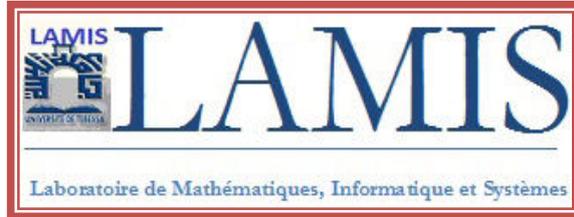


RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE  
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ LARBI TEBESSI, TEBESSA  
Département des Mathématiques et Informatique  
Laboratoire de Mathématiques, d'Informatique et Systèmes (LAMIS)



## Thèse

Présentée pour l'obtention du diplôme de

DOCTORAT LMD

En

**Informatique**

Spécialité : **Systemes d'information coopératifs**

Titre

# **Opinion Mining : Extraction & structuration**

Présentée et soutenue par

**Bouras Dalila**

Devant le jury :

Pr. LAOUAR Med Ridda	Prof Université Larbi Tébessi, Tébessa	Président
Dr. ZAROOUR Karim	MCA Université Constantine 2, Constantine	Examineur
Dr. LAIMECHE Lakhdar	MCA Université Larbi Tébessi, Tébessa	Examineur
Dr. SIAM Abderahim	MCA Université Abbes Laghrour, Khenchela	Examineur
Pr. BENDJENNA Hakim	Prof Université Larbi Tébessi, Tébessa	Directeur de thèse
Dr. AMROUNE Mohamed	MCA Université Larbi Tébessi, Tébessa	Co-encadreur

Le .../.../.....

## *Dédicaces*

*J'ai voulu commencer par ma petite boule familiale :*

*Je dédie ce modeste travail, spécialement à mon cher mari pour son soutien sans faille et ses encouragements, je le remercie d'avoir toujours trouvé les mots justes dans les moments difficiles et m'avoir apporté le réconfort dont j'avais besoin.*

*A mon petit soleil Iskandar et mes trois étoiles Nouha, Maram & Yasmina.*

*A mes parents, Je me souviendrai toujours de tous les moments où ils m'ont poussé à travailler, si je suis aujourd'hui arrivé à ce niveau et à ce degré de réussite c'est grâce à leurs efforts et conseils.*

*A mes très chers frères, sœurs, belle mère, belles sœurs et beau frère avec qui j'apprends toujours dans la vie, merci.*

*A la mémoire de ma tante Nadja.*

*"Ma deuxième mère, comme le temps passe, la douleur en moi reste vivace car rien ne peut combler le vide ainsi que l'infinie tristesse que tu laisses dans ma vie.  
« Allah yarhmak ».*

## *Remerciements*

Je tiens tout d'abord à remercier le Bon Dieu tout-puissant de m'avoir éclairé de son savoir, guidé tout au long de ce travail, et de m'avoir donné le courage et la puissance pour terminer ce travail.

J'exprime mes chaleureux remerciements à mon directeur de thèse Prof. BENDJENNA Hakim Professeur à l'Université de Tébessa, qui a accordé de son temps et qui a contribué pour l'orientation de ce travail dans la bonne direction pour ses précieux conseils et ses encouragements.

Je tiens à exprimer ma profonde gratitude à Dr. Amroune Mohamed, pour avoir co-encadré et dirigé mes recherches. Je le remercie de m'avoir soutenu et appuyé tout au long de ma thèse. Ses précieux conseils, son exigence et ses commentaires ont permis d'améliorer grandement la qualité de mes travaux.

Je remercie le plus sincèrement possible les membres de jury Prof. LAOUAR Med Ridda, Dr. ZAROOUR Karim, Dr. LAIMECHE Lakhdar et Dr. SIAM Abderahim : pour leur attention et l'intérêt porté à cette thèse et qui ont eu l'obligeance d'accepter de juger ce travail.

Je présente mes sincères remerciements à tous les membres de LAMIS qu'ils soient professeurs, docteurs et doctorants qui ont contribué à élaborer ce laboratoire de recherche et qui ont su rendre mon travail agréable à travers leurs conseils, leur simple présence ainsi que les différentes journées doctorales et les conférences élaborées.

## Abstract

What others think has always been a very important piece of information for most of us during the decision-making process. Long before the knowledge of the World Wide Web became widespread, many of us asked their friends' opinions to recommend a product or use a service. Nowadays, with the evolution of Web technologies, internet users are constantly invited to share their opinions and preferences with the rest of the world, which has led to an increase in opinion blogs, product reviews, services and comments on virtually everything. This type of web content is increasingly recognized as a source of data adding value to multiple application domains. However, people often find it difficult to manually read each journal, to identify and summarize detailed feelings or opinions buried in rich resources. The traditional search of opinions, which focuses on general opinions, fails to reveal feelings about aspects of the entities examined. To do this, the aspect-based opinion search aims to extract product-related aspects and the corresponding opinions expressed by users.

As part of this thesis we study the Aspect Based Opinion Search problem (ABOM), which consists in extracting the aspects relating to the entities criticized and the corresponding opinions to each aspect expressed by the users in texts written in natural language. ABOM includes several subtasks: 1) Aspect extraction, 2) Aspect category classification, and 3) Polarity classification of feelings at the aspect level. We focused on extracting aspects and detecting the polarity of feelings about aspects of each entity. To do this, we propose in this thesis a Deep Learning approach, based on Tree-Long Short Term Memory (LSTM) deep neural networks, including the extraction of aspect and classification of opinions where they are simultaneously managed in the same network of deep neurons of the Tree-LSTM type. The results on the SemEval 2014 benchmark dataset suggest the effectiveness of our proposed approach and demonstrate that it performs better in relation to various proposed architectures.

**Keywords:** Opinion Mining, sentiment analysis, aspect extraction, neural networks, Deep Learning, word embedding, Tree-LSTM.

