés les plus robustes pour extraire des caractéristiques efficaces et discriminantes à partir d'un grand ensemble de données non étiqueté (Baghaei et al. 2022). Leur architecture générale est constituée de deux composants, la fonction **encodeur** $f(x)$ qui vise à transformer l'entrée x en une variable latente h de dimensions inférieures et la fonction **décodeur** g dont le but est de reconstruire l'entrée $\hat{x}$ étant donné la variable précédente h comme le montre la figure 2.6.

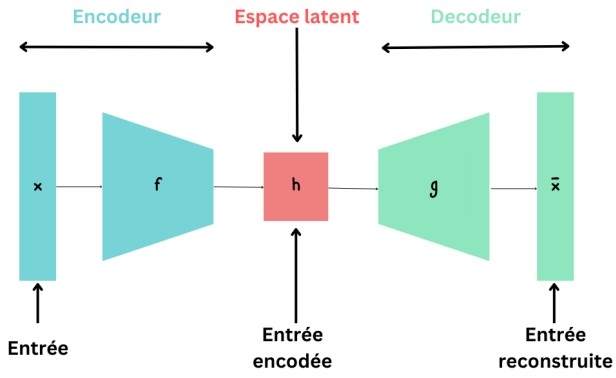


Fig. 2.6 : Architecture simplifiée d'un auto-encodeur

Le processus d'apprentissage consiste à ajuster les poids de ces composantes en fonction de la fonction de perte de la reconstruction, et il existe de nombreuses variables d'auto-encodeurs : les sous-complets, les débruiteurs (DAE), les clairsemés (SAE), les variationnels (VAE) et les contractifs (CAE) (Baghaei et al. 2022). (Dong et al. 2020) ont proposé une architecture de forêt décisionnelle d'auto-encodeur neuronal et rapportent une mesure F-score de 95,11%. (Saumya et Singh 2022), de leur côté, ont utilisé un auto-encodeur LSTM dans le but d'apprendre la structure interne des avis spams tout en

conservant les longues séquences existantes dans ces revues, et ils ont signalé un F-score de 99% en utilisant une représentation One Hot Embedding.

2.7.2.2 Word Embeddings :

Les Word Embeddings, ou prolongements de mots, consistent en des représentations des mots comme des vecteurs denses dans un espace de grande dimension, dans le but de capturer les relations sémantiques et syntaxiques entre les mots. Chaque mot est mappé sur un vecteur, et les mots ayant des significations similaires sont mappés sur des vecteurs proches les uns des autres dans l'espace vectoriel. Dans d'autres études, les chercheurs ont également appliqué des intégrations de phrases (Sentence embeddings) via des RNNs, des CNNs et des réseaux d'auto-attention (Zhou et al. 2020). Dans la littérature, il existe deux familles principales de Word embeddings, celle des embeddings statiques et l'autre dynamique (Y. Wang et al. 2019). Ces prolongements jouent souvent le rôle de la première couche de traitement de données pour les approches d'apprentissage profond.

- **Word Embeddings statiques** :

Les Word Embeddings statiques sont un type de technique de word embedding utilisée dans le traitement du langage naturel pour représenter les mots sous forme de vecteurs denses de dimension fixe dans un espace de grande dimension. Ces incorporations sont « statiques » dans le sens où elles sont pré-formées sur un grand corpus de texte et ne changent pas au cours d'une tâche ou d'un processus d'apprentissage spécifique. Les vecteurs capturent les relations sémantiques et syntaxiques entre les mots en fonction de leurs modèles de co-occurrence dans le corpus d'entraînement, et peuvent être utilisés pour représenter la signification de mots individuels ou de documents entiers. Dans les intégrations de mots statiques, un mot est mappé sur une représentation vectorielle fixe, quel que soit son contexte ou la manière dont il est utilisé dans une phrase ou un document spécifique. Cela signifie qu'un mot aura toujours la même représentation vectorielle, quelle que soit sa signification dans différents contextes. Par exemple, dans la phrase "I rode the train to get to the gym so I could train with my peers" le mot **train** sera mappé sous le sens 'train' et donc ces autres significations (e.g. 'formation') ne seront pas prises en considération. Les techniques d'incorporation de mots statiques populaires incluent Word2Vec (Mikolov et al. 2013) et ses extensions (par exemple Sent2Vec (Le et Mikolov 2014) et Doc2Vec (Le et Mikolov 2014)), GloVe (Pennington et al. 2014) et fastText (Bojanowski et al. 2017).

- **Word2vec** a été introduit pour la première fois en 2013 (Mikolov et al. 2013), c'est l'un des modèles de word embeddings les plus populaires qui est basé sur un réseau neuronal pour générer des intégrations de mots statiques. Le modèle Word2vec apprend à représenter des mots dans un espace vectoriel continu où les mots ayant des significations similaires sont plus proches les uns des autres. L'un des exemples fréquemment donnés est l'équation *" Word2vec(king) - Word2vec(man) + Word2vec(woman) = Word2vec(queen)"* c'est-à-dire que Word2vec mappe les mots de manière à ce que la valeur vectorielle obtenue par la soustraction et l'addition de ces vecteurs est égale au vecteur correspondant à l'expression "queen". Le modèle est entrainé sur un grand corpus de texte, et l'objectif est soit de prédire la probabilité d'un mot compte tenu de son contexte, soit la probabilité d'un contexte compte tenu d'un mot. Sur cette base, il existe deux algorithmes principaux utilisés dans Word2vec : Sac continu de mots (CBOW : Continuous Bag of Words), dans lequel le modèle est formé pour prédire un mot en fonction de son contexte (l'ensemble

de mots qui apparaissent dans une fenêtre fixe du mot cible), et Skip-Gram, qui prédit le contexte étant donné le mot cible (Mohawesh et al. 2021b). La figure 2.7 présente une architecture simplifiée de ces deux algorithmes pour une meilleure compréhension. Afin de minimiser la différence entre les probabilités prédites et les probabilités réelles, le modèle est formé en ajustant les poids du réseau de neurones à l'aide de la rétropropagation. Parmi les travaux ayant utilisé Word2vec on peut citer (Andresini et al. 2022 ; Shehnepoor et al. 2021) pour l'algorithme CBOW et (Hajek et al. 2020 ; Kennedy et al. 2020) pour celui de Skip-Gram.

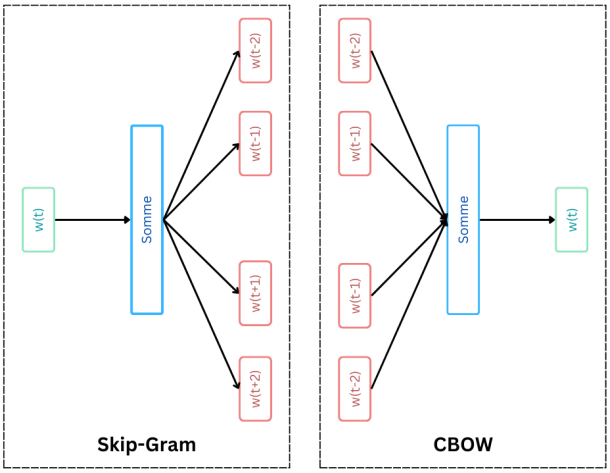


Fig. 2.7 : Architecture simplifiée des algorithmes Word2vec

- **GloVe**, abréviation de Global Vectors for Word Representation, a été introduit pour la première fois dans (Pennington et al. 2014) suite au succès de Word2vec. Ce modèle d'intégration est basé sur une matrice globale de co-occurrence de mots qui utilise une approche similaire à Word2Vec, mais au lieu de prédire les mots en s'appuyant uniquement sur les informations locales, il vise à optimiser directement les vecteurs de mots afin qu'ils capturent les propriétés distributionnelles du corpus incorporant à la fois des relations locales et globales entre les mots. Il s'agit d'une méthode basée sur le comptage, ce qui signifie qu'elle utilise le nombre de co-occurrences de mots pour apprendre les vecteurs de mots. L'idée principale derrière GloVe est que le rapport des probabilités de co-occurrence de deux mots doit être égal au rapport de leurs longueurs vectorielles. Cette approche permet à GloVe de capturer les similitudes sémantiques et syntaxiques entre les mots d'une manière plus directe et efficace que d'autres méthodes basées sur le comptage comme LSA par exemple.

Le problème rencontré par Word2vec et Glove est leur incapacité à traiter les mots hors vocabulaire (OOV), donc si le modèle ne voyait pas un terme pendant la formation, il ne serait pas en mesure de le représenter par la suite.

- **FastText** tente de résoudre ce probleme. **FastText** est une variante de Word2vec basée sur le modèle Skip-Gram, et il a été introduit dans (Bojanowski et al. 2017). L'une des principales différences entre eux est que FastText apprend les représentations vectorielles des unités de sous-mots, telles que les n-grammes de caractères, en plus des mots complets. Cela permet au modèle de mieux gérer les mots OOV ainsi que de capturer la morphologie des mots. De plus, les représentations de sous-mots de FastText peuvent capturer la signification de mots qui ne sont pas compositionnels, tels que le préfixe et le suffixe d'un mot. FastText représente chaque mot avec un sac de n-grammes, par exemple

le mot 'Delight' est représenté pour n=4 par le 4-gramme : *'<Del', 'Deli', 'elig', 'ligh', 'ight', 'ght>'* où '<' et '>' indiquent le début et la fin du terme, respectivement. L'un des avantages de FastText est sa vitesse et son évolutivité, ce qui lui permet de s'entraîner efficacement sur de grands ensembles de données contenant des millions ou des milliards de mots. FastText comprend également des modèles pré-formés pour plusieurs langues, ce qui facilite son utilisation pour les langues autres que l'anglais.

- **Word embeddings dynamiques** :

Contrairement aux plongements de mots statiques, dans les plongements de mots dynamiques, la représentation vectorielle d'un mot n'est pas figée et peut varier en fonction du contexte dans lequel il apparaît. Cela signifie qu'un mot peut avoir différentes représentations vectorielles en fonction du contexte dans lequel il apparaît, ce qui peut capturer les différentes significations du mot dans différents contextes. Les Word embeddings dynamiques, également appelées Word embeddings contextualisées, sont générées en formant un réseau de neurones sur un grand corpus de données textuelles à l'aide d'une tâche de modélisation du langage, comme prédire le mot suivant dans une phrase ou remplir un blanc. Il en résulte un modèle qui peut générer une représentation vectorielle pour n'importe quel mot donné, en tenant compte de ses mots environnants dans le contexte donné. Des exemples courants d'incorporations de mots dynamiques incluent ELMo (Peters et al. 2018), ULMFiT (Howard et Ruder 2018) et BERT (Devlin et al. 2018).

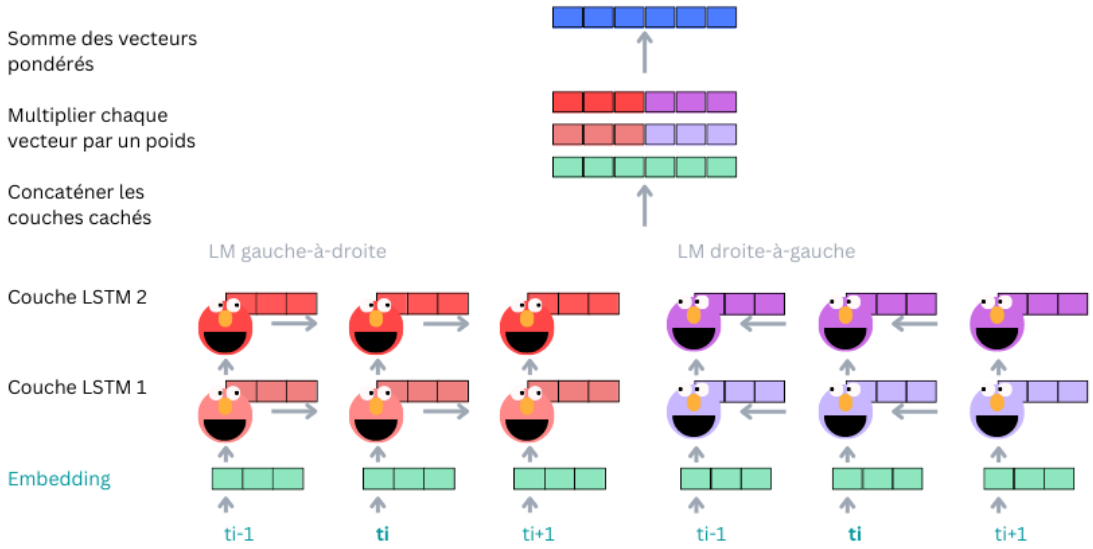


Fig. 2.8 : Architecture simplifiée de ELMo

- **ELMo**, qui signifie "Embeddings from Language Models", est une représentation de mots contextualisée profonde introduite pour la première fois en 2018 (Peters et al. 2018), elle génère des incorporations de mots en fonction des contextes dans lesquels ils sont utilisés pour enregistrer le sens du mot et récupérer des informations contextuelles supplémentaires. ELMo est basé sur un modèle de langage neuronal qui apprend à prédire le mot suivant dans une séquence de mots et est entraîné sur de grands corpus de texte à l'aide d'un réseau bidirectionnel profond LSTM (mémoire longue à court terme), ce qui lui permet de capturer les dépendances contextuelles des mots dans une phrase. La sortie de ce réseau est un ensemble de word embeddings contextualisés qui représentent la signification d'un terme dans le contexte des mots environnants dans une phrase, cette architecture est présentée (de manière simplifiée) dans la figure 2.8. Grâce à la pré-formation, il peut

représenter plus précisément des mots polysémiques dans une variété de contextes et est plus informatif sur la sémantique de niveau supérieur du texte.

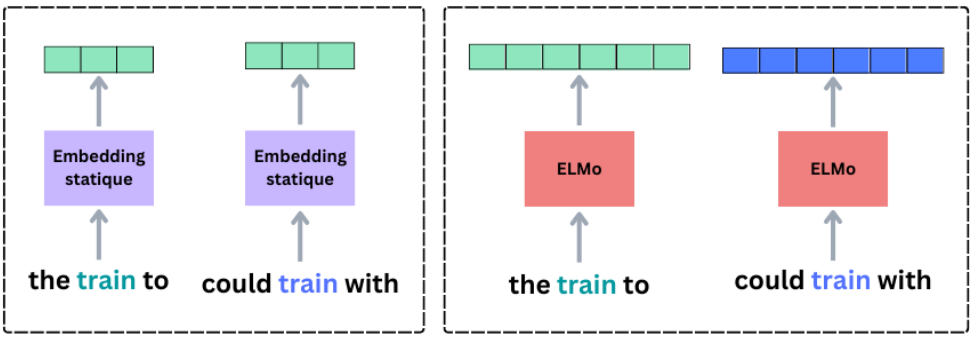


Fig. 2.9 : Exemple du fonctionnement de ELMo

Si nous utilisons la phrase "I rode the *train* to get to the gym so I could *train* with my peers" comme exemple une fois de plus, les prolongements présentés précédemment mapperaient le mot "train" avec le même vecteur, ELMo de son côté forme son réseau bidirectionnel afin que son modèle de langage ait un sens à la fois du mot précédent et du suivant par conséquent le mot 'train' sera représenté avec un vecteur différent selon le contexte, comme l'illustre la figure 2.9.