## Résumé

Ce que les autres pensent a toujours été un élément d'information très important pour la plupart d'entre nous au cours du processus de la prise de décision. Bien avant que la connaissance du World Wide Web ne se généralise, beaucoup d'entre nous ont demandé l'avis de leurs amis pour recommander un produit ou utiliser un service. De nos jours, et avec l'évolution des technologies Web, les utilisateurs d'internet sont constamment invités à partager leurs opinions et leurs préférences avec le reste du monde, ce qui a entraîné une augmentation de blogs d'opinions, de critiques de produits, de services et de commentaires sur pratiquement tout. Ce type de contenu Web est de plus en plus reconnu comme une source de données apportant une valeur ajoutée à plusieurs domaines d'application. Cependant, les gens ont souvent du mal à lire manuellement chaque revue, à identifier et à résumer des sentiments ou des opinions détaillés enfouis dans les ressources riches en opinions. La fouille traditionnelle des opinions, qui se concentre sur les opinions générales, ne parvient pas à révéler les sentiments concernant les aspects des entités examinées. Pour ce faire, la fouille d'opinion basée sur les aspects vise à extraire les aspects relatifs aux produits et les opinions correspondantes exprimées par les utilisateurs.

Dans le cadre de cette thèse nous étudions le problème de recherche de la fouille d'opinion basée sur les aspects (Aspect based opinion mining (ABOM)), qui consiste à extraire les aspects relatifs aux entités critiquées et les opinions correspondantes à chaque aspect exprimées par les utilisateurs dans des textes porteurs d'opinion écrites en langue naturelle. ABOM comprend plusieurs sous-tâches: 1) extraction d'aspect, 2) classification des catégories d'aspect, et 3) classification de la polarité des sentiments au niveau d'aspect. Nous nous sommes concentrés sur l'extraction des aspects et la détection de la polarité des sentiments concernant les aspects de chaque entité. Pour ce faire, nous proposons dans cette thèse une approche Deep Learning, basée sur les réseaux de neurones en profondeur de type LSTM arborescente (Tree-Long Short Term Memory), dont laquelle l'extraction des d'aspect et la classification des opinions sont simultanément gérées dans un même réseau de neurones en profondeur de type Tree-LSTM. Les résultats sur l'ensemble de données de référence SemEval 2014 suggèrent l'efficacité de notre approche proposée et démontrent qu'elle réalise de meilleures performances par rapport à diverses architectures présentées dans le domaine.

**Mots-clés :** Fouille d'opinion, analyse des sentiments, extraction d'aspect, réseaux de neurone, apprentissage en profondeur, prolongement des mots, Tree-LSTM.

## المخلص

ما يفكر فيه الآخرون كان دوماً عنصراً معلوماً جيداً لمعظمنا خلال عملية اتخاذ القرار. قبل تعميم المعرفة بشبكة الويب العالمية، كان الكثير منا يلجأ إلى طلب آراء أصدقائهم أو أقاربهم للتوصية بمنتج أو خدمة. لكن في الوقت الحاضر، مع تطور تقنيات الويب، أصبح مستخدمو الإنترنت مطالبون بمشاركة آرائهم مع بقية العالم باستمرار، وهذا ما أدى إلى زيادة كبيرة في مدونات الرأي، انتقادات المنتجات، الخدمات والتعليقات على كل شيء تقريباً. وقد أصبح هذا النوع من محتوى الويب مصدر للبيانات التي تضيف قيمة إلى مجالات تطبيق متعددة. ومع ذلك، غالباً ما يجد الأشخاص صعوبة في قراءة هذه المدونات، تحديد وتلخيص المشاعر والآراء المفصلة ضمن هذه الموارد الغنية بالآراء. إن البحث التقليدي الذي يركز على الآراء العامة، لم يتمكن من الكشف عن المشاعر الموجهة لخصائص العنصر المراد انتقاده. للقيام بذلك، يهدف البحث عن الرأي القائم على الخصائص إلى استخراج الخصائص المتعلقة بالمنتج وقطبية الآراء المقابلة لكل خاصية أعرب عنها المستخدمون.

في إطار هذه الأطروحة، ندرس مشكلة البحث عن الرأي القائم على الخصائص، والذي يتمثل في استخراج الخصائص المتعلقة بالمنتجات التي تم انتقادها والآراء المقابلة لكل خاصية تم التعبير عنها من قبل المستخدمين ضمن نصوص مكتوب بلغة المستخدم. يتضمن البحث عن الرأي القائم على الخصائص عدة مهام فرعية وهي: (1) استخراج الخاصية، (2) تصنيف فئة الخاصية، و (3) تصنيف قطبية المشاعر بالنسبة لكل خاصية. ركزنا في هذه الأطروحة على استخراج الخصائص واكتشاف قطبية المشاعر التابعة لكل خاصية في المنتج. للقيام بذلك، نقترح في هذه الأطروحة طريقة للتعلم العميق، تعتمد على الشبكات العصبية العميقة للذاكرة من نوع شجرة (LSTM)، أين تتم كل من استخراج الخصائص وتصنيف قطبية المشاعر المتعلقة بكل خاصية في آن واحد و بنفس الشبكة العصبية العميقة من نوع شجرة LSTM. تشير نتائج الطريقة المقترحة و المطبقة على مجموعة البيانات المرجعية SemEval 2014 إلى فعالية المقترح الذي حقق ضمن هذه الأطروحة أداء أفضل مقارنة بالعديد من الدراسات المقدمة في هذا المجال.

## الكلمات المفتاحية:

بحث الرأي، تحليل المشاعر، استخراج الجانب، الشبكات العصبية، التعلم العميق، امتداد الكلمات، LSTM شجرة.