- **BERT**, ou Bidirectional Encoder Representations from Transformers (Représentations d'encodeurs bidirectionnels à partir de transformateurs), est un réseau neuronal profond qui a été introduit en 2018 pour les tâches de traitement du langage naturel (Devlin et al. 2018), il s'agit d'un modèle de langage pré-entraîné développé par Google qui utilise une architecture de transformateur bidirectionnel combinant l'apprentissage non supervisé et supervisé pour générer des représentations de mots contextualisées de haute qualité. L'architecture de BERT comprend un encodeur avec plusieurs couches d'attention multi-têtes. BERT a été entraîné sur des quantités massives de données textuelles (paragraphes de texte Wikipédia anglais de 2500 millions de mots et corpus de livres de 800 millions de mots), l'une de ses caractéristiques déterminantes est sa stratégie de préformation, appelée Masked Language Model (MLM) où il masque de manière aléatoire des mots dans la phrase d'entrée, et l'objectif du modèle est de prédire ces mots masqués en fonction de leur contexte environnant. De plus, il utilise une tâche de prédiction de phrase suivante, où il entraîne le modèle pour déterminer si deux phrases sont consécutives ou non.

Une fois que tout cela est fait, l'encodeur peut être réglé avec précision pour la tâche spécifique à accomplir. Cela implique l'entraînement du modèle à l'aide d'un ensemble de données plus petit et spécifique à la tâche. Au cours de ce processus, les poids du modèle BERT pré-entrainé sont ajustés pour faire des prédictions sur la tâche spécifique. La figure 2.10 est directement extraite de l'article original (Devlin et al. 2018) présentant des exemples de fine-tuning de BERT sur différentes tâches.

Dans le même article (Devlin et al. 2018), les auteurs ont introduit deux variantes de BERT, à savoir **BERT-BASE** et **BERT-LARGE**. BERT-BASE se compose d'une pile de 12 blocs de Transformations (une sequence d'encodeurs) ainsi que 12 têtes d'attention. BERT-LARGE, quant à lui, se distingue de BERT-BASE par 24 blocs de tranformations et 16 têtes d'attention, mais pas que ! BERT-LARGE a été entraîné sur un corpus de texte beaucoup plus grand que la version BERT-BASE, ce qui lui permet de capturer

des relations plus complexes entre les mots, cependant il nécessite plus de mémoire et de puissance de calcul pour l'entraînement et l'inférence que la version BERT-BASE.

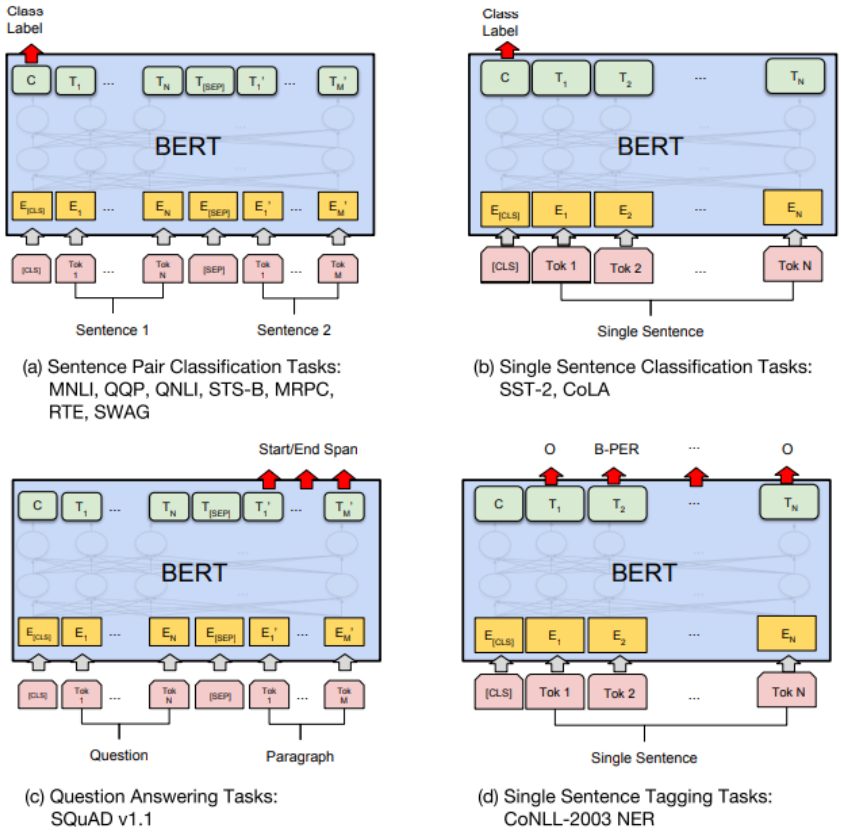


Fig. 2.10 : Exemples d'affinage de BERT sur différentes tâches (Devlin et al. 2018)

D'autres variantes de BERT sont apparues, telles que **DistilBERT** qui en est une version allégée visant à atténuer certaines des limites de BERT telles que la complexité de calcul, la taille de longueur d'entrée fixe et le problème d'intégration de mots, et **RoBERTa** une version étendue de BERT qui peut surpasser les performances de BERT en entraînant le modèle plus longtemps et sur des séquences plus longues (Mohawesh et al. 2021b). Nous pouvons également mentionner **DziriBERT**, le premier modèle de langage basé sur Transformer qui a été pré-entraîné spécifiquement pour le dialecte algérien (Abdaoui et al. 2021).

L'approche de fine-tuning n'est pas la seule façon d'utiliser BERT. Tout comme ELMo, le modèle BERT pré-entraîné peut être directement utilisé pour extraire des représentations de mots contextualisées riches à partir de données textuelles, sans avoir besoin d'une formation supplémentaire, avant d'alimenter ces prolongements dans un modèle de classification. L'un des avantages de cette approche est qu'elle ne nécessite pas de données étiquetées pour l'affinage, ce qui peut être particulièrement utile dans des contextes où les données étiquetées sont rares ou coûteuses à obtenir. Cependant, l'utilisation de cette méthode peut nécessiter des ressources de calcul importantes, en particulier pour les modèles plus grands tels que **BERT-LARGE**, en raison de la complexité de la tâche de calcul des word embeddings contextualisées.

Parmi les travaux qui ont exploité les performances exceptionnelles de BERT pour la problématique de détection de spam d'opinion, nous citons : (Andresini et al. 2022 ; Kennedy et al. 2020 ; Mir et al. 2023 ; Mohawesh et al. 2021b ; Shaalan et al. 2021).

Nous avons passé en revue les principales techniques de représentation et leur classification. L'extraction et l'apprentissage de caractéristiques transforment tous deux les données brutes, comme expliqué ci-dessus, afin de trouver des caractéristiques pertinentes, tandis que la sélection de caractéristiques et la réduction de dimensionnalité ont pour objectif de rendre le processus moins complexe sans sacrifier trop de puissance de prédiction. Il peut être intéressant de noter que les techniques d'extraction et d'apprentissage produisent souvent un grand nombre de caractéristiques qui peuvent être redondantes et entraîner des exigences de calcul excessives qui, à leur tour, peuvent rendre leur application en temps réel peu pratique ou inutilement difficile, par conséquent, il est courant d'utiliser la sélection de caractéristiques et la réduction de la dimensionnalité pour remédier à ce problème. La combinaison de différentes caractéristiques et différentes représentations peut améliorer les performances obtenues (Andresini et al. 2022).

2.8 Approches de classification des spams d'opinion

L'objectif principal de la détection de spam d'opinions est d'identifier chaque avis spam ou chaque spammeur, les approches adoptées pour réaliser cela sont diverses et peuvent être classifiées selon plusieurs critères : l'entité à détecter (opinion spam, auteur spammeur, produit ciblé,...), le type d'approche de représentation avec qui cette classification est utilisée, etc. Dans cette section, nous nous intéresserons aux approches les plus citées dans la littérature pour la tache de classification selon leur architecture, nous les catégoriserons donc en deux familles : celle des modèles ML traditionnels et celle des modèles neuronaux (Mohawesh et al. 2021b ; Ren et Ji 2019).

2.8.1 Modèles ML traditionnels

Les modèles de classification traditionnels ont souvent une structure simple et un temps d'apprentissage et de prédiction relativement rapides, ils sont souvent associés à des approches d'ingénierie de caractéristiques (généralement Bag-Of-Words et les n-grammes) qui pourraient améliorer les performances des modèles, en particulier dans le cas des données structurées. Cependant, ils pourraient ne pas être en mesure de capturer des relations plus complexes et non linéaires dans les données, sans oublier que l'utilisation d'approches d'ingénierie de caractéristiques nécessite du temps et une expertise du domaine (comme indiqué précédemment dans la sous-section 2.7.1). Du point de vue ML, ces modèles peuvent être divisés en 3 classes, celle de l'apprentissage supervisé, semi-supervisé, et non supervisé (Ren et Ji 2019).

2.8.1.1 Apprentissage supervisé

Puisque le problème de détection de spam d'opinion a été considéré comme un problème de classification binaire avec deux classes "Spam" et "Non Spam" (si l'entité considérée est l'opinion) (Jindal et B. Liu 2008), de nombreux chercheurs ont opté pour des méthodes d'apprentissage supervisé. Ces méthodes entraînent un modèle d'apprentissage automatique à l'aide de données étiquetées pour classer les spams d'opinion. L'entrée du modèle se compose de caractéristiques extraites des données brutes telles que les n-grammes, les sacs de mots, les POS tags et d'autres caractéristiques textuelles et comportementales jugées importantes. Les algorithmes couramment utilisés dans l'apprentissage supervisé comprennent les machines à vecteurs de support (SVM), les arbres de décision, Naïve Bayes et la régression logistique. Le principal avantage de l'utilisation

de ces modèles pour la détection des spams d'opinion est qu'ils sont relativement faciles à mettre en œuvre et peuvent atteindre de bonnes performances avec de petits ensembles de données. De plus, ils permettent l'interprétabilité, ce qui signifie qu'il est possible de comprendre comment le modèle fait des prédictions et quelles caractéristiques sont les plus importantes. Cependant, les ensembles de données étiquetés de qualité nécessaires pour entraîner ces modèles sont difficiles à acquérir.

Parmi les multiples approches qui utilisent l'apprentissage supervisé, mentionnons d'abord (Ott et al. 2011), dans ce travail, les auteurs ont construit un ensemble de données de référence qui est maintenant largement utilisé dans la détection du spam d'opinion, et est couramment appelé OpSpam [9]. Les auteurs de cet article ont testé à la fois les classifieurs **Naïve Bayes** et **SVM** avec des combinaisons de caractéristiques (identification de genre avec POS, psycholinguistiques avec LIWC, textuelles avec n-grammes, ...). Leurs exprimentations ont montré des résultats qui ont surpassé les résultats obtenus par le jugement humain et les méthodes de la littérature, leur approche la plus performante était la combinaison de LIWC et de bi-grammes avec le classificateur SVM, qui a donné une exactitude de 89,8%.

(Shahariar et al. 2019) ont appliqué plusieurs techniques sur les données acquises de Yelp[10], après leur pré-traitement. Ils ont appliqué à la fois les techniques ML traditionnelles (SVC, KNN, NB) et des méthodes d'apprentissage approfondi. Ces modèles ont été alimentés avec des caractéristiques sélectionnées via TF-IDF, N-Grammes et Word2vec. La combinaison la plus performante a été de jumeler Uni-Gram et Naïve Bayes, l'exactitude signalée était de 91,73%.

(Rastogi et al. 2020) est la première étude de son genre qui considère trois perspectives à la fois, celui de l'opinion, de son auteur et du produit qu'elle cible, afin d'analyser l'efficacité de la détection de spam en utilisant des caractéristiques textuelles et comportementales ainsi que 4 classifieurs (SVM, LR, MLP et NB) sur YelpZIP et YelpNYC (Rayana et Akoglu 2015). L'architecture de l'approche adoptée dans cette étude est illustrée dans 2.11.

Les résultats indiquent que les caractéristiques comportementales sont plus efficaces que les caractéristiques textuelles pour détecter le spam d'opinion dans les trois configurations avec les 4 classificateurs. De plus, les modèles formés sur des caractéristiques hybrides surpassent légèrement ceux formés sur des caractéristiques comportementales et il y a des cas où ces derniers font aussi bien, voire mieux que les hybrides. Les caractéristiques utilisées dans ce travail apportent une amélioration par rapport aux caractéristiques existantes utilisées dans d'autres travaux connexes. De plus, l'analyse du temps de calcul pour la phase d'extraction des caractéristiques montre une meilleure rentabilité des caractéristiques comportementales par rapport au textuel. (Rastogi et al. 2020)

[9] https ://myleott.com/op-spam.html
[10] https://www.yelp.com/

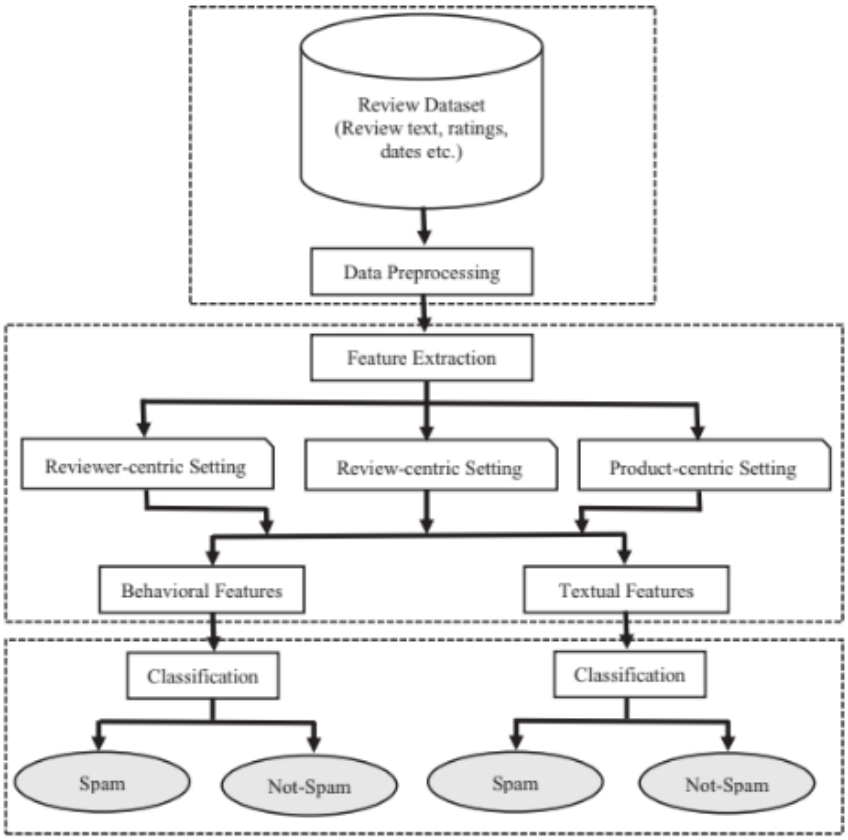


Fig. 2.11 : Architecture de l'approche adoptée dans (Rastogi et al. 2020)

Enfin, (Yao et al. 2021) proposent un modèle d'ensemble combinant 5 classifieurs supervisés (RF, Xgboost, Lightgbm, Catboost, GBDT) et des caractéristiques basées sur l'opinion (Notation (rating), Nombre de mots, TF-IDF) et son auteur (Nombre d'avis, Nombre de premiers avis, note moyenne,...), et les ont testées sur deux sous-ensembles de YelpCHI (Mukherjee et al. 2013b) (un pour les restaurants et l'autre pour les hôtels). Les auteurs ont traité le déséquilibre des données en combinant la méthode de recherche de grille et le rééchantillonnage en trouvant le meilleur rapport d'échantillonnage pour chaque classificateur. Ensuite, les caractéristiques extraites sont transmises séparément à chaque classifieur. Enfin, ils ont proposé deux méthodes afin d'améliorer les performances du modèle de classification : une méthode de vote majoritaire et d'empilement. Les résultats expérimentaux ont montré que le modèle proposé fonctionnait assez bien, mais ne surpassait pas les méthodes de pointe. Ils rapportent une exactitude de 72.06% sur le sous-ensemble YelpCHI-Restaurant et de 79.46% sur YelpCHI-Hôtel en suivant la stratégie d'empilement.