## Liste des figures

<b>Figure 1.1</b> : Les données de Google Trends sur les tendances liées à l'analyse des sentiments .....	12
<b>Figure 1.2</b> : Les niveaux d'analyse dans le domaine de la fouille d'opinion (B. Liu, 2012).....	13
<b>Figure 1.3</b> : Domaine d'application de la fouille d'opinion (B. Liu, 2012).....	15
<b>Figure 1.4</b> : Les différentes tâches de la fouille d'opinion (Pozzi, Fersini, Messina, & Liu, 2017).....	16
<b>Figure 1.5</b> : Le processus général de la fouille d'opinion (B. Liu, 2012).....	17
<b>Figure 1.6</b> : Les techniques de classification des opinions(Medhat, Hassan, & Korashy, 2014).....	21
<b>Figure 1.7</b> : Support à vaste marge dans un problème de classification (Cortes & Vapnik, 1995).....	24
<b>Figure 2.1</b> Illustration du problème de la fouille d'opinion basé sur aspects. ....	32
<b>Figure 3.1</b> L'intelligence artificielle (IA), l'apprentissage par la machine (ML) et l'apprentissage en profondeur (DL) (Goodfellow et al., 2016).....	53
<b>Figure 3.2</b> La différence entre apprentissage par la machine (ML) et l'apprentissage en profondeur (DL) (Yann LeCun et al., 2015) (Goodfellow et al., 2016) .....	54
<b>Figure 3.3</b> : Description des modèles CBOW et Skip-gram (Tomas Mikolov, Sutskever, Chen, Corrado, & Dean, 2013).....	57
<b>Figure 3.4</b> : Réseau de neurone pour traitement de la langue naturelle au niveau fenêtre (Collobert, Weston, et al., 2011).....	58
<b>Figure 3.5</b> : Réseau de neurone pour traitement de la langue naturelle au niveau phrase(Collobert, Weston, et al., 2011).....	59
<b>Figure 3.6</b> : Architecture d'un réseau de neurone simple Vs réseau de neurone approfondi (Yann LeCun, Yoshua Bengio, & Geoffrey Hinton, 2015). ....	59
<b>Figure 3.7</b> : Architecture du réseau de neurone convolutionnel (Y. Kim, 2014b).....	60
<b>Figure 3.8</b> : Architecture du réseau de neurone récurrent (Tomáš Mikolov, Karafiát, Burget, Černocký, & Khudanpur, 2010) .....	62
<b>Figure 3.9</b> : Architecture d'un réseau de neurone de type LSTM (Hochreiter & Schmidhuber, 1997).....	65
<b>Figure 3.10</b> : Comportement d'un réseau de neurone type LSTM (Hochreiter & Schmidhuber, 1997). ....	65
<b>Figure 3.11</b> : Un réseau LSTM standard (a) & Un réseau LSTM arborescent avec facteur de branchement arbitraire $N = 2$ (b) (Tai, Socher, & Manning, 2015). ....	67
<b>Figure 3.12</b> : Stratégie d'apprentissage classique & d'apprentissage en profondeur (Thanaki, 2017).....	71
<b>Figure 3.13</b> : Architecture globale de l'AEN (Song et al., 2019).....	72
<b>Figure 3.14</b> : Modèle de classification des sentiments par niveau d'aspect (B. Huang et al., 2018).....	73
<b>Figure 3.15</b> : Architecture du modèle de(Chen, Sun, Bing, & Yang, 2017. p. 452-461).....	73
<b>Figure 3.16</b> : la classification des sentiments au niveau aspect(Tang, Qin, & Liu, 2016). ....	74
<b>Figure 3.17</b> : Modèle général pour la fouille d'opinion basée aspect(Xu, Liu, Shu, & Yu, 2018) .....	75
<b>Figure 3.18</b> : Illustration de modèle GCAE pour Tâche ACSA(Xue & Li, 2018). ....	75
<b>Figure 4.1</b> Approches supervisées pour la Fouille d'opinion basée aspects (Bouras et al., 2019) .....	78
<b>Figure 4.2</b> Approches semi supervisées pour la fouille d'opinion basée aspects (Bouras et al., 2019).....	80
<b>Figure 4.3</b> Approches non supervisées pour la fouille d'opinion basée aspects (Bouras et al., 2019).....	81
<b>Figure 4.4</b> : Les différents types du réseau Tree-LSTM (Tai et al., 2015).....	86
<b>Figure 4.5</b> : L'architecture de l'approche proposée (Bouras, Amroune, Bendjenna, & Bendib, 2020).....	87
<b>Figure 4.6</b> : Le bloc de mémoire à chaque nœud du réseau de type 'circonscription arborescente-LSTM (Bouras et al., 2020).....	89
<b>Figure 4.7</b> : (a) Un exemple d'arbre de circonscription, (b) Le réseau Tree-LSTM correspondant (Bouras et al., 2020).....	90
<b>Figure 4.8</b> : Le comportement de notre modèle Tree-LSTM sur une partie de l'arbre de circonscription de la phrase «The laptop is great» (Bouras et al., 2020). ....	90
<b>Figure 4.9</b> Exemple de jeu de données SemEval 2014 Task 4 sur les ordinateurs portables (Pontiki et al., 2014).....	92