2.8.1.2 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est un type d'apprentissage qui combine l'apprentissage supervisé et non supervisé, où un algorithme apprend à partir d'un petit ensemble de données étiquetées et d'un ensemble de données non étiquetées beaucoup plus grand. L'idée est d'exploiter la grande quantité de données non étiquetées pour améliorer les performances du modèle dans l'apprentissage des données étiquetées. Ceci est particulièrement utile lorsque le coût de l'étiquetage des données est élevé ou lorsque les données étiquetées sont rares, c'est donc plus efficace que d'étiqueter toutes les données à la main, mais cela nécessite une sélection minutieuse des exemples à étiqueter et à réapprendre.

D'après (Ligthart et al. 2021a), on peut trouver dans cette catégorie plusieurs méthodes telles que l'apprentissage multi-vues, les modèles génératifs, l'apprentissage basé sur des graphes, ainsi que le **Self-Training** (Yarowsky 1995), le **Co-Training** (Blum et Mitchell 1998) et le **PU-Learning** (X.-L. Li et B. Liu 2005 ; B. Liu et al. 2002) auxquels nous allons nous intéresser.

L'approche **Self-Training** (Auto-apprentissage) consiste à former un classificateur sur une petite quantité de données étiquetées, puis à utiliser le classificateur pour prédire les étiquettes des données non étiquetées. Les prédictions à haute confiance sont ajoutées aux données étiquetées, et le processus est répété de manière itérative jusqu'à ce qu'un critère d'arrêt soit satisfait (toutes les données ont été étiquetées, aucune observation supplémentaire ne satisfait les critères, le nombre maximum d'itérations a été atteint,...). L'approche **Co-Training** (Co-Apprentissage) implique l'entraînement de deux classificateurs sur différents ensembles de caractéristiques et l'utilisation des prédictions de haute confiance de chaque classificateur pour étiqueter les données de l'autre. Cela vient de l'hypothèse que s'il y a deux vues indépendantes du même objet (elles utilisent chacune un ensemble de caractéristiques différent) alors deux classificateurs correctement entraînés sur ces vues doivent étiqueter cet objet de la même manière. Cela peut être efficace lorsque les caractéristiques ont des forces différentes dans des contextes différents. L'approche **PU-learning** (Positive-Unlabeled learning), quant à elle, traite d'un ensemble d'échantillons positifs étiquetés (opinions spam) et d'un ensemble d'échantillons non étiquetés (qui peuvent inclure à la fois des avis de spam et non-spam). L'objectif de PU-learning est d'entraîner un classificateur binaire capable de classer avec précision les échantillons positifs et de les distinguer des échantillons négatifs, sans utiliser d'exemples étiquetés négatifs. L'idée de base est de différencier entre les instances négatives et les instances positives appartenant à l'ensemble des données non étiquetées, puis utiliser le résultat de cette classification comme un jeu de données annotées pour apprendre un classifieur supervisé. Le PU-learning est particulièrement utile dans les cas où les échantillons négatifs sont difficiles à identifier ou à collecter, ce qui est souvent le cas dans les tâches de détection de spam ou de fraude. En utilisant cette approche, il est possible d'obtenir des résultats précis en termes de classification des échantillons positifs, sans avoir besoin d'une grande quantité d'exemples négatifs étiquetés.

Il existe plusieurs travaux ayant exploité les techniques semi-supervisées, (Ligthart et al. 2021a) de leur part en ont étudié plusieurs, à savoir l'auto-apprentissage, le co-apprentissage et le SVM transductif sur le dataset OpSpam. Ils ont aussi testé ces modèles avec différentes combinaisons de caractéristiques dont les Uni-grammes, les Bi-grammes, la combinaison de TF-IDF et des Bi-grammes ainsi que la combinaison de POS et les Uni-grammes. Le meilleur résultat obtenu sur le dataset OpSpam était une exactitude de 93% en combinant les Bi-Grammes avec l'algorithme de **Self-training** (avec Naïve Bayes comme classifieur de base), cette méthode a donné les meilleures performances même durant les tests additionnels sur des sous-ensembles de YelpCHI.

(J. Wang et al. 2020), de leur part, ont proposé un modèle combinant plusieurs caractéristiques (Sémantiques, lexicales, basées sur les sentiments, sur des informations externes telles que le comportement anormal de l'auteur, etc), ils ont même proposé une méthode afin de juger si l'émotion exprimée par l'auteur pourrait être une caractéristique significative. Ils ont jugé que la combinaison donnerait une bonne performance et qu'une

combinaison optimale de classifieurs n'est sélectionnée qu'à travers plusieurs expérimentations, pour cela, ils ont utilisé 7 modèles ML (SVM, Naïve Bayes, Random Forest, KNN, LDA, et les arbres de décision) afin d'identifier la crédibilité de l'opinion. En expérimentant sur YelpCHI, l'approche de **Co-training** avec fusion des multiples caractéristiques a donné les meilleurs résultats avec une exactitude de 84.45% et une mesure F-score de 81.89%.

En ce qui est de l'approche **PU-Learning**, nous pouvons citer le travail de (Fusilier et al. 2015) qui ont proposé une variante conservatrice de la méthode originale (B. Liu et al. 2002) utilisant le dataset OpSpam. En ce qui est du classifieur utilisé, selon leurs expérimentations, Naïve Bayes a surpassé SVM en utilisant une combinaison d'Uni-grammes et de Bi-grammes. Ils ont aussi essayé d'analyser le rôle de la polarité pour la détection de spam d'opinion dans un dataset de polarité mixte avec deux configurations : soit en utilisant un seul classificateur pour les spams négatifs et positifs, soit en considérant les spams positifs et négatifs comme deux problèmes différents (donc 2 classificateurs). Les résultats de cette expérimentation montrent que le traitement de la détection de spam positif et négatif comme un seul problème donne de meilleurs résultats, on constate donc que ces deux types de spam ont des points en commun qui aident à améliorer la classification. L'approche adoptée a donné une mesure F-score de 81.1% pour la détection de spams positifs et de 72.3% pour les spams négatifs, ces résultats montrent que cette approche modifiée surpasse le travail original et que la détection du spam négatif est plus difficile que la détection du spam positif, mais qu'il reste préférable d'avoir un seul classificateur pour analyser les deux types d'opinions que d'utiliser deux classificateurs distincts.

Plus récemment, (Tian et al. 2020) ont proposé une approche robuste non-convexe semi-supervisée sous le nom de "Ramp One-Class SVM". Elle est nommée comme cela car elle combine SVM à une classe et la fonction Ramp. **One-Class SVM** est une technique d'apprentissage non supervisée pour apprendre à différencier les échantillons de test d'une classe particulière d'autres classes, ils l'ont appliqué afin de gérer le manque de données étiquetées pour les opinions spams. En ce qui concerne la fonction de perte **Ramp**, il s'agit d'une fonction de perte non convexe, et ils exploitent cela afin d'éliminer les effets des valeurs aberrantes et des non-avis. Ils ont obtenu une exactitude de 74.37% sur Yelp et de 92.13% sur OpSpam.

Nous avons déjà mentionné le travail de (Ziani et al. 2021) en ce qui est des caractéristiques sémantiques qu'ils ont proposées pour la langue arabe. En combinant ces caractéristiques avec d'autres statiques ils ont appliqué un modèle de classification SVM semi-supervisé, ou S3VM (Bennett et Demiriz 1998), sur plusieurs datasets dont une version d'OpSpam traduite en arabe, pour lequel ils ont obtenu une exactitude de 93%.

2.8.1.3 Apprentissage non-supervisé

Comme nous pouvons le conclure de ce qui a été présenté jusqu'à présent, l'un des problèmes majeurs que rencontrent les méthodes supervisées et semi-supervisées est le fait qu'elles dépendent des datasets annotés pendant la phase d'apprentissage, et que la qualité desdits datasets peut affecter leur performance. Ainsi, de nombreux chercheurs ont décidé d'aborder le problème de détection de spam d'opinion sous l'angle de classification non supervisée. Ces modèles peuvent être utiles lorsque les données étiquetées sont rares ou coûteuses à obtenir, mais ils peuvent être moins efficaces que les méthodes supervisées

car ils n'ont pas accès à des données étiquetées sur lesquelles s'entraîner et ils nécessitent souvent plus de connaissances et d'affinage pour obtenir de bonnes performances.

Parmi les approches non supervisées, nous pouvons citer '**RLOSD**', acronyme de "Representation Learning based Opinion Spam Detection", qui a été proposée par (Sedighi et al. 2017). Cette étude propose une approche basée sur les arbres de décision ainsi qu'un apprentissage de représentation non supervisé. Pour la sélection des caractéristiques, ils utilisent des méthodes traditionnelles dont POS, TF-IDF et N-grammes, PCA et MMI. Le modèle a obtenu de bonnes performances durant les expérimentations sur Yelp et le dataset construit par (Jiwei Li et al. 2014) avec une mesure F-score toujours supérieure ou égale à 75%. Et selon (Mohawesh et al. 2021b), ce modèle proposé peut être amélioré en prenant en compte la corrélation des données dans le choix des caractéristiques appropriées.

L'étude de (Z. Wang et al. 2020) présente une approche basée sur l'intégration de réseau non supervisée pour apprendre les intégrations d'utilisateurs en exploitant conjointement les relations directes et indirectes entre les utilisateurs par paires. La figure ci-dessous 2.12 montre l'approche proposée au sein de l'architecture unifiée proposée nommée COSD, son objectif est de combiner conjointement l'exploration de voisinage directe et indirecte pour apprendre la représentation intégrée de chaque utilisateur pour identifier plus précisément les spammeurs. L'intégration de pertinence directe contrôle l'apprentissage des intégrations d'utilisateurs vers les utilisateurs par paires avec une forte intensité des caractéristiques collusoires. L'intégration indirecte de la pertinence, de son côté, a tendance à rapprocher les utilisateurs par paires partageant plus fréquemment les voisins de coévaluation. Les auteurs de ce papier ont opté pour l'utilisation des ypes de relations parce que ces deux types se renforcent mutuellement pour rapprocher les utilisateurs concernés des relations directes et indirectes des utilisateurs. Les auteurs de ce papier combinent 4 types différents de caractéristiques hétérogènes par paire (la proximité de la notation du produit, la proximité temporelle du produit, la catégorie cote proximité, la catégorie proximité horaire). Les résultats des expérimentations, évalués sur deux jeux de données AmazonCn et YelpHotel, montrent la supériorité de COSD comparée à d'autres méthodes de la littérature. On peut remarquer une amélioration de l'AUC de 12.04% sur AmazonCn et de 12.78% sur YelpHotel.

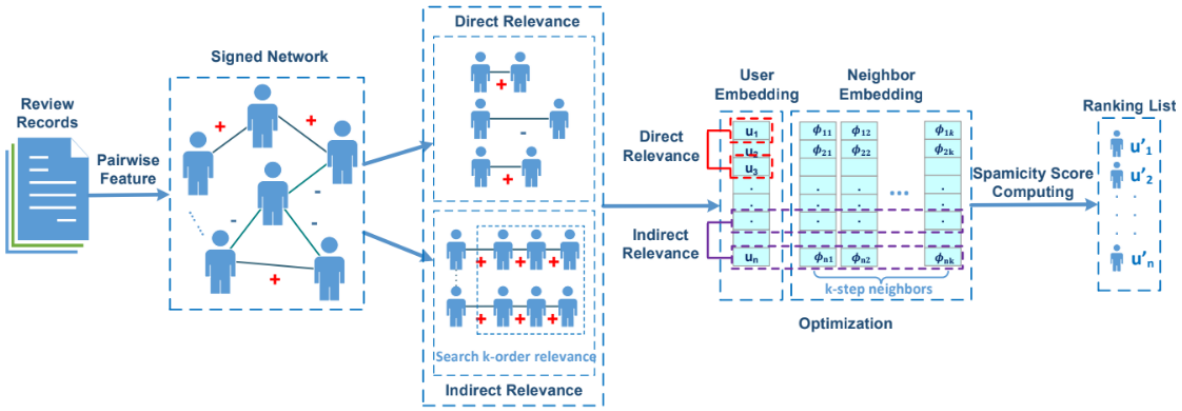


Fig. 2.12 : Architecture de l'approche adoptée dans (Z. Wang et al. 2020)

Le tableau 2.3 résume quelques travaux présents dans la littérature en mettant en

valeur, en addition à l'approche de classification utilisée, les caractéristiques représentées, le dataset exploité et certains résultats obtenus. Afin de simplifier la visualisation, nous nous contenterons de mentionner les résultats de l'approche la plus performante (pour certains travaux présentant les résultats des différentes expérimentations menées).

Tab. 2.3 : Travaux exploitant les modèles ML traditionnels pour le problème de détection de spam d'opinion

Référence	Dataset	Caractéristique	Classifieur	Résultats
Apprentissage supervisé				
(Ott et al. 2011)	OpSpam	Bi-Grammes	SVM	Accuracy = 89.6%
		LIWC + Bi-Grammes		Accuracy = 89.8%
(Etaiwi et Naymat 2017)	OpSpam	BOW + Suppression des mots vides	NB	Accuracy = 88.7%
(Shahariar et al. 2019)	Yelp	Unigrammes	NB	Accuracy = 91.73%
(Mohawesh et al. 2021a)	YelpCHI	TF-IDF	Analyse de la dérive de concept (SVM, LR, PNN)	Accuracy = 68.17%
	YelpNYC			Accuracy = 84.85%
	YelpZIP			Accuracy = 91.35%
	Yelp consumer electronic			Accuracy = 76.72%
(Hassan et Islam 2021)	OpSpam	Empath (Fast et al. 2016), Sentiment score	SVM	Accuracy = 64.06%
		Empath, Probabilistic sentiment score		Accuracy = 66.56%
		TF-IDF, Sentiment score		Accuracy = 86.25%
		TF-IDF, Probabilistic sentiment score		Accuracy = 89.37%
(Yao et al. 2021)	YelpCHI Restaurant	Caractéristiques sur le contenu et l'auteur	Modèle d'ensemble	Accuracy = 72.06%
	YelpCHI Hôtel			Accuracy = 79.46%

(Goel et al. 2021)	Yelp	Matrice Uni-gramme, Matrice Bi-gramme, Pourcentage POS, etc.	SVM	Accuracy = 70.03%
(Mir et al. 2023)	Multi-Dataset [Hôtel, Restaurant, Médecin]	BERT word embeddings	SVM	Accuracy = 87.81%
Apprentissage semi-supervisé				
(Fusilier et al. 2015)	OpSpam [Avis positifs]	Uni-grammes, Bi-grammes	PU learning modifiée	F-score = 81.1%
	OpSpam [Avis négatifs]			F-score = 72.3%
(Tian et al. 2020)	Yelp	TF-IDF	Ramp One-Class SVM	Accuracy = 74.37%
	OpSpam			Accuracy = 92.13%
(J. Wang et al. 2020)	YelpCHI	Corps de l'opinion + Caractéristiques de l'auteur	Co-Training avec multi-fusion des caractéristiques	F-score = 81.89%
(Ligthart et al. 2021a)	OpSpam	Bi-grammes	Self-Training avec NB comme classifieur de base	Accuracy = 93%
(Ziani et al. 2021)	OpSpam traduit en arabe	Caractéristiques sémantiques et statistiques	S3VM	Accuracy = 93%
Apprentissage non supervisé				
(Sedighi et al. 2017)	Yelp	Méthode de sélection de caractéristiques [POS, TF-IDF et N-grammes, PCA et MMI]	RLOSD [Arbre de décision]	F-score = 76.91%
	(Jiwei Li et al. 2014) Hôtels			F-score = 78.2%
	(Jiwei Li et al. 2014) Médecins			F-score = 75%
	(Jiwei Li et al. 2014) Restaurants			F-score = 81.8%

(Noekhah et al. 2020)	OpSpam	Caractéristiques basées sur le contenu, le comportement et les relations	Multi-iterative Graph-based opinion Spam Detection (MGSD)	Accuracy = 95.3%
(Z. Wang et al. 2020)	AmazonCn	Caractéristiques hétérogènes par paires	COSD	AUC = Amélioration de 12.04%
	YelpHotel			AUC = Amélioration de 12.78%
(Jiandun Li et al. 2021)	Dataset à partir de JD.com	LDA	Grouped Spam Detection approach based on Nominated Topics (GSDNT)	Accuracy = 96.42%

2.8.2 Modèles de Deep Learning

Les modèles d'apprentissage en profondeur sont principalement conçus pour apprendre automatiquement les caractéristiques les plus informatives directement à partir des données brutes, éliminant ainsi le besoin d'ingénierie manuelle des caractéristiques. Ceci est réalisé grâce à l'utilisation de plusieurs couches de transformations non linéaires, la première consistant généralement en une technique d'apprentissage de caractéristiques comme mentionné précédemment, ce qui permet au modèle de capturer des relations complexes dans les données plus efficacement que les modèles traditionnels. Cependant, il convient de mentionner que ces techniques nécessitent généralement de grandes quantités de données et des ressources importantes. Des exemples de tels modèles incluent CNN, RNN, LSTM, et GAN (Mohawesh et al. 2021b ; Ren et Ji 2019). Ces modèles ont atteint de bonnes performances sur une variété de tâches de traitement du langage naturel, y compris l'analyse des sentiments et la détection des spams d'opinion.