## Liste des tableaux

<b>Tableau 2.1</b> Différents jeux de données de SemEval (Pontiki et al., 2014) .....	36
<b>Tableau 2.2</b> Équipes de recherche ayant contribué à la création des jeux de données pour chaque langue. (Pontiki et al., 2015).....	37
<b>Tableau 2.3</b> : Synthèse des approches supervisées(Bouras, Amroune, & Ben Djanna, 2019).....	40
<b>Tableau 2.4</b> : Synthèse des approches semi-supervisées(Bouras et al., 2019). .....	42
<b>Tableau 2.5</b> : Synthèse des approches non supervisées(Bouras et al., 2019).....	45
<b>Tableau 4.1:</b> Statistique SemEval 2014 du jeu de données des ordinateurs portable. ....	91
<b>Tableau 4.2:</b> Table de comparaison de note approche avec les approches existantes en termes de ...	95
<b>Tableau 4.3:</b> Comparaison de notre approche avec les approches existantes en terme d'Accuracy sur une prédiction à 3 classes concernant le jeu de données SemEval 2014 Task 4 (Bouras et al., 2020)...	96
<b>Tableau 4.4:</b> Résultats de notre approche en termes de Précision (P), Recall (R),.....	96



# Sommaire

<b>Introduction générale .....</b>	<b>1</b>
▪ Contexte du travail .....	1
▪ Motivation & contributions.....	2
▪ Publications.....	5
▪ Structure de la thèse.....	5
<b>1 Chapitre I : La Fouille d'opinion : un domaine émergent.....</b>	<b>7</b>
<b>1.1 Introduction .....</b>	<b>7</b>
<b>1.2 Définitions d'une opinion .....</b>	<b>8</b>
1.2.1 Définition 1 .....	8
1.2.2 Définition 2 .....	8
<b>1.3 Différents types d'opinion .....</b>	<b>9</b>
1.3.1 Opinion régulière versus opinion comparative.....	9
1.3.2 L'opinion implicite versus explicite .....	10
<b>1.4 Définition de la fouille d'opinion .....</b>	<b>10</b>
<b>1.5 Les niveaux d'analyse d'opinion .....</b>	<b>12</b>
1.5.1 Niveau document.....	13
1.5.2 Niveau phrase.....	13
1.5.3 Niveau aspect .....	13
<b>1.6 Domaine d'application de la fouille d'opinion .....</b>	<b>14</b>
<b>1.7 Les différentes tâches de la fouille d'opinion .....</b>	<b>15</b>
<b>1.8 Processus de classification des opinions.....</b>	<b>16</b>
1.8.1 Acquisition des données.....	17
1.8.2 Prétraitement des données.....	17
1.8.3 Extraction des caractéristiques .....	18
1.8.4 Classification d'opinion .....	20
<b>1.9 Les techniques de classification d'opinion.....</b>	<b>20</b>
1.9.1 Les approches par apprentissages de la machine.....	22
1.9.2 Les approches basées sur le lexique.....	26
<b>1.10 Conclusion .....</b>	<b>29</b>
<b>2 Chapitre II : La Fouille d'opinion basée sur les aspects.....</b>	<b>30</b>
<b>2.1 Introduction .....</b>	<b>30</b>
<b>2.2 Formalisation du problème et concepts de base .....</b>	<b>31</b>
2.2.1 Le concept d'entité .....	32
2.2.2 Le concept d'aspect.....	32
2.2.3 Le concept terme d'aspect .....	32
<b>2.3 Mesures d'évaluation de la performance des systèmes de la fouille d'opinion.....</b>	<b>34</b>
<b>2.4 Les datasets d'expérimentation .....</b>	<b>35</b>
<b>2.5 Etat de l'art sur la fouille d'opinion basée sur les aspects .....</b>	<b>37</b>
2.5.1 Approches supervisées .....	37
2.5.2 Approches semi-supervisées .....	40
2.5.3 Approches non supervisées.....	42
<b>2.6 Conclusion .....</b>	<b>45</b>
<b>3 Chapitre III : Les modèles d'apprentissage approfondi pour la fouille d'opinion basée aspect ..</b>	<b>47</b>

<b>3.1</b>	<b>Introduction .....</b>	<b>47</b>
<b>3.2</b>	<b>Les techniques traditionnelles pour l'extraction des caractéristiques .....</b>	<b>49</b>
3.2.1	La présence et fréquence des termes.....	49
3.2.2	Les parties du discours (POS) .....	49
3.2.3	Les mots et expressions d'opinion .....	50
3.2.4	Les négations.....	50
<b>3.3</b>	<b>Le traitement du langage naturel (TLN (NLP)).....</b>	<b>50</b>
<b>3.4</b>	<b>Les problèmes de traitement du langage naturel .....</b>	<b>51</b>
<b>3.5</b>	<b>L'apprentissage par la machine .....</b>	<b>51</b>
<b>3.6</b>	<b>Définition de l'apprentissage approfondi (en profondeur).....</b>	<b>52</b>
<b>3.7</b>	<b>Les avantages de l'apprentissage en profondeur.....</b>	<b>53</b>
<b>3.8</b>	<b>Les modules de base d'apprentissage en profondeur.....</b>	<b>54</b>
3.8.1	Les mots plongés (Word Embedding).....	54
3.8.2	L'architecture du réseau de neurone pour le traitement du langage naturel.....	57
3.8.3	Les réseaux de neurone convolutionnels (CNN) .....	60
3.8.4	Les réseaux de neurone récurrent (RNN).....	62
3.8.5	Les réseaux de mémoire à long-court terme (LSTM) .....	63
<b>3.9</b>	<b>La stratégie d'apprentissage des réseaux en profondeur .....</b>	<b>70</b>
<b>3.10</b>	<b>Quelques travaux récents basés Deep Learning pour la fouille d'opinion basée aspects .....</b>	<b>71</b>
3.10.1	Réseau de codeurs attentionnels pour la classification de sentiments ciblés.....	72
3.10.2	Classification des opinions avec des réseaux de neurone Attention-Sur-Attention .....	72
3.10.3	Réseau d'attention récurrent en mémoire pour l'analyse des sentiments.....	73
3.10.4	Classification de sentiment au niveau Aspect avec Deep Memory Network .....	74
3.10.5	Double embeddings et étiquetage de séquence basée sur CNN pour l'extraction d'aspect .....	74
3.10.6	Analyse du sentiment basée sur les aspects avec des réseaux convolutionnels à porte .....	75
<b>3.11</b>	<b>Conclusion .....</b>	<b>76</b>
<b>4</b>	<b><i>Chapitre IV : une approche basée Tree-LSTM pour la fouille d'opinion basée sur les aspects ....</i></b>	<b>77</b>
<b>4.1</b>	<b>Introduction .....</b>	<b>77</b>
<b>4.2</b>	<b>Contribution 1 : Etude comparative et synthèse sur des approches existantes de la fouille d'opinion basée aspects.....</b>	<b>77</b>
4.2.1	Synthèse.....	82
<b>4.3</b>	<b>Contribution 2 : Une approche basée Tree-LSTM pour la fouille d'opinion basée aspect.....</b>	<b>84</b>
4.3.1	Présentation de l'approche proposée .....	84
<b>4.4</b>	<b>Expérimentation.....</b>	<b>91</b>
4.4.1	Jeux de données ( datasets) .....	91
4.4.2	Environnement et paramètres d'exécution .....	92
4.4.3	Mesures d'évaluation .....	92
<b>4.5</b>	<b>Discussion.....</b>	<b>93</b>
4.5.1	Comparaison des modèles .....	93
<b>4.6</b>	<b>Analyse des résultats.....</b>	<b>96</b>
4.6.1	La sous tâche d'extraction d'aspect.....	96
4.6.2	La sous tâche de la détection de la polarité.....	97
<b>4.7</b>	<b>Conclusion .....</b>	<b>98</b>
	<b><i>Conclusion générale.....</i></b>	<b>99</b>