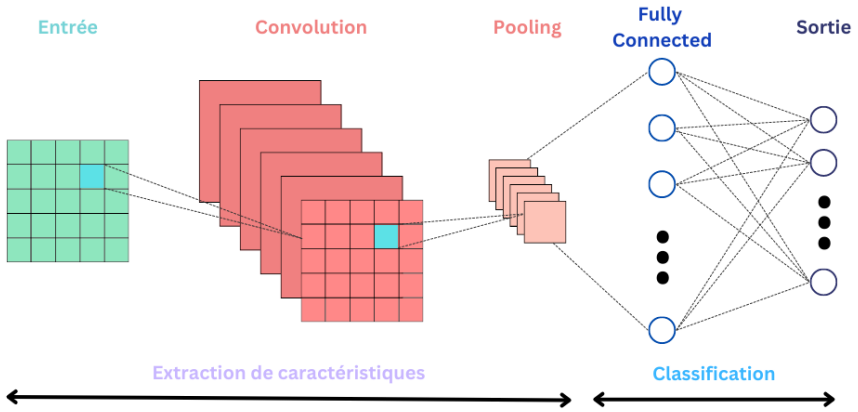


Fig. 2.13 : Architecture générale d'un réseau de neurones convolutif

En ce qui est de **CNN**, ou réseau de neurones convolutif, c'est un type de réseaux de neurones acycliques (Feed-Forward) basé sur l'idée de convolution qui est une opération

mathématique qui consiste à appliquer un filtre ou un noyau à une image d'entrée pour extraire des caractéristiques. Il est couramment utilisé dans les tâches de Computer vision telles que la classification des images et la détection d'objets, mais peut jouer un rôle important dans la capture de caractéristiques locales significatives pour la classification des tâches de traitement du langage naturel (Mohawesh et al. 2021b). L'architecture générale d'un CNN est présentée dans la figure 2.13.

Parmi les couches qui composent ce réseau de neurones, on trouve la **couche d'entrée** qui reçoit les données d'entrée et est suivie par la **couche convolutionnelle** qui applique un ensemble de filtres (appelées Noyaux ou kernels) sur l'entrée afin d'en extraire les différentes caractéristiques. Chaque filtre est une petite matrice de pondérations appliquée à une petite partie des données d'entrée. La sortie de la couche convolutionnelle est un ensemble de cartes de caractéristiques, chacune correspondant à un filtre différent. Ces cartes de caractéristiques générées deviennent ensuite l'entrée de la **couche de regroupement** (Pooling) qui vise à réduire leur taille en les sous-échantillonnant, i.e. en prenant la valeur maximale ou moyenne dans chaque région locale, dans le but de diminuer le nombre de paramètres dans le réseau pour le rendre plus efficace en termes de calcul. Parmi les méthodes de Pooling les plus utilisées, on trouve Max-pooling. Le pooling est suivi par la **couche d'activation** qui consiste en l'application d'une fonction d'activation non-linéaire (comme ReLU, Sigmoid ou la Tangente hyperbolique) sur la sortie des couches précédentes, ce qui aide à introduire la non-linéarité dans le réseau et permet d'apprendre des modèles plus complexes dans les données. Cette couche peut être appliquée avant ou après la couche de Pooling, ce choix dépend de l'application spécifique et de l'architecture du CNN. Et enfin, la dernière couche est la couche **entièrement connectée** (Fully connected) qui est une couche standard des réseaux de neurones dans laquelle chaque neurone est connecté à chaque neurone de la couche précédente afin de générer la sortie qui est un vecteur de scores de classe, qui est utilisé pour faire des prédictions ou des classifications. En plus de ces couches, un CNN peut également inclure d'autres types de couches, telles que des couches de normalisation, des couches d'abandon (dropout), etc dans le but d'améliorer ses performances et réduire le surajustement.

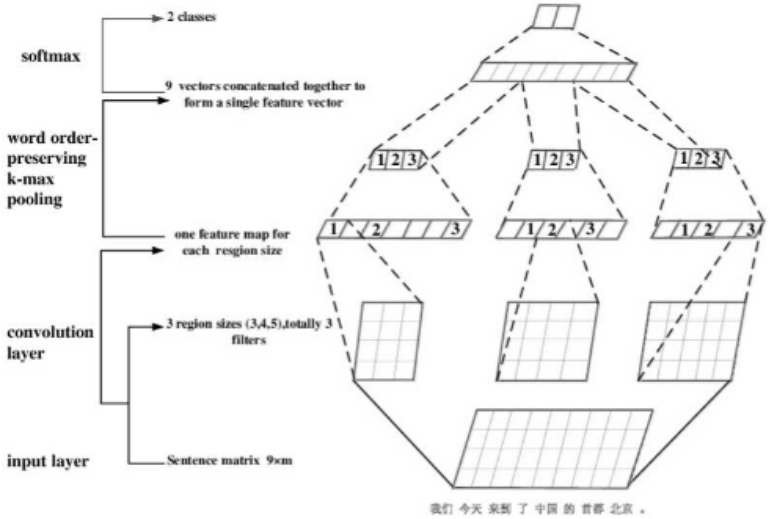


Fig. 2.14 : Architecture générale de OpCNN (Zhao et al. 2018)

De nombreux travaux ont utilisé les CNN, citons par exemple le modèle **OpCNN** introduit par (Zhao et al. 2018), qui est un modèle qui prend en considération les ca-

ractéristiques de l'ordre des mots pendant le processus d'analyse. Ce modèle CNN est composé de quatre couches, une couche d'entrée, une couche de convolution, une couche de regroupement et une couche de sortie comme l'illustre la figure 2.14.

Le modèle prend en entrée les avis avec un certain ordre de mots, puis dans les couches de convolution et de regroupement, l'ordre consécutif des mots est conservé en appliquant la méthode de pooling "word order-preserving k-max" qu'ils ont conçue. Afin d'évaluer leur modèle, ils ont construit un corpus contenant 24,166 avis sur des hôtels collectés du site chinois dianping.com puis les ont annotés en suivant la méthode de (F. H. Li et al. 2011), 4,132 de ces avis ont été annotés comme Spam. Les auteurs ont constaté une exactitude de 70.02% surpassant les performances d'autres méthodes de l'état de l'art (TF-IDF + SVM, Bi-grammes + SVM, CNN). Ils ont aussi testé sa capacité de généralisation sur le dataset OpSpam, OpCNN a obtenu de meilleurs résultats que CNN avec une exactitude de 84.50% et une mesure F-score de 82.84%.

Les **Réseaux de neurones récurrents** (RNN), de leur part, sont un type de réseau neuronal artificiel conçu pour traiter des données séquentielles en maintenant et en mettant à jour un état interne, qui sert de mémoire au réseau. Cet état permet au réseau de traiter et d'analyser des séquences d'entrées de longueur arbitraire, ce qui le rend adapté à des tâches telles que le traitement du langage naturel, la reconnaissance vocale et la prédiction de séries chronologiques. L'architecture de base d'un RNN se compose de trois composants principaux : la *couche d'entrée* qui prend une séquence d'entrées (sous forme de vecteurs), et les transmet à la couche suivante qui est la couche récurrente. La *couche récurrente* contient un ensemble d'unités récurrentes, dont chacune a un état interne qui est mis à jour en fonction de l'entrée actuelle et de l'état précédent, sa sortie est ensuite transmise à la couche finale, *la couche de sortie*, qui produit la sortie finale du réseau.

Comme nous l'avons expliqué, RNN peut utiliser les informations dans des séquences de longueurs arbitraires, mais c'est surtout en théorie. En pratique, un RNN standard se limite à ne regarder que quelques étapes en arrière en raison du problème de disparition du gradient ou de gradient explosif. Les chercheurs ont développé des types de RNN plus sophistiqués pour éviter ces défauts tels que la mémoire à long court terme (LSTM), l'unité récurrente fermée (GRU)(Cho et al. 2014), le LSTM bidirectionnel, le LSTM empilé et le LSTM avec méthode d'attention (Mohawesh et al. 2021b). **LSTM**, ou Long Short-Term Memory (Hochreiter et Schmidhuber 1997), est un type de RNN qui peut apprendre des dépendances à long terme et, contrairement à un RNN traditionnel, il dispose d'un mécanisme de déclenchement qui lui permet d'oublier ou de conserver de manière sélective les informations des étapes de temps précédentes, il a été introduit afin de résoudre le problème de disparition du gradient que nous avons mentionné ci-dessus. L'architecture de base d'un LSTM comprend une cellule mémoire (qui stocke les informations de l'instant précèdent), une porte d'entrée (qui détermine quelles informations doivent être stockées dans la cellule mémoire à l'instant actuel), une porte de sortie (qui décide quelles informations de la cellule mémoire doivent être transmises à la couche ou l'instant suivant) et la porte d'oubli (qui décide quelles informations doivent être supprimées de la cellule mémoire).**Bi-LSTM**, d'autre part, est l'abréviation de LSTM bidirectionnel, et c'est un type de LSTM qui traite les entrées dans les sens avant et arrière (Schuster et Paliwal 1997). Il se compose de deux LSTM, l'un traitant l'entrée vers l'avant et l'autre vers l'arrière. La sortie des deux LSTM est concaténée pour former la sortie finale.

(Zeng et al. 2019) ont proposé une approche qui applique un modèle Bi-LSTM afin de détecter les faux avis en fonction de la structure de l'avis. La méthode d'ensemble proposée, **RSBE**, est composée de quatre encodeurs LSTM bidirectionnels distincts, qui encodent la première phrase de l'examen, le contexte intermédiaire, la dernière phrase et le texte entier en quatre représentations de document, ainsi que deux couches de mécanismes d'attention qui intègrent les 4 représentations précédentes en une représentation finale qui est ensuite transmise à une fonction softmax comme illustré dans la figure 2.15 ci-dessous. L'intérêt pour la structure des critiques est venu de leurs conclusions selon lesquelles les fausses critiques exprimaient des émotions plus fortes que les véritables critiques, que ces émotions étaient principalement présentes dans les premières et dernières phrases et que les fausses critiques commençaient ou se terminaient par des phrases similaires. Ils ont évalué leur modèle à l'aide du corpus construit par (Jiwei Li et al. 2014) et de Word embeddings pré-entraînés sur le corpus Wikipédia. Le modèle proposé a obtenu de meilleurs résultats que d'autres méthodes de l'état de l'art dans un domaine (hôtel, médecin et restaurant) avec une précision de 85,7%, 84,7% et 85,5%, respectivement. De plus, dans le domaine mixte, il atteint une précision de 83,4%. Cependant, le modèle proposé n'a pas réussi à obtenir de bons résultats dans les domaines croisés avec seulement 71,6%.

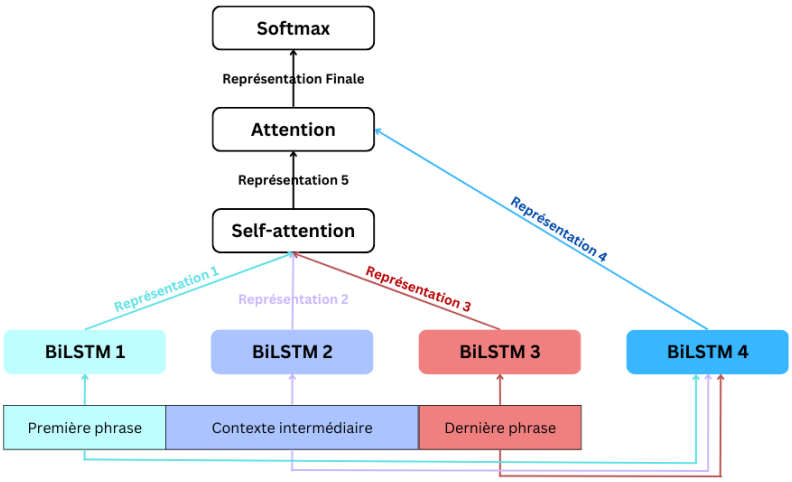


Fig. 2.15 : Architecture générale de l'approche RSBE de (Zeng et al. 2019)

En ce qui est des simples LSTMs, (Shahariar et al. 2019), dont nous avons présenté les méthodes traditionnelles utilisées précédemment, ont aussi expérimenté avec des classificateurs neuronaux, à savoir CNN et LSTM, et cela, avec des embeddings Word2vec. Les meilleurs résultats ont été obtenus avec LSTM avec une exactitude de 94.565% sur le corpus OpSpam et 96.75% sur Yelp. Plus récemment et toujours concernant les LSTMs, les auteurs de (Harris 2022) ont proposé une approche combinant les caractéristiques linguistiques et comportementales avec une architecture exploitant LSTM. Ils entraînent d'abord trois modèles LSTM : *LSTM1* utilisant 12 caractéristiques linguistiques (dont le nombre de mots, de verbes, et de mots spatio-temporels, etc), *LSTM2* utilisant 6 caractéristiques comportementales (MRD, BST,...) et *LSTM3* utilisant l'approche K-L Divergence, qui est une mesure de l'information perdue lorsqu'une distribution est estimée par une autre. Cette séparation leur a permis de configurer et d'évaluer chaque modèle LSTM séparément. Ces modèles sont ensuite combinés dans un modèle d'ensemble, comme le montre la figure 2.16, avec deux couches d'auto-attention présentes, l'auto-attention étant un type de mécanisme d'attention qui peut extraire les dépendances positionnelles (sa sor-

tie est la moyenne pondérée des différentes positions dans la séquence d'entrée) tout en nécessitant moins de paramètres et a une complexité de calcul inférieure. Afin d'évaluer leur approche, ils ont choisi les avis de YelpZIP portants sur les restaurants. Et étant donné que les données authentiques sont plus fréquentes que les fausses (l'ensemble de données est déséquilibré) ils ont utilisé une technique largement appliquée pour suréchantillonner la classe minoritaire, appelée SMOTE (Chawla et al. 2002), séparément pour les modèles linguistiques et comportementaux.

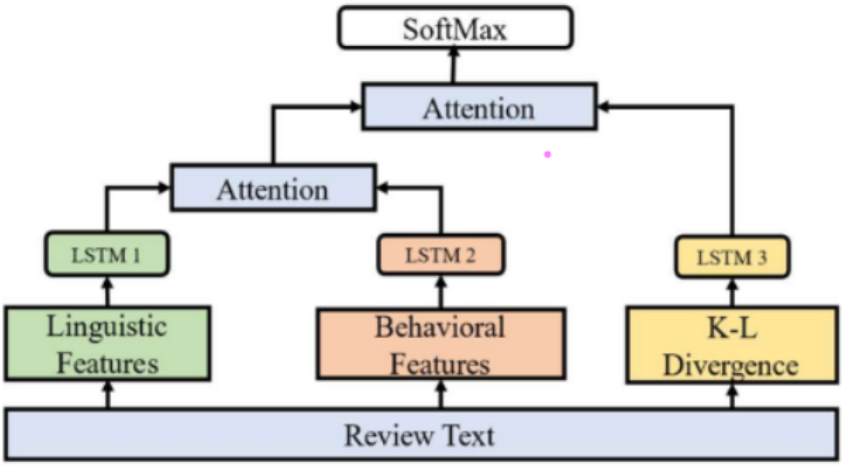


Fig. 2.16 : Architecture générale de l'approche de (Harris 2022)

En ce qui est des résultats de l'étude (Harris 2022), l'évaluation des 12 caractéristiques linguistiques montre que les avis signalés comme spam ont moins de diversité lexicale et de contenu, utilisent moins de mots spatio-temporels, ont une longueur de mot/terme légèrement plus courte et utilisent globalement moins de clauses. L'évaluation des 6 caractéristiques comportementales, quant à elle, montre que les fenêtres d'activité des spammeurs ont tendance à être beaucoup plus courtes, utilisent des contenus plus similaires, sont plus extrêmes et utilisent davantage les majuscules. Ils ont trouvé que les caractéristiques comportementales ont de meilleurs résultats que les linguistiques et que le meilleur résultat a été observé lors de la combinaison de modèles, ce modèle a surpassé chacune des 3 approches indépendantes et d'autres approches de l'état de l'art basées sur la divergence K-L appliquées sur le même ensemble de données avec une AUC de 89.30%.

Passons maintenant à **GAN**, qui est l'abréviation de "Generative Adversarial Networks". Il s'agit d'un type de modèle d'apprentissage profond qui se compose de deux réseaux de neurones, un générateur et un discriminateur, qui sont entraînés ensemble pour générer de nouvelles données similaires aux données d'entraînement. Le générateur prend un bruit d'entrée aléatoire et produit un échantillon qui ressemble aux données d'apprentissage, tandis que le discriminateur prend à la fois des échantillons réels et générés et essaie de les distinguer. Le générateur est formé pour produire des échantillons qui peuvent tromper le discriminateur, tandis que le discriminateur est formé pour identifier correctement les vrais échantillons à partir de ceux générés. Au fur et à mesure que les deux réseaux sont entraînés ensemble, ils se font concurrence et s'améliorent dans leurs tâches respectives. Finalement, le générateur est capable de produire des échantillons très similaires aux données d'apprentissage, et le discriminateur n'est pas capable de les distinguer des données réelles.

(Tang et al. 2020) se sont intéressés au manque de caractéristiques comportementales

efficaces pour les nouveaux utilisateurs qui ne publient qu'un seul avis. Afin de gérer le problème des démarrages à froid (ou des spammeurs singleton), ils ont exploité le réseau GAN, l'idée de base est de générer des caractéristiques de comportement synthétique (SBFs) pour les nouveaux utilisateurs à partir de leurs caractéristiques facilement accessibles (EAFs). D'abord, 6 caractéristiques réelles (RBFs) ont été extraites pour les utilisateurs réguliers existants (Activity Window - Maximum Number of Reviews - Percentage of Positive Reviews - Review Count - Reviewer Deviation - Maximum Content Similarity) ainsi que 3 caractéristiques facilement accessibles pour les nouveaux utilisateurs : des caractéristiques textuelles (TF) via des emmbeddings à 100 dimensions pré-formés à l'aide de CNN (puisqu'un évaluateur doit publier au moins un avis), des caractéristiques d'évaluation (RF) via un vecteur à 100 dimensions représentant l'écart de note (puisqu'un évaluateur doit évaluer au moins un produit) et des caractéristiques d'aspect (AF) via un vecteur à 100 dimensions représentant la déviation temporelle (puisqu'un utilisateur doit avoir des horodatages d'inscription et de publication). En utilisant ces caractéristiques, ils entraînent le modèle GAN, y compris un *générateur* utilisé pour apprendre le mappage des EAFs d'entrée aux SBFs sous la contrainte de maintenir les SBFs proches des RBFs et un *discriminateur* qui consomme les SBFs du générateur, ainsi que les EAFs et les RBFs à partir des données de formation en entrée et vise à distinguer les caractéristiques de comportement synthétiques et réelles. L'architecture globale de cette approche peut être visualisée dans la figure 2.17.

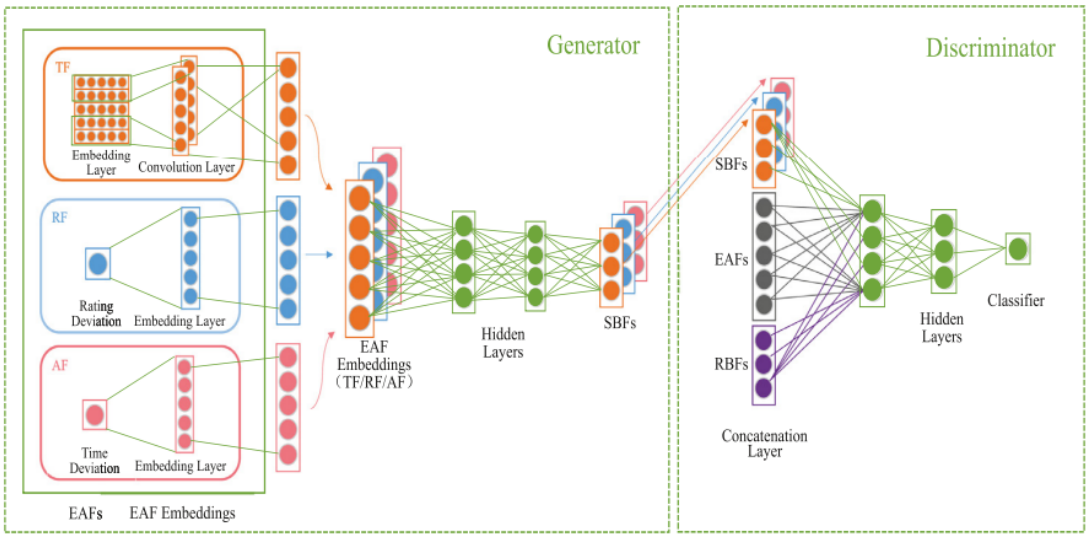


Fig. 2.17 : Architecture générale de l'approche de (Tang et al. 2020)

Le modèle proposé a été évalué sur deux sous-ensembles de YelpCHI, Hôtel et Restaurant, et a surpassé certaines méthodes de pointe avec une exactitude de 83% sur le domaine hôtelier et 75,7% sur le domaine restaurant. Les caractéristiques combinées ont amélioré les performances du modèle de classification.

À la suite de cela, (Kennedy et al. 2020) ont analysé et comparé plusieurs méthodes neuronales et traditionnelles afin de distinguer entre les avis véridiques et spam ainsi que différentes caractéristiques. Les auteurs de cet article ont expérimenté sur Yelp et OpSpam, et ont proposé les caractéristiques suivantes : Bag-Of-Words, Word2vec pré-entraîné avec une dimensionnalité de 300 sur un dataset de Google Actualités[11] ainsi que des ca-

[11] https://code.google.com/archive/p/word2vec/

ractéristiques structurelles (Longueur de l'avis, longueur moyenne des mots et des phrases, pourcentage de mots en majuscules et pourcentage de chiffres), des caractéristiques comportementales (nombre maximal d'avis en une journée, durée moyenne des avis, écart type des notes et pourcentage de notes positives et négatives), des pourcentages Part-Of-Speech et des pourcentages de polarité pour chaque mot. Cependant, après une sélection de caractéristiques à l'aide de la régression logistique, il s'est avéré que ces deux dernières caractéristiques n'étaient pas prédicatives et que de meilleures performances peuvent être obtenues en combinant les caractéristiques comportementales et BOW. À noter que les caractéristiques comportementales ne peuvent être exploitées qu'avec le dataset Yelp, et que pour la combinaison de ces caractéristiques avec BOW, elles sont directement concaténées. En ce qui est des modèles utilisés, on trouve FFNNs (Réseaux de neurones Feed-Forward), CNNs, LSTMs ainsi qu'une version affinée de BERT (bert-base-uncased). Les auteurs précisent dans leur article les spécifications techniques de chacun de ces modèles pour chacun des datasets et pour chaque ensemble de caractéristiques. En comparant les performances des classificateurs FFNN, CNN et LSTM, les meilleures performances ont été données par FFNN et BOW sur OpSpam (une exactitude de 88,8%), suivis de LSTM et BOW (exactitude = 87,6%). En revanche, sur Yelp, LSTM+BOW et CNN+Word2vec ont donné les meilleurs résultats avec 73,1%. Les meilleures performances sur l'ensemble de données OpSpam, qui sont compétitifs par rapport aux techniques de l'état de l'art, sont obtenues en affinant BERT avec une précision de 90,5% (en utilisant l'implémentation TensorFlow).

La dernière étude que nous allons présenter est l'approche appelée **EUPHORIA**, "nEural mUlti-view aPproacH fOr RevIew spAm" (Andresini et al. 2022) qui est une approche qui utilise l'apprentissage multi-vues en exploitant plusieurs vues (chevauchées) des données afin que le modèle de classification ne manque aucun aspect important. Dans cette méthode, chaque vue de données est représentée par un vecteur de caractéristiques distinct qui est ensuite traité par le réseau neuronal à entrées multiples. Actuellement, le modèle prend en charge trois vues de données différentes et en accepte d'autres. Deux de ces vues sont liées aux caractéristiques textuelles et utilisent Word2Vec (CBOW) pour une vue spécifique au domaine du texte de révision, et BERT pour une vue plus générale. La troisième vue, quant à elle, comporte 6 caractéristiques comportementales qui représentent le profil de l'examinateur au fil du temps (nombre maximal d'avis par jour (MRD), rapport positif (PR), durée moyenne de l'examen (ARL), écart de l'examinateur (RD), similarité moyenne de l'examen (ARS), similarité maximale des avis (MRS)).

Comme l'illustre la figure 2.18, après l'extraction des caractéristiques, les vecteurs d'intégration et les caractéristiques comportementales sont transmis à un réseau de neurones à entrées multiples. Comme indiqué, la première couche du réseau est constituée de 3 couches distinctes entièrement connectées avec dropout, et crée une vue compressée et spécifique à la tâche de chaque vecteur de caractéristiques. Les trois vues sont concaténées ensemble, puis transmises par la couche de classification. L'utilisation d'un réseau de neurones à entrées multiples s'est avérée, après les expérimentations présentées, aider réellement à bénéficier de la richesse des données multi-vues par rapport au réseau de neurones à entrée unique dont les performances sont dégradées par les effets de la malédiction de la dimensionnalité.

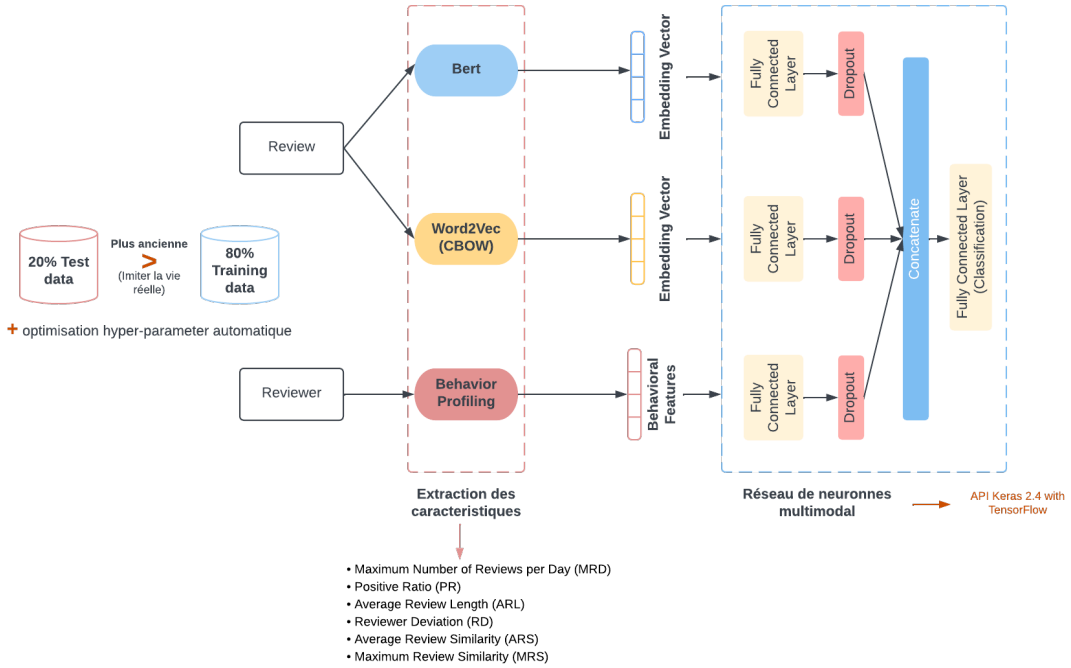


Fig. 2.18 : Architecture générale de l'approche de (Andresini et al. 2022)

Les données utilisées dans cette étude ont été extraites de deux sous ensembles de YelpCHI, à savoir Hôtel and Restaurant, contenant des avis sur 85 hôtels et 130 restaurants, respectivement, dans la région de Chicago. Le cadre expérimental adopté dans cette étude est différent de celui couramment adopté dans la littérature de la détection de spam, où les divisions des datasets Entraînement-Test sont générées de manière aléatoire, négligeant ainsi la date/heure de publication des avis. Afin d'adapter les expériences à un scénario réaliste où le modèle ne serait pas en mesure d'utiliser les informations produites après la rédaction de l'avis, dans les ensembles de validation et de test, les valeurs des caractéristiques comportementales pour chaque opinion sont mises à jour en fonction de l'ordre dans lequel ils ont été écrits. Une autre particularité de la méthode proposée est que même si les deux ensembles de données sont classifiés séparément, car tout domaine a ses caractéristiques spécifiques, le classifieur considère également les connaissances hors domaine en tenant compte des critiques écrites par le même auteur concernant d'autres domaines, plus tard leurs expériences ont montré que c'était un pas dans la bonne voie, car l'acquisition de plus de connaissances a reflété un gain de performances.