*"La connaissance des mots conduit à la connaissance des choses."*

«Platon»

### 1. Contexte du travail

Le Web 2.0 (Graham, 2005; O'Reilly, 2005) a fait son émergence, avec des espaces plus interactifs comme les sites sociaux, les blogs et les forums. Les utilisateurs qui étaient jusque-là passifs dans le Web 1.0, sont devenus actifs dans le Web 2.0. Où chacun peut ajouter des informations et partager ses opinions (Birdsall, 2007). De ce fait, le nombre de documents contenant des informations exprimant des opinions, des sentiments ou des jugements d'évaluation devient de plus en plus important. L'émergence de ces contenus générés par les utilisateurs via les médias sociaux a eu un impact indéniable sur les environnements commerciaux. En fait, les médias sociaux ont déplacé la publication de contenu des entreprises vers les clients (O'Brien, 2011). Bien qu'il existe plusieurs sources de contenu généré par l'utilisateur (forums de discussion, tweets, blogs, actualités et rapports, commentaires des consommateurs sur les courriels, centres d'appels, etc.), aucune d'entre elles n'est aussi ciblée que les critiques en ligne. C'est pourquoi les consommateurs lisent et consultent principalement les avis en ligne avant d'acheter des produits ou d'utiliser des services. Cependant, le nombre croissant d'opinions en ligne rend de plus en plus difficile la prise de décisions éclairées.

D'autre part, les revues en ligne offrent aux entreprises une riche source d'opinions gratuites des consommateurs. Des études marketing montrent que les avis en ligne influencent considérablement le comportement des consommateurs (Lipsman, 2007). De nombreuses entreprises suivent maintenant les commentaires des clients via des sources en ligne. Amazon<sup>1</sup>, Cnet<sup>2</sup>, Rateitall<sup>3</sup>, Epinions<sup>4</sup>, TripAdvisor<sup>5</sup>, Yelp<sup>6</sup> et Ciao UK<sup>7</sup> sont des exemples de ces ressources internet contenant des avis de consommateurs.

---

<sup>1</sup> <http://www.amazon.com/>

<sup>2</sup> <http://www.cnet.com/>

<sup>3</sup> <http://www.rateitall.com/>

<sup>4</sup> <http://www.epinions.com/>

<sup>5</sup> <http://www.tripadvisor.com>

<sup>6</sup> <http://www.yelp.com/>

<sup>7</sup> <http://www.ciao.co.uk/>

Ces besoins ont attiré les chercheurs vers un nouveau domaine d'étude, appelé « fouille d'opinion », afin d'analyser des opinions, des sentiments, des évaluations, des attitudes et des émotions tirés du contenu généré par l'utilisateur.

## 2. Motivations & contributions

La fouille d'opinion (également connu sous le nom d'analyse de sentiment) est une étude informatique d'informations subjectives vers différentes entités. Les entités font généralement référence à des produits, organisations, services ou / et à leurs aspects, fonctions, composants et attributs.

La fouille d'opinions est une tâche majeure du traitement de la langue naturelle (NLP) qui étudie des méthodes permettant d'identifier et d'extraire des opinions à partir de textes écrits, telles que des critiques de produits, des groupes de discussion, des forums et des blogs. Il fait du Web une source d'information étendue et excellente pour recueillir des opinions sur un objet spécifique. Avec la croissance indéniable du Web, les particuliers et les organisations utilisent le contenu en ligne pour leurs décisions d'achat et de fabrication.

Chaque fois que quelqu'un tente de découvrir ce que les autres personnes pensent de quelque chose sur le Web, la réponse fournie est une énorme quantité de données, ce qui rend difficile la recherche d'informations utiles facilement. Pour les organisations, le suivi des commentaires des clients peut aider à mesurer le niveau de satisfaction et à prendre des décisions optimales en matière de fabrication et de vente. En raison de limitations mentales et physiques, il est difficile de rassembler et d'analyser manuellement la masse d'informations sur le Web. Par conséquent, un système capable de résumer automatiquement les documents est de plus en plus souhaitable. Un tel système extrait les informations pertinentes et les présentes de manière facile à lire et à comprendre afin de prendre des décisions éclairées.