Afin de savoir dans quelle mesure chaque point de vue individuel influence la précision du modèle de classification, les auteurs ont analysé comment les connaissances contenues dans les vues textuelles et de comportement peuvent influencer les performances d'EUPHORIA. Les résultats ont montré que les caractéristiques comportementales transmettent les informations les plus pertinentes nécessaires pour détecter les faux avis surpassant à la fois Word2Vec et Bert et leur combinaison également. D'autre part, il a été démontré que les caractéristiques de contenu extraites conjointement par Word2Vec et BERT donnaient de meilleurs résultats que de les traiter séparément. Le principal résultat de cette analyse a montré que la combinaison des trois vues surpasse toutes les vues précédentes et, par conséquent, Euphoria a obtenu la meilleure performance avec une AUC de 70,8% sur le corpus de restaurants et de 81,3% sur le corpus d'hôtels. Euphoria a aussi eu de meilleurs résultats par rapport à deux algorithmes de l'état de l'art basés sur SVM, dans le premier, les vecteurs de caractéristiques ont d'abord été concaténés en

un seul vecteur d'entrée, puis traités comme par le classificateur SVM tandis que dans le second (noté Ens-SVM) un ensemble de trois SVM distincts sont formés séparément des trois vecteurs de caractéristiques, puis la règle de majorité d'ensemble est utilisée pour la classification finale. Les résultats ont montré que l'apprentissage de classificateurs séparés pour chaque vue est moins performant que l'apprentissage d'un seul classificateur en concaténant toutes les vues ensemble. Cela ne pose pas de problème en ce qui concerne Euphoria en raison de la rétro-propagation dans le réseau neuronal à entrées multiples, contrairement à Ens-SVM où les informations ne peuvent pas être partagées dans ce cadre.

Le tableau 2.4 résume certains travaux exploitant les modèles Deep Learning pour le problème de détection de spam d'opinion, dont certains que nous avons expliqués précédemment et d'autres non.

Tab. 2.4 : Travaux exploitant les modèles Deep Learning pour le problème de détection de spam d'opinion

Référence	Dataset	Caractéristiques	Classifieur	Résultats
(Shahariar et al. 2019)	OpSpam	Word2vec	LSTM	Accuracy = 94.565%
	Yelp			Accuracy = 96.75%
(Zeng et al. 2019)	(Jiwei Li et al. 2014) [Hôtel]	Word embeddings pré-entraînés sur le corpus Wikipédia	RSBE : LSTM bidirectionnel avec un mécanisme d'auto-attention	Même domaine, Accuracy = 85.7%
	(Jiwei Li et al. 2014) [Restaurant]			Même domaine, Accuracy = 85.5%
				Inter-domaines, Accuracy = 71.6%
	(Jiwei Li et al. 2014) [Médecin]			Même domaine, Accuracy = 84.7%
				Inter-domaines, Accuracy = 60.5%
	(Jiwei Li et al. 2014)			Domaine Mixte, Accuracy = 83.4%
(You et al. 2020)	TripAdvisor	Évaluation de l'aspect + TF-IDF	AR-LOF+ : Aspect-Rating Local Outlier Factor Reinforced algorithm	F-score = 79.8%

(M. Liu et al. 2020)	YelpCHI	Caractéristiques liées à l'aspect à l'aide d'un modèle de fusion au niveau du mot	MIANA : Réseau neuronal multiniveau basé sur l'attention	AUC = 86.65%
	YelpNYC			AUC = 91.89%
	YelpZIP			AUC = 93.26%
(Tang et al. 2020)	YelpCHI	6 Caractéristiques de comportement réel (RBF) + Caractéristiques facilement accessibles (EAF)	bfGAN : Behaviour featuresgenerative adversarial network	[Hôtel] Accuracy = 83%
				[Restaurant] Accuracy = 75.7%
(Cao et al. 2020)	YelpCHI	LDA-BP +TextCNN + Word2vec	Deceptive reviews detection frameword using Coarse and Fine-grained fusion	Accuracy = 84.5%
	(Jiwei Li et al. 2014) [Hôtel]			Accuracy = 85.9%
	(Jiwei Li et al. 2014) [Restaurant]			Accuracy = 81.5%
	(Jiwei Li et al. 2014) [Médecin]			Accuracy = 82.7%
(Fahfouh et al. 2020)	OpSpam	BOW distribué par vecteur de paragraphe (PV-DBOW) + Auto-encodeur de débruitage (DAE)	PV-DAE : Modèle d'apprentissage profond hybride	Accuracy = 92.5%
(Mohawesh et al. 2021b)	OpSpam	RoBERTa + XLNet + ALBERT	Approche d'ensemble combinant les 3 architectures Transformer	Accuracy = 94.08%
	Deception dataset			Accuracy = 92.07%

(Harris 2022)	YelpZIP	12 Caractéristiques linguistiques (Nombre de mots, Nombre de verbes, Nombre de mots spatio-temporels ,...) + 6 Caractéristiques comportementales (MRD, BST,…)	Combinaison de 3 modèles (Linguistique, Comportemental et Divergence K-L) dans un réseau LSTM	AUC = 89.30%
(Andresini et al. 2022)	Sous-ensemble de YelpCHI [Restaurant]	BERT - Word2vec - 6 Caractéristiques comportementales (MRD, PR, ARL, RD, ARS, MRS)	EUPHORIA : Réseau de neurones multi-entrées	AUC = 70.8%
	Sous-ensemble de YelpCHI [Hôtels]			AUC = 81.3%

2.9 Métriques et méthodes d'évaluation

Puisque le problème de la détection de spam d'opinions est considéré comme un problème de classification binaire, les chercheurs évaluent les performances de leurs approches en utilisant les métriques utilisées pour l'évaluation des différents modèles de classification citées dans la littérature. Il existe deux moyens d'évaluer un modèle de classification, en utilisant des métriques d'évaluation ou à l'aide d'une approche graphique (Rastogi et al. 2020).

En ce qui est de la première approche, le tableau 2.5 illustre les métriques les plus utilisées dans les travaux de la recherche, parmi ces métriques, nous retrouvons la précision P, le rappel R et le F-score que nous avons déjà abordé dans la section 1.7 du chapitre précèdent. Le tableau 2.6, représente une table de confusion adaptée au problème de détection de spam.

Quant à la deuxième approche, nous citerons deux courbes, la première étant celle du ROC_AUC que nous avions déjà expliquée dans la section 1.7 du chapitre précédent, et la seconde étant la courbe précision-rappel (ou Courbe PR) qui est un tracé qui montre le compromis entre la précision (axe des ordonnées) et le rappel (axe des abscisses) pour différents seuils de probabilité comme illustré à travers la figure 2.19. Une seule mesure, l'aire sous la courbe (AUC), est utilisée pour résumer les résultats des courbes ROC et PR (Rastogi et al. 2020).

Tab. 2.5 : Exemples de métriques d'évaluation des méthodes de détection de spam d'opinion

Métriques	Formules	Objectifs de l'évaluation
Exactitude	$\frac{TP+TN}{TP+FP+TN+FN}$	Mesure la proportion des instances correctement classées (à ne pas utiliser avec les datasets) datasets déséquilibrés)
Précision (P)	$\frac{TP}{TP+FP}$	Mesure la proportion des prédictions correctes dans la classe positive (Spam)
Rappel (R)	$\frac{TP}{TP+TN}$	Mesure la proportion des instances positives (spam) correctement classées
F-score	$2\frac{PR}{P+R}$	Représente la moyenne harmonique entre la précision et le rappel
G-mean	$\sqrt{TPTN}$	Vise à maximiser TP et TN, tout en maintenant les deux valeurs relativement équilibrées

Tab. 2.6 : Table de confusion - Détection de spam

Classe Réelle / Classe Prédite	Spam	Non Spam
Spam	TP (Vrai Positif)	FN (Faux Négatif)
Non Spam	FP (Faux Positif)	TN (Vrai Négatif)

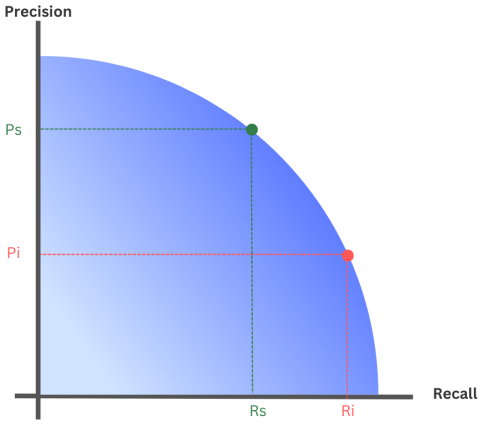


Fig. 2.19 : La courbe Précision-Rappel

2.10 Défis et discussions

Comme présenté précédemment, les études antérieures ont construit des bases solides pour résoudre le problème de détection d'opinions frauduleuses. Cependant, il reste de nombreux défis à surmonter dans ce domaine et comme l'a souligné (Meel et Vishwakarma 2020), ces challenges peuvent être considérés comme de nouvelles opportunités de recherche. Dans ce qui suit, nous présenterons certains de ces défis selon la phase du processus affectée :

2.10.1 L'acquisition des données :

D'après la littérature et ce que nous avons discuté, on constate que la plupart des datasets utilisés dans les travaux dans ce domaine sont créés de manière synthétique puisque l'acquisition d'une large quantité de données non étiquetées est facile, cependant la collecte de données étiquetées est difficile et coûteuse (Ma et F. Li 2012). Cependant, ces données ne sont pas nécessairement représentatives du spam d'opinion du monde réel. D'où, l'établissement de datasets de référence de qualité est une tache de haute importance afin d'assurer la qualité des modèles de détection de spam entraînés sur ces données (Meel et Vishwakarma 2020).

En plus de cela, il peut s'avérer intéressant de comparer les opinions et les auteurs sur plusieurs sites. Comme l'indique (B. Liu 2020), cette comparaison pourrait nous aider à découvrir des anomalies telles que des avis similaires publiés sur plusieurs plates-formes, à peu près au même moment, avec des adresses IP et des ID utilisateur similaires, un avis comme celui-ci est peu susceptible d'être authentique par exemple, pourquoi un utilisateur prendrait-il la peine de publier le même avis positif sur un produit sur plusieurs sites ? ne serait-il pas embêtant de s'inscrire sur plusieurs plateformes pour publier cette critique si ce n'est pour un certain gain ? Cependant, la collecte de toutes ces données à travers plusieurs sites, leur analyse et leur comparaison peut être difficile et coûteux, surtout avec la quantité immense de données présentes sur les réseaux qui ne cesse d'augmenter de jour en jour.

Le dernier point que nous voudrions mentionner concernant les défis en matière de données spam serait leur détection. Comme les modèles mettent un certain temps à se former dans les comportements des spammeurs et les spams textuels, il est difficile de détecter les avis trompeurs rapidement, si nous attendons trop longtemps, il pourrait être trop tard, car le mal serait déjà fait (B. Liu 2020). C'est pourquoi que les algorithmes de détection doivent être conçus pour réduire le temps nécessaire pour identifier efficacement le spam.

2.10.2 Les caractéristiques :

D'autant plus, du point de vue des caractéristiques (features), déterminer la meilleure combinaison de caractéristiques décrivant au mieux un contenu spam ou l'attitude d'un spammeur est aussi un défi majeur dans la détection d'opinions frauduleuses. La difficulté de choisir ces caractéristiques est qu'il est également difficile pour les humains de distinguer si un avis est un spam ou non vu que les spammeurs fournissent des efforts considérables pour que leur style de rédaction ressemble le plus possible à celles d'un être humain qui a vécu l'expérience avec le produit. Nous pouvons aussi remarquer que la plupart des travaux réalisés se basent sur les caractéristiques textuelles et non pas comportementales, or certaines études montrent qu'une combinaison des deux donnent de meilleurs résultats (Andresini et al. 2022).

Un autre défi critique qui pourrait entraver la pertinence des caractéristiques, en particulier les caractéristiques comportementales, est l'existence des opinions spam singletons (Aliarab et Fouladi 2022). Une opinion spam singleton est un commentaire dont l'auteur n'en a publié qu'un seul, il n'a donc pas laissé beaucoup d'informations utiles pour modéliser. Dans la plupart des études, ce type de commentaire est supprimé de

l'ensemble de données, seuls quelques-uns d'entre eux ont essayé de résoudre ce problème, comme (Shaalan et al. 2021) par exemple qui ont utilisé des schémas d'aspect-sentiment temporels anormaux profonds et des auto-encodeurs. Sans oublier le problème de dimensionnalité qui, comme l'affirme (Andresini et al. 2022), détériore les performances des modèles d'apprentissage avec chaque caractéristique considérée.

2.10.3 Les modèles de classification :

D'après (Meel et Vishwakarma 2020), les travaux présents dans la littérature sont principalement concentrés sur les techniques d'apprentissage supervisé, or ces dernières sont dépendantes des ensembles de données étiquetées et, comme nous l'avons discuté dans les points précédents, peuvent être difficile, voire impossible, à acquérir. Pour cela, il serait plus intéressant de développer des approches d'apprentissage non supervisé.

N'oublions pas que puisque la polarité d'une opinion peut être une caractéristique utile à exploiter dans le processus de détection de spam, tout défi rencontré durant l'analyse des sentiments (Section 1.8) est un défi pour la détection de spam d'opinion.

2.11 Conclusion

Au cours de ce chapitre, nous avons exploré les différents travaux de l'état de l'art traitants le problème de l'identification de la crédibilité d'une opinion. Nous avons d'abord commencé par définir quelques généralités sur le domaine pour une meilleure compréhension. Puis, nous avons présenté le processus de détection de spam d'opinion et ses différentes étapes en passant par les différents jeux de données disponibles et les métriques d'évaluation utilisées. Enfin, nous avons évoqué les différents défis présents et que les prochaines contributions doivent relever afin d'assurer la crédibilité de l'opinion.

Conclusion et perspectives

Conclusion générale

Dans ce rapport, nous nous sommes intéressées aux méthodes présentes dans la littérature qui répondent à la problématique d'analyse des sentiments et celle de détection de spam d'opinion, pour cela, nous l'avons divisé en deux chapitres respectivement. Le premier chapitre porte sur la classification des opinions selon le critère de polarité, nous y présentons les approches abordées pour effectuer cette classification, ainsi que les différents concepts nécessaires à la compréhension du domaine, les méthodes d'évaluation et les défis rencontrés. Le second chapitre aborde, quant à lui, la classification des avis selon le critère de crédibilité. Nous avons d'abord défini ce qui est considéré comme crédible dans ce domaine avant de présenter les techniques utilisées pour identifier les opinions frauduleuses, les métriques d'évaluation et les défis rencontrés. Cette recherche documentaire nous a permis d'élaborer un état de l'art qui montre l'avancement des recherches pour remédier aux problématiques citées et les défis rencontrés afin d'établir une base solide pour garantir une bonne contribution au domaine.

Perspectives

À travers cette étude de l'état de l'art, nous avons pu mettre en avant les avancées récentes, ainsi que les obstacles et les perspectives de développement pour l'analyse des sentiments et la détection de spam d'opinion.

- Les résultats obtenus par les méthodes de classification des opinions selon la polarité montrent qu'il y a encore des limites à l'utilisation de ces techniques dans des contextes réels. Ainsi, les perspectives futures peuvent inclure la recherche de nouvelles approches pour améliorer les performances de la classification et l'adaptation de ces techniques à des langues autres que l'anglais. De plus, l'application de ces méthodes dans des contextes multi-domaines peut également être explorée.