Pour collecter les informations initiales, la plupart des pages Web de commerce électronique fournissent un espace de commentaire ou de révision dans lequel les clients individuels peuvent échanger leurs expériences et leurs opinions. Le contenu du commentaire client est souvent présenté sous forme de texte en langage naturel sous une forme non structurée. Par conséquent, analyser et extraire des informations utiles est une tâche difficile qui nécessite des techniques spécialisées.

Il existe généralement deux types d'informations textuelles dans les avis clients : les déclarations objectives, qui représentent des faits, et les déclarations subjectives, qui symbolisent des opinions ou des perceptions.

L'extraction d'opinions peut être étudiée à trois niveaux différents, à savoir les niveaux de document, de phrase et d'aspect (Pang & Lee, 2008). Une analyse des sentiments au niveau du document permet de classer un document contenant une opinion (par exemple, une analyse de produit) en fonction du sentiment général du document entier. Cela suppose que tout le document exprime une opinion unique. De même, une analyse de sentiment au niveau de la phrase classe le sentiment au niveau de la phrase. Cependant, toutes les phrases ne sont pas des phrases subjectives (Wilson, Wiebe, & Hoffmann, 2005).

Bien que la fouille d'opinions au niveau des documents et des phrases soit précieuse, elle ne découvre jamais ce que les gens aiment ou n'aiment pas. Un texte d'opinion sur une entité particulière ne signifie pas que l'auteur aime ou n'aime pas tous les aspects de l'entité. « L'appareil photo Canon est incroyable, elle est mieux que l'appareil photo Samsung », est un exemple d'avis qui exprime une opinion positive sur un produit et une opinion négative sur un autre produit. Il n'est pas valide de classer et de généraliser le sentiment sur les deux produits. Pour obtenir des opinions détaillées, il est nécessaire d'examiner le niveau d'aspect (K. Liu, Xu, & Zhao, 2015).

Le niveau d'aspect est également connu sous le nom d'extraction d'opinion basée sur les aspects (Aspect Based Opinion Mining ABOM), qui identifie et extrait les opinions et leurs cibles. ABOM a trois tâches principales : premièrement, identifier et extraire les aspects du produit ; ensuite, déterminez les mots d'opinion et leurs polarités ; et enfin cartographier la relation entre les aspects et les opinions. Si ABOM était appliqué à l'exemple précédent, le système devrait identifier «appareil photo Canon », puis identifier l'opinion «incroyable » et enfin cartographier la relation dans laquelle l'opinion correspond à « appareil photo Canon » et non à « appareil photo Samsung ».

Notre étude s'articule autour d'une problématique principale qui est l'extraction d'une manière algorithmique des opinions où l'objectif principal est d'analyser spécifiquement les avis d'opinions des clients, d'identifier et d'extraire des aspects et des opinions et de les présenter sous forme de paires aspect-opinion. Cette recherche explore l'utilisation de techniques de traitement du langage naturel et se concentre principalement sur l'analyse et l'extraction d'opinions et d'aspects à partir de textes écrits en langage naturel.

À cet effet, les questions scientifiques que nous abordons dans le cadre de cette thèse sont les suivantes :

- ❖ Quels sont les travaux réalisés dans le domaine de la fouille d'opinion spécifiquement au niveau fin de granularité (au niveau aspect) ?

- ❖ Comment concevoir un système robuste d'analyse d'opinions basée sur les aspects en faisant appel à technique de Deep Learning ?
- ❖ Comment peut-on introduire un ensemble riche de caractéristique avec moins d'effort et de temps tout en analysant les avis d'une manière approfondie ?
- ❖ Comment peut-on présenter nos données d'une manière structurée ?

Afin de répondre aux questions ci-dessus et cela par ordre de priorité, nous avons pu cerner nos objectifs par les points suivants :

- ❖ Une étude comparative entre les différents travaux réalisés par les chercheurs de la communauté pour faire face aux défis du domaine et pour ne pas répéter les notions qui sont déjà introduites par d'autres chercheurs.
- ❖ Concevoir un système performant pour la fouille d'opinion basée sur les aspects, qui implique les deux sous-tâches principales suivantes :
  - Identification des aspects : à partir d'une revue de produit ou de service spécifique (par exemple, un téléphone intelligent ou un restaurant), l'objectif est d'extraire automatiquement de la revue tous les aspects de produit possibles et pertinents. Par exemple, dans « Les images sont claires » et « La qualité d'image est excellente », extrayez le mot « images » et le mot « image » qui présentent les aspects de l'appareil photo. Connaître les aspects pertinents nous amène à développer des méthodes pour détecter les aspects d'un produit et les opinions correspondantes à partir du texte en langage naturel.
  - Identification des sentiments : La deuxième tâche consiste à identifier les opinions correspondantes pour chaque aspect extrait. Dans l'exemple précédent, le système détecte les sentiments, le mot « clair » et le mot « grand », puis analyse la polarité comme négative ou positive.
- ❖ Opter pour une analyse des opinions basée sur les aspects de manière plus approfondis en appliquant les techniques de traitement de la langue naturelle et de l'apprentissage approfondis (Deep Learning).
- ❖ Présenter les données d'entrées d'une manière structurée tout en gardant l'aspect sémantique entre les différentes parties de la phrase.

### 3. Publications

#### ▪ Article de journal

- ❖ Bouras, D., Amroune, M., Bendjenna, H, and Azizi, Nabiha, “Techniques and Trends for Fine-Grained Opinion Mining and Sentiment Analysis: Recent Survey”, Recent Advances in Computer Science and Communications (2019), 13.  
<https://doi.org/10.2174/2213275912666181227144256>
- ❖ Bouras, D., Amroune, M., Bendib, I & Hakim Bendjenna (2020). “Improving Fine-Grained Opinion Mining based Approach with a Deep Constituency Tree-Long Short Term Memory Network and Word Embedding”: Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) (2020) 13: 1.  
<https://doi.org/10.2174/2666255813999200922142212>.