- Pour la détection de spam d'opinion, bien que des progrès significatifs aient été réalisés, il reste encore beaucoup à faire pour atteindre des résultats optimaux dans des contextes réels. Les perspectives futures peuvent inclure la recherche de nouvelles approches pour détecter des opinions frauduleuses plus sophistiquées, à savoir l'apprentissage continu (online learning) et par transfert (Andresini et al. 2022), et l'application de ces méthodes à des plateformes de médias sociaux en temps réel.

Enfin, des études supplémentaires peuvent être menées pour mieux comprendre les biais et les limites de ces techniques, en particulier en ce qui concerne les problèmes de représentativité des données et les biais culturels. La collaboration avec des experts du domaine peut aider à mieux comprendre ces questions et à développer des solutions plus adaptées.

Bibliographie

Abdaoui, Amine et al. (2021). « Dziribert : a pre-trained language model for the algerian dialect ». In : *arXiv preprint arXiv :2109.12346*.

Ahmed, Hadeer (2017). « Detecting opinion spam and fake news using n-gram analysis and semantic similarity ». Thesis.

Alhaj, Yousif A et al. (2022). « A novel text classification technique using improved particle swarm optimization : A case study of Arabic language ». In : *Future Internet* 14.7, p. 194.

Aliarab, Mahmoud et Kazim Fouladi (2022). « A survey on review spam detection methods using deep learning approach ». In : *International Journal of Web Research* 5.1, p. 19-24.

Alsubari, Saleh Nagi, Mahesh B Shelke et Sachin N Deshmukh (2020). « Fake reviews identification based on deep computational linguistic ». In : *International Journal of Advanced Science and Technology* 29.8s, p. 3846-3856.

Aly, Mohamed et Amir Atiya (2013). « Labr : A large scale arabic book reviews dataset ». In : *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 494-498.

Aminuddin, Raihah et al. (2021). « Sentiment analysis of online product reviews using Lexical Semantic Corpus-Based technique ». In : *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. IEEE, p. 233-238.

Andresini, Giuseppina et al. (2022). « EUPHORIA : A neural multi-view approach to combine content and behavioral features in review spam detection ». In : *Journal of Computational Mathematics and Data Science* 3, p. 100036.

Baccianella, Stefano, Andrea Esuli et Fabrizio Sebastiani (mai 2010). « SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining ». In : *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta : European Language Resources Association (ELRA).

Baghaei, Kourosh T et al. (2022). « Deep representation learning : Fundamentals, Perspectives, Applications, and Open Challenges ». In : *arXiv preprint arXiv :2211.14732*.

Becchetti, Luca et al. (2008). « Link analysis for web spam detection ». In : *ACM Transactions on the Web (TWEB)* 2.1, p. 1-42.

Bengio, Yoshua, Aaron Courville et Pascal Vincent (2013). « Representation learning : A review and new perspectives ». In : *IEEE transactions on pattern analysis and machine intelligence* 35.8, p. 1798-1828.

Bennett, Kristin et Ayhan Demiriz (1998). « Semi-supervised support vector machines ». In : *Advances in Neural Information processing systems* 11.

Blum, Avrim et Tom Mitchell (1998). « Combining labeled and unlabeled data with co-training ». In : *Proceedings of the eleventh annual conference on Computational learning theory*, p. 92-100.

Bojanowski, Piotr et al. (2017). « Enriching word vectors with subword information ». In : *Transactions of the association for computational linguistics* 5, p. 135-146.

Brownlee, Jason (2014). « Discover feature engineering, how to engineer features and how to get good at it ». In : *Machine Learning Process*.

Cao, Ning et al. (2020). « A deceptive review detection framework : Combination of coarse and fine-grained features ». In : *Expert Systems with Applications* 156, p. 113465.

Castillo, Carlos, Marcelo Mendoza et Barbara Poblete (2011). « Information credibility on twitter ». In : *Proceedings of the 20th international conference on World wide web*, p. 675-684.

Chaturvedi, Iti et al. (2018). « Distinguishing between facts and opinions for sentiment analysis : Survey and challenges ». In : *Information Fusion* 44, p. 65-77.

Chawla, Nitesh V et al. (2002). « SMOTE : synthetic minority over-sampling technique ». In : *Journal of artificial intelligence research* 16, p. 321-357.

Cho, Kyunghyun et al. (2014). « Learning phrase representations using RNN encoder-decoder for statistical machine translation ». In : *arXiv preprint arXiv :1406.1078*.

Comon, Pierre (1994). « Independent component analysis, a new concept ? » In : *Signal processing* 36.3, p. 287-314.

Dai, Yong et al. (2021). « Unsupervised sentiment analysis by transferring multi-source knowledge ». In : *Cognitive Computation* 13, p. 1185-1197.

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.

Deerwester, Scott et al. (1990). « Indexing by latent semantic analysis ». In : *Journal of the American society for information science* 41.6, p. 391-407.

Devlin, Jacob et al. (2018). « Bert : Pre-training of deep bidirectional transformers for language understanding ». In : *arXiv preprint arXiv :1810.04805*.

Ding, Xiaowen et Bing Liu (s. d.). « The utility of linguistic rules in opinion mining ». In : *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 811-812.

Dong, Manqing et al. (2020). « Opinion fraud detection via neural autoencoder decision forest ». In : *Pattern Recognition Letters* 132, p. 21-29.

Etaiwi, Wael et Ghazi Naymat (2017). « The impact of applying different preprocessing steps on review spam detection ». In : *Procedia computer science* 113, p. 273-279.

Fahfouh, Anass et al. (2020). « PV-DAE : A hybrid model for deceptive opinion spam based on neural network architectures ». In : *Expert Systems with Applications* 157, p. 113517.

Fast, Ethan, Binbin Chen et Michael S Bernstein (2016). « Empath : Understanding topic signals in large-scale text ». In : *Proceedings of the 2016 CHI conference on human factors in computing systems*, p. 4647-4657.

Fusilier, Donato Hernández et al. (2015). « Detecting positive and negative deceptive opinions using PU-learning ». In : *Information processing management* 51.4, p. 433-443.

Ghelani, Pooja H et Tosal M Bhalodia (2017). « Opinion mining and opinion spam detection ». In : *International Research Journal of Engineering and Technology (IRJET)* 4, p. 11.

Go, Alec, Richa Bhayani et Lei Huang (2009). « Twitter sentiment classification using distant supervision ». In : *CS224N project report, Stanford* 1.12, p. 2009.

Goel, Anurag et al. (2021). « Classification of Positive and Negative Fake Online Reviews using Machine Learning Techniques ». In : *International Journal of Advanced Networking and Applications* 12.6, p. 4746-4749.

Golub, GH et CF Van Loan (1996). « Matrix computations 3rd edition the john hopkins university press ». In : *Baltimore, MD*.

Hajek, Petr, Aliaksandr Barushka et Michal Munk (2020). « Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining ». In : *Neural Computing and Applications* 32, p. 17259-17274.

Harris, Christopher G (2022). « Combining Linguistic and Behavioral Clues to Detect Spam in Online Reviews ». In : *2022 IEEE International Conference on e-Business Engineering (ICEBE)*. IEEE, p. 38-44.

Hassan, Rakibul et Md Rabiul Islam (2021). « Impact of sentiment analysis in fake online review detection ». In : *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. IEEE, p. 21-24.

Hercig, Tomás et Ladislav Lenc (2017). « The Impact of Figurative Language on Sentiment Analysis. » In : *RANLP*, p. 301-308.

Hochreiter, Sepp et Jürgen Schmidhuber (1997). « Long short-term memory ». In : *Neural computation* 9.8, p. 1735-1780.

Hossen, Md Sharif et Niloy Ranjan Dev (2021). « An improved lexicon based model for efficient sentiment analysis on movie review data ». In : *Wireless Personal Communications* 120, p. 535-544.

Howard, Jeremy et Sebastian Ruder (2018). « Universal language model fine-tuning for text classification ». In : *arXiv preprint arXiv :1801.06146*.

Hussain, Naveed et al. (2019). « Spam review detection techniques : A systematic literature review ». In : *Applied Sciences* 9.5, p. 987.

Hussein, Doaa Mohey El-Din Mohamed (2018). « A survey on sentiment analysis challenges ». In : *Journal of King Saud University-Engineering Sciences* 30.4, p. 330-338.

Jindal, Nitin et Bing Liu (2006a). « Identifying comparative sentences in text documents ». In : *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 244-251.

— (2006b). « Mining comparative sentences and relations ». In : *Aaai*. T. 22. 13311336, p. 9.

— (2008). « Opinion spam and analysis ». In : *Proceedings of the 2008 international conference on web search and data mining*, p. 219-230.

Joyce, James (2003). « Bayes' theorem ». In.

Kennedy, Stefan et al. (2020). « Fact or factitious ? Contextualized opinion spam detection ». In : *arXiv preprint arXiv :2010.15296*.

Kim, Seongsoon et al. (2015). « Deep semantic frame-based deceptive opinion spam analysis ». In : *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, p. 1131-1140.

Le, Quoc et Tomas Mikolov (2014). « Distributed representations of sentences and documents ». In : *International conference on machine learning*. PMLR, p. 1188-1196.

Li, Fangtao Huang et al. (2011). « Learning to identify review spam ». In : *Twenty-second international joint conference on artificial intelligence*.

Li, Huayi et al. (2014). « Spotting fake reviews via collective positive-unlabeled learning ». In : *2014 IEEE international conference on data mining*. IEEE, p. 899-904.

Li, Jiandun et al. (2021). « Exploring groups of opinion spam using sentiment analysis guided by nominated topics ». In : *Expert Systems with Applications* 171, p. 114585.

Li, Jiwei et al. (2014). « Towards a general rule for identifying deceptive opinion spam ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1566-1576.

Li, Luyang et al. (2015). « Learning document representation for deceptive opinion spam detection ». In : *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data : 14th China National Conference, CCL 2015 and Third International Symposium, NLP-NABD 2015, Guangzhou, China, November 13-14, 2015, Proceedings 14*. Springer, p. 393-404.

Li, Ning, Chi-Yin Chow et Jia-Dong Zhang (2020). « SEML : A semi-supervised multi-task learning framework for aspect-based sentiment analysis ». In : *IEEE Access* 8, p. 189287-189297.

Li, Ruohan et Ayoung Suh (2015). « Factors influencing information credibility on social media platforms : Evidence from Facebook pages ». In : *Procedia computer science* 72, p. 314-328.

Li, Xiao-Li et Bing Liu (2005). « Learning from positive and unlabeled examples with different data distributions ». In : *Machine Learning : ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16*. Springer, p. 218-229.

Ligthart, Alexander, Cagatay Catal et Bedir Tekinerdogan (2021a). « Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification ». In : *Applied Soft Computing* 101, p. 107023.

— (2021b). « Systematic reviews in sentiment analysis : a tertiary study ». In : *Artificial Intelligence Review* 54.7, p. 4997-5053.

Liu, Bing (2006). *Web data mining : exploring hyperlinks, contents, and usage data*. Springer.

— (2011). *Web data mining : exploring hyperlinks, contents, and usage data*. T. 1. Springer.

— (2012). « Sentiment analysis and opinion mining ». In : *Synthesis lectures on human language technologies* 5.1, p. 1-167.

— (2020). « Sentiment Analysis ». In : *Sentiment Analysis : Mining Opinions, Sentiments, and Emotions*. 2e éd. Studies in Natural Language Processing. Cambridge University Press, p. i-i.

Liu, Bing et al. (2002). « Partially supervised classification of text documents ». In : *ICML*. T. 2. 485. Sydney, NSW, p. 387-394.

Liu, Meiling et al. (2020). « Detecting fake reviews using multidimensional representations with fine-grained aspects plan ». In : *IEEE Access* 9, p. 3765-3773.

Ma, Yingying et Fengjun Li (2012). « Detecting review spam : Challenges and opportunities ». In : *8th International Conference on Collaborative Computing : Networking, Applications and Worksharing (CollaborateCom)*. IEEE, p. 651-654.

Maas, Andrew et al. (2011). « Learning word vectors for sentiment analysis ». In : *Proceedings of the 49th annual meeting of the association for computational linguistics : Human language technologies*, p. 142-150.

Maynard, Diana G et Mark A Greenwood (2014). « Who cares about sarcastic tweets ? investigating the impact of sarcasm on sentiment analysis ». In : *Lrec 2014 proceedings*. ELRA.

Meel, Priyanka et Dinesh Kumar Vishwakarma (2020). « Fake news, rumor, information pollution in social media and web : A contemporary survey of state-of-the-arts, challenges and opportunities ». In : *Expert Systems with Applications* 153, p. 112986.

Mehta, Pooja et Sharnil Pandya (2020). « A review on sentiment analysis methodologies, practices and applications ». In : *International Journal of Scientific and Technology Research* 9.2, p. 601-609.

MI, Blei DM Ng AY Jordan et al. (2003). « Latent dirichlet allocation ». In : *J. Mach. Learn. Res* 3.993, p. 1022.

Mikolov, Tomas et al. (2013). « Efficient estimation of word representations in vector space ». In : *arXiv preprint arXiv :1301.3781*.

Mir, Abrar Qadir, Furqan Yaqub Khan et Mohammad Ahsan Chishti (2023). « Online Fake Review Detection Using Supervised Machine Learning And BERT Model ». In : *arXiv preprint arXiv :2301.03225*.

Mohawesh, Rami et al. (2021a). « Analysis of concept drift in fake reviews detection ». In : *Expert Systems with Applications* 169, p. 114318.

Mohawesh, Rami et al. (2021b). « Fake reviews detection : A survey ». In : *IEEE Access* 9, p. 65771-65802.

Mukherjee, Arjun, Bing Liu et Natalie Glance (2012). « Spotting fake reviewer groups in consumer reviews ». In : *Proceedings of the 21st international conference on World Wide Web*, p. 191-200.

Mukherjee, Arjun et al. (2011). « Detecting group review spam ». In : *Proceedings of the 20th international conference companion on World wide web*, p. 93-94.

Mukherjee, Arjun et al. (2013a). « Spotting opinion spammers using behavioral footprints ». In : *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 632-640.

Mukherjee, Arjun et al. (2013b). « What yelp fake review filter might be doing ? » In : *Proceedings of the international AAAI conference on web and social media*. T. 7. 1, p. 409-418.

Nandwani, Pansy et Rupali Verma (2021). « A review on sentiment analysis and emotion detection from text ». In : *Social Network Analysis and Mining* 11.1, p. 81.

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.

Ni, Jianmo, Jiacheng Li et Julian McAuley (2019). « Justifying recommendations using distantly-labeled reviews and fine-grained aspects ». In : *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, p. 188-197.

Noekhah, Shirin, Naomie binti Salim et Nor Hawaniah Zakaria (2020). « Opinion spam detection : Using multi-iterative graph-based model ». In : *Information Processing & Management* 57.1, p. 102140.

Ott, Myle, Claire Cardie et Jeffrey T Hancock (2013). « Negative deceptive opinion spam ». In : *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics : human language technologies*, p. 497-501.

Ott, Myle et al. (2011). « Finding deceptive opinion spam by any stretch of the imagination ». In : *arXiv preprint arXiv :1107.4557*.

Oueslati, Oumayma, Ahmed Ibrahim S Khalil et Habib Ounelli (2018). « Sentiment analysis for helpful reviews prediction ». In : *Int. J* 7, p. 34-40.

Özçelik, Merve et al. (2021). « HisNet : a polarity lexicon based on wordnet for emotion analysis ». In : *Proceedings of the 11th Global Wordnet Conference*, p. 157-165.

Padmaja, S et S Sameen Fatima (2013). « Opinion mining and sentiment analysis-an assessment of peoples' belief : A survey ». In : *International Journal of Ad hoc, Sensor Ubiquitous Computing* 4.1, p. 21.