#### ▪ Articles dans une conférence internationale

- ❖ Bouras, D., Amroune, M., & Bendjanna, H, “A review of recent aspect extraction techniques for opinion mining systems”, 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), Algeria, Algiers, 2018, pp. 1-6, doi: 10.1109/ICNLSP.2018.8374382.(2018)
- ❖ Dalila, B., Bendjanna, H., Mohamed, A., “Arabic Opinion Mining: State of the Art”. In the 3rd International Conference on Networking and Advanced Systems (ICNAS 2017),Algeria, Annaba. (2017)

### 4. Structure de la thèse

Cette thèse est organisée en quatre chapitres comme suit :

**Les trois premiers chapitres** clarifient les éléments nécessaires pour la bonne compréhension des sujets abordés dans la suite de cette thèse, où :

- **Le chapitre I** aborde le contexte général de ce travail qui est le domaine de l’analyse des sentiments ou la fouille d’opinions, on présentant les informations essentielles nécessaires pour avoir une idée générale sur le domaine de recherche. En mettant l’accent sur ses concepts



fondamentaux et les différentes techniques utilisées dans le domaine de la fouille d'opinion d'une manière générale.

- **Le chapitre II** est destiné à présenter le domaine de la fouille d'opinion au niveau fin de granularité avec un état de l'art sur les différents travaux réalisés par les chercheurs de la communauté pour faire face aux défis du domaine de la fouille d'opinion et spécifiquement au niveau aspect.
- **Le chapitre III** est consacré à la présentation des différents modèles basés sur l'apprentissage approfondi (Deep Learning) pour la fouille d'opinion basée sur les aspects.

**Le quatrième chapitre** est consacré à la proposition de deux contributions permettant la fouille d'opinion basée sur les aspects par une approche basée Deep Learning.

- La première contribution consiste à présenter une étude comparative et un état synthèse des différents travaux présenté dans le chapitre 02.
- La deuxième contribution est une présentation d'une approche Deep learning basée sur les réseaux de neurone de type mémoire à long-court terme (LSTM) pour l'extraction et la classification des textes porteurs d'opinion au niveau aspect.

Les reconnaissances acquises à partir de notre étude comparative illustrée dans la première contribution, nous ont encouragées de travailler suivant la nouvelle tendance technologique tout on implémentant une approche basée sur les réseaux LSTM arborescente qui ont présenté leur performance dans différents domaines de recherche et cela est confirmé par notre approche proposée à comparer avec d'autres travaux d'état de l'art.

## 1 Chapitre I : La Fouille d'opinion : un domaine émergent

### 1.1 Introduction

Au cours des dernières décennies, la plupart des gens devaient demander à leurs amis ou se tourner vers les responsables de magazines pour décider quels produits valaient la peine d'être achetés. Mais avec l'essor du Web 2.0 (Graham, 2005; O'Reilly, 2005) et du commerce électronique, les modes d'achat et de vente ont considérablement changé, où leurs utilisateurs sont devenus capables d'exprimer des critiques ou des commentaires sur les produits qu'ils ont achetés. Plusieurs rapports et enquêtes ont prouvé que plus de 5000 procédures d'achat sont faites en ligne (Stevens, 2016) sur Yelp<sup>8</sup> et Amazon<sup>9</sup> sites web et selon le sondage du PewResearch Center<sup>10</sup> réalisé en 2016 : 82 % des personnes de nationalité américaine interrogées ont déclaré consulter les commentaires et les critiques en ligne lorsqu'elles achètent un produit pour la première fois (Smith & Anderson, 2016). L'information provenant de ces commentaires est importante pour aider d'autres utilisateurs potentiels à prendre une décision d'acheter ou non un produit en se basant sur l'expérience et les opinions des utilisateurs précédent au sujet d'un article particulier. Avec l'augmentation du nombre d'utilisateurs, le volume de contenu brut généré par ces utilisateurs va augmenter rapidement au fil du temps avec une taille énorme ce qui rend l'accès, la lecture et le traitement de cette quantité de données très difficile, **donc pourquoi pas un processus d'analyse d'opinion pour acquérir facilement des informations utiles à partir d'une grande masse de commentaire, ce qui aide les utilisateurs à prendre une décision dans les meilleurs délais avec moins d'effort en se basant sur les opinions des autres ?** Une telle question a permis l'apparition d'un émergent et moderne domaine de recherche nommé fouille d'opinion [FO] ou analyse des sentiments [SA].

Ce chapitre présente un aperçu général sur le domaine de la fouille d'opinions et offre une historique perspective sur la manière dont ce domaine est apparu, puis on se mettant l'accent sur ses concepts de base, ses taches, ses domaines d'application et finalement on donnera un aperçu des tâches successives nécessaires à la construction d'un système de fouille d'opinions.

---

<sup>8</sup><https://www.Yelp.com/>

<sup>9</sup><https://www.amazon.fr/>

<sup>10</sup><http://www.pewresearch.org/>

## 1.2 Définitions d'une opinion

### 1.2.1 Définition 1

Selon le dictionnaire Larousse en ligne : avant d'approfondir dans les différents concepts du domaine de la fouille d'opinion, concentrons-nous d'abord sur la notion clé "opinion". Selon le dictionnaire Larousse, une opinion signifie :

- Un « jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense ». La définition positionne l'opinion au niveau de l'individu / adresser une question à une personne (une opinion individuelle). Exemple : Quelle est votre opinion sur ce film ?
- « L'ensemble des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc. ». La définition évoque une opinion collective et partagée (une opinion collective). Exemple : Quelles sont tes opinions politiques ?