Patel, Rinki et Priyank Thakkar (2014). « Opinion spam detection using feature selection ». In : *2014 International Conference on Computational Intelligence and Communication Networks*. IEEE, p. 560-564.

Pearson, Karl (1901). « LIII. On lines and planes of closest fit to systems of points in space ». In : *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11, p. 559-572.

Pennington, Jeffrey, Richard Socher et Christopher D Manning (2014). « Glove : Global vectors for word representation ». In : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532-1543.

Peters, ME et al. (2018). « Deep contextualized word representations. arXiv 2018 ». In : *arXiv preprint arXiv :1802.05365* 12.

Putra, Muhammad Mukhtar Dwi, Wikky Fawwaz Al Maki et Ade Romadhony (2021). « Sentiment Analysis on Marketplace Review using Hybrid Lexicon and SVM Method ». In : *2021 9th International Conference on Information and Communication Technology (ICoICT)*. IEEE, p. 66-70.

Ramesh, Bhavana et Charles M Weber (s. d.). « State-of-art methods used in sentiment analysis : A literature review ». In : *2022 Portland International Conference on Management of Engineering and Technology (PICMET)*. IEEE, p. 1-13.

Rastogi, Ajay et Monica Mehrotra (2018). « Impact of behavioral and textual features on opinion spam detection ». In : *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, p. 852-857.

Rastogi, Ajay, Monica Mehrotra et Syed Shafat Ali (2020). « Effective opinion spam detection : A study on review metadata versus content ». In : *Journal of Data and Information Science* 5.2, p. 76-110.

Ravi, Kumar et Vadlamani Ravi (2015). « A survey on opinion mining and sentiment analysis : tasks, approaches and applications ». In : *Knowledge-based systems* 89, p. 14-46.

Ravi Kumar, G, K Venkata Sheshanna et G Anjan Babu (2021). « Sentiment analysis for airline tweets utilizing machine learning techniques ». In : *International Conference on Mobile Computing and Sustainable Informatics : ICMCSI 2020*. Springer, p. 791-799.

Rayana, Shebuti et Leman Akoglu (2015). « Collective opinion spam detection : Bridging review networks and metadata ». In : *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, p. 985-994.

Ren, Yafeng et Donghong Ji (2019). « Learning to Detect Deceptive Opinion Spam : A Survey ». In : *IEEE Access* 7, p. 42934-42945.

Ren, Yafeng, Donghong Ji et Liang Yin (2014). « Deceptive reviews detection based on semi-supervised learning algorithm ». In : *J. Sichuan Univ.(Eng. Sci. Ed.)* 46.3, p. 62-69.

Sakurada, Mayu et Takehisa Yairi (2014). « Anomaly detection using autoencoders with nonlinear dimensionality reduction ». In : *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, p. 4-11.

Saumya, Sunil et Jyoti Prakash Singh (2022). « Spam review detection using LSTM autoencoder : an unsupervised approach ». In : *Electronic Commerce Research* 22.1, p. 113-133.

Schölkopf, Bernhard, Alexander Smola et Klaus-Robert Müller (1998). « Nonlinear component analysis as a kernel eigenvalue problem ». In : *Neural computation* 10.5, p. 1299-1319.

Schuster, Mike et Kuldip K Paliwal (1997). « Bidirectional recurrent neural networks ». In : *IEEE transactions on Signal Processing* 45.11, p. 2673-2681.

Sedighi, Zeinab, Hossein Ebrahimpour-Komleh et Ayoub Bagheri (2017). « RLOSD : Representation learning based opinion spam detection ». In : *2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)*, p. 74-80.

Shaalan, Yassien et al. (2021). « Detecting singleton spams in reviews via learning deep anomalous temporal aspect-sentiment patterns ». In : *Data Mining and Knowledge Discovery* 35.2, p. 450-504.

Shah, Arkesha (2021). « Sentiment analysis of product reviews using supervised learning ». In : *Reliability : Theory & Applications* 16.SI 1 (60), p. 243-253.

Shahariar, GM et al. (2019). « Spam review detection using deep learning ». In : *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, p. 0027-0033.

Shan, Guohou, Lina Zhou et Dongsong Zhang (2021). « From conflicts and confusion to doubts : Examining review inconsistency for fake review detection ». In : *Decision Support Systems* 144, p. 113513.

Shehnepoor, Saeedreza et al. (2021). « HIN-RNN : A graph representation learning neural network for fraudster group detection with no handcrafted features ». In : *IEEE Transactions on Neural Networks and Learning Systems*.

Shojaee, Somayeh et al. (2013). « Detecting deceptive reviews using lexical and syntactic features ». In : *2013 13th International Conference on Intellient Systems Design and Applications*. IEEE, p. 53-58.

Socher, Richard et al. (2013). « Recursive deep models for semantic compositionality over a sentiment treebank ». In : *Proceedings of the 2013 conference on empirical methods in natural language processing*, p. 1631-1642.

Sun, Shiliang, Chen Luo et Junyu Chen (2017). « A review of natural language processing techniques for opinion mining systems ». In : *Information fusion* 36, p. 10-25.

Tang, Xiaoya, Tieyun Qian et Zhenni You (2020). « Generating behavior features for cold-start spam review detection with adversarial learning ». In : *Information Sciences* 526, p. 274-288.

Tenenbaum, Joshua B, Vin de Silva et John C Langford (2000). « A global geometric framework for nonlinear dimensionality reduction ». In : *science* 290.5500, p. 2319-2323.

Tian, Yingjie et al. (2020). « A non-convex semi-supervised approach to opinion spam detection by ramp-one class SVM ». In : *Information Processing & Management* 57.6, p. 102381.

Tiffani, Ilham Esa (2020). « Optimization of naïve bayes classifier by implemented unigram, bigram, trigram for sentiment analysis of hotel review ». In : *Journal of Soft Computing Exploration* 1.1, p. 1-7.

Van Engelen, Jesper E et Holger H Hoos (2020). « A survey on semi-supervised learning ». In : *Machine learning* 109.2, p. 373-440.

Wang, Jingdong et al. (2020). « Fake review detection based on multiple feature fusion and rolling collaborative training ». In : *IEEE Access* 8, p. 182625-182639.

Wang, Tao et Hua Zhu (2014). « Voting for deceptive opinion spam detection ». In : *arXiv preprint arXiv :1409.4504*.

Wang, Yile, Leyang Cui et Yue Zhang (2019). « Using Dynamic Embeddings to Improve Static Embeddings ». In : *arXiv Preprint*.

Wang, Ziyang et al. (2020). « User-based Network Embedding for Collective Opinion Spammer Detection ». In : *arXiv preprint arXiv :2011.07783*.

Washha, Mahdi et al. (2017). « A topic-based hidden Markov model for real-time spam tweets filtering ». In : *Procedia Computer Science* 112, p. 833-843.

Yadollahi, Ali, Ameneh Gholipour Shahraki et Osmar R Zaiane (2017). « Current state of text sentiment analysis from opinion to emotion mining ». In : *ACM Computing Surveys (CSUR)* 50.2, p. 1-33.

Yan, Weizhong et Lijie Yu (2019). « On accurate and reliable anomaly detection for gas turbine combustors : A deep learning approach ». In : *arXiv preprint arXiv :1908.09238*.

Yao, Jianrong, Yuan Zheng et Hui Jiang (2021). « An ensemble model for fake online review detection based on data resampling, feature pruning, and parameter optimization ». In : *IEEE Access* 9, p. 16914-16927.

Yarowsky, David (1995). « Unsupervised word sense disambiguation rivaling supervised methods ». In : *33rd annual meeting of the association for computational linguistics*, p. 189-196.

You, Lan et al. (2020). « Integrating aspect analysis and local outlier factor for intelligent review spam detection ». In : *Future Generation Computer Systems* 102, p. 163-172.

Zeng, Zhi-Yuan et al. (2019). « A review structure based ensemble model for deceptive review spam ». In : *Information* 10.7, p. 243.

Zhang, Wenping, Mengna Xu et Qiqi Jiang (2018). « Opinion Mining and Sentiment Analysis in Social Media : Challenges and Applications ». In : *STACS 98*. STACS 98, p. 536-548.

Zhao, Siyuan et al. (2018). « Towards accurate deceptive opinions detection based on word order-preserving CNN ». In : *Mathematical Problems in Engineering* 2018.

Zheng, Hongwen et Yanxia Zhang (2008). « Feature selection for high-dimensional data in astronomy ». In : *Advances in Space Research* 41.12, p. 1960-1964.

Zhong, Guoqiang, Xiao Ling et Li-Na Wang (2019). « From shallow feature learning to deep learning : Benefits from the width and depth of deep architectures ». In : *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery* 9.1, e1255.

Zhou, Ming et al. (2020). « Progress in neural NLP : modeling, learning, and reasoning ». In : *Engineering* 6.3, p. 275-290.

Ziani, Amel et al. (2021). « Deceptive Opinions Detection Using New Proposed Arabic Semantic Features ». In : *Procedia Computer Science* 189, p. 29-36.

Webographie

Kaggle (2015). *Twitter US Airline Sentiment*. URL : https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment.

Li, Jessica (2018). *UCI ML Drug Review Dataset*. URL : https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018.

Vimal et Tarun (2019). *Amazon Reviews : Unlocked Mobile Phones*. URL : https://www.kaggle.com/datasets/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones.

Annexes

Annexe A
Extraction de caractéristiques

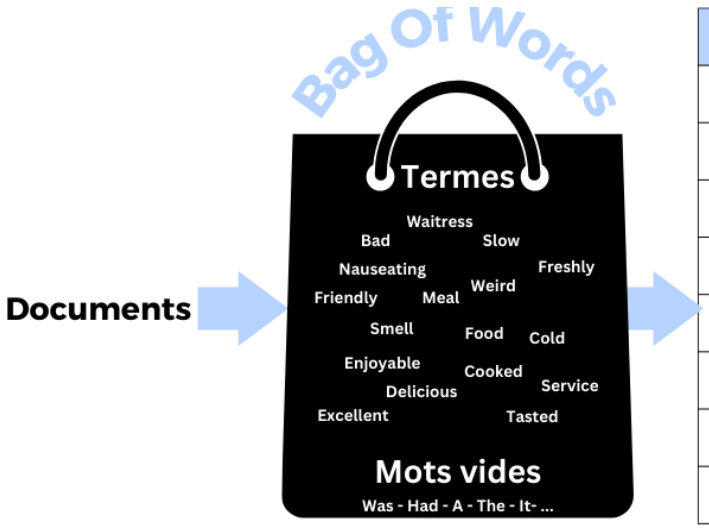
Parmi les méthodes d'extraction de caractéristiques que nous avons mentionné dans la section 2.7.1.3, nous avons cité Tf-Idf (Term Frequency-Inverse Document Frequency). Pour bien expliquer ces méthodes, nous avons cité deux exemples d'avis, extraits de TripAdvisor :

- **D_1 :** *« Had a meal here, food was cold and tasted bad. The smell was weird and nauseating. Our waitress was friendly but the service was very slow. »*

- **D_2 :** *« We had an excellent meal, food was freshly cooked and delicious. Our waitress was very friendly. It was enjoyable. »*

A.1 Bag Of Words

La figure A.1 présente le résultat de l'application de BOW sur les deux documents D_1 et D_2.



Terme	D1	D2	Nauseating	1	0
Food	1	1	Bad	1	0
Friendly	1	1	Excellent	0	1
Freshly	0	1	Smell	1	0
Weird	1	0	Delicious	0	1
Cold	1	0	Cooked	0	1
Tasted	1	0	Service	1	0
Meal	1	1	Slow	1	0
Waitress	1	1	Enjoyable	0	1

Fig. A.1 : Exemple du fonctionnement de Bag-of-Words (BOW)

A.2 N-grammes

La figure A.2 ci-dessous de l'annexe est un exemple simple de n-gramme appliqué sur une phrase extraite de D_1.

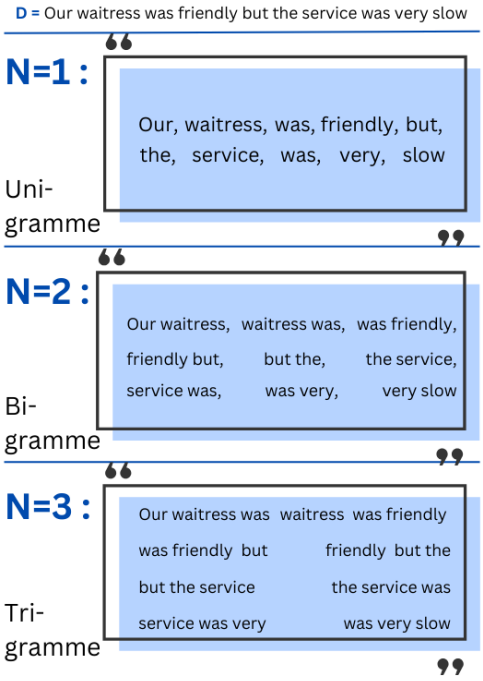


FIG. A.2 : Exemple du fonctionnement des n-grammes

A.3 Part-of-speech tagging

La figure A.3 de l'annexe représente un exemple simple où nous avons appliqué la technique POS sur une D_1 et nous avons représenté les étiquettes par différentes couleurs.

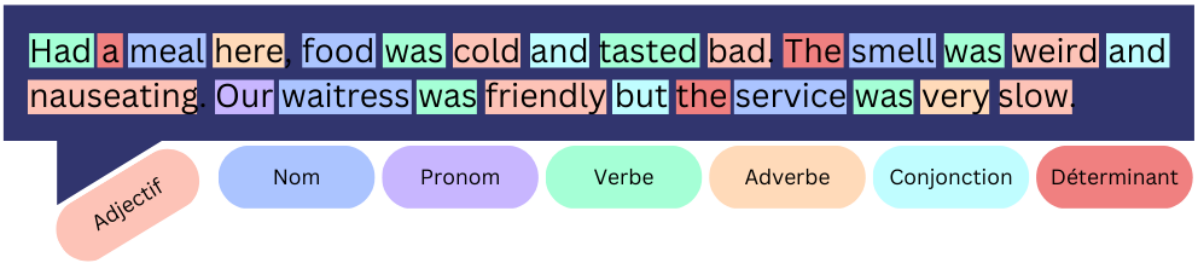


FIG. A.3 : Exemple du fonctionnement de la technique POS (Part-of-speech)

A.4 TF-IDF

La figure A.4 illustre un exemple d'application de TF et de TF-IDF sur D_1 et D_2.

Terme	Terme Frequency (TF)		N	IDF	TF-IDF	
	C1	C2			C1	C2
meal	0.038	0.052	2	0.301	0.012	0.016
food	0.038	0.052	2	0.301	0.012	0.016
cold	0.038	0	1	0.602	0.023	0
tasted	0.038	0	1	0.602	0.023	0
bad	0.038	0	1	0.602	0.023	0
smell	0.038	0	1	0.602	0.023	0
weird	0.038	0	1	0.602	0.023	0
nauseating	0.038	0	1	0.602	0.023	0
waitress	0.038	0.052	2	0.301	0.012	0.016
friendly	0.038	0.052	2	0.301	0.012	0.016
service	0.038	0	1	0.602	0.023	0
slow	0.038	0	1	0.602	0.023	0
excellent	0	0.052	1	0.602	0	0.032
freshly	0	0.052	1	0.602	0	0.032
cooked	0	0.052	1	0.602	0	0.032
delicious	0	0.052	1	0.602	0	0.032
enjoyable	0	0.052	1	0.602	0	0.032

Fig. A.4 : Exemple du fonctionnement de TF et de TF-IDF