### 1.2.2 Définition 2

Selon Lui. Bing (B. Liu, 2012), une opinion est une déclaration subjective présentée par le quintuple  $(\mathbf{e}_i, \mathbf{a}_{ij}, \mathbf{s}_{ijkl}, \mathbf{h}_k, \mathbf{t}_l)$  où :

1.  $\mathbf{e}_i$  est le nom d'une entité,
2.  $\mathbf{a}_{ij}$  est un aspect de  $\mathbf{e}_i$ ,
3.  $\mathbf{s}_{ijkl}$  est le sentiment sur l'aspect  $\mathbf{a}_{ij}$  de l'entité  $\mathbf{e}_i$  exprimé par le détenteur  $\mathbf{h}_k$  en temps  $\mathbf{t}_l$
4.  $\mathbf{h}_k$  est le détenteur de l'opinion.
5.  $\mathbf{t}_l$  est le moment où l'opinion est exprimée par  $\mathbf{h}_k$ .
6.  $\mathbf{e}_i$  et  $\mathbf{a}_{ij}$  ensemble présente l'objet cible.

Le sentiment  $\mathbf{s}_{ijkl}$  est positif, négatif ou neutre, ou est exprimé avec différents niveaux d'intensité, par exemple 1 à 5 étoiles, tels qu'utilisés par la plupart des commentaires sur le Web.

Les aspects  $\mathbf{a}_{ij}$ s sont également appelés caractéristiques (comme dans les caractéristiques du produit). Cependant, les caractéristiques ici peuvent être confondues avec les caractéristiques utilisées dans l'apprentissage automatique, où une caractéristique signifie un attribut de données. Pour éviter toute confusion, le mot aspect est devenu plus populaire ces dernières

années. Notez que certains chercheurs utilisent également les termes facettes, attributs ou sujets, et dans des applications spécifiques, les entités et les aspects peuvent également être appelés autrement en fonction du domaine d'application.

Cette définition fournit une plateforme pour transformer le texte non structuré en données structurées. Le quintuple ci-dessus est fondamentalement un schéma de base de données, sur lequel les opinions extraites peuvent être placées dans une table de base de données. Ensuite, une analyse qualitative, quantitative des opinions peut être effectuée à l'aide d'un système de gestion de base de données.

L'opinion définie dans cette section n'est qu'un type d'opinion, appelé une opinion régulière, Pour simplifier, nous l'appelons simplement opinions. Un autre type est l'opinion comparative (Jindal & Liu, 2006b; B. Liu, 2007, 2012), qui nécessite une définition différente. La section suivante présentera les différents types d'opinions.

### 1.3 Différents types d'opinion

Nous pouvons également classer les opinions en fonction de la manière dont elles sont exprimées dans un texte, une opinion explicite et une opinion implicite.

#### 1.3.1 Opinion régulière versus opinion comparative

Une opinion peut prendre plusieurs images et peut être affectée à l'un des groupes suivants :

- **Opinion régulière** : une opinion régulière est souvent qualifiée dans la littérature comme étant une opinion standard et elle comporte deux sous-types principaux (B. Liu, 2012) :
  - *Opinion directe* : Une opinion directe fait référence à une opinion exprimée directement sur une entité ou sur un aspect de celle-ci, par exemple "La résolution de l'écran de la tablette est formidable".
  - *Opinion indirecte* : Une opinion indirecte est une opinion exprimée indirectement sur une entité ou un aspect de cette entité sur la base de ses effets sur certaines autres entités. Ce sous-type d'opinion se produit souvent dans le domaine médical. Exemple " après avoir pris le médicament, j'ai perdu mon équilibre", cela décrit un effet non souhaitable de la prise du médicament ce qui provoque indirectement une opinion négative envers le médicament.

- **Opinion comparative** : d'après (Jindal & Liu, 2006a, 2006b), une opinion comparative exprime une relation de similitude ou de différences entre deux ou plusieurs entités et / ou une préférence du détenteur d'opinion basée sur certains aspects communs des entités. Une opinion comparative est généralement exprimée en utilisant la forme comparative d'un adjectif ou d'un adverbe, exemple : " l'analyse au niveau d'aspect est plus difficile qu'au niveau document" ou superlative, exemple : "l'apprentissage approfondi est la meilleure technique de traitement d'image".

### 1.3.2 L'opinion implicite versus explicite

Les opinions peuvent être également classées en fonction de la manière dont elles sont exprimées dans un texte (B. Liu, 2012) en :

- **Opinion explicite** : Une opinion explicite est une déclaration subjective qui donne une opinion régulière ou comparative.
- **Opinion implicite** : Un avis implicite est un énoncé objectif cela implique une opinion régulière ou comparative. Une telle déclaration objective exprime généralement un fait souhaitable ou indésirable. Exemple " le PC portable est assez léger pour tenir toute la journée sans encombre", on parle de l'aspect poids, mais cette fois-ci aucun mot explicite n'a été utilisé pour exprimer cet aspect.

D'après les définitions et les exemples présentés, nous constatons que les opinions explicites sont plus faciles à détecter et à classer que les opinions implicites. Cela justifie le grand nombre des travaux de recherche actuelle focalisés sur des opinions explicites à comparer avec les travaux qui ont été faits sur les opinions implicites (Zhang & Liu, 2011).

## 1.4 Définition de la fouille d'opinion

La fouille d'opinion est également appelée analyse de sentiment, extraction d'opinion, analyse de subjectivité, analyse d'affect, analyse d'émotions, etc., développée à la fin des années 1990 (Hatzivassiloglou & McKeown, 1997). Cependant, ces nominations relèvent toutes de l'analyse du sentiment ou de l'analyse d'opinions (B. Liu, 2012). Tandis que dans l'industrie, le terme analyse de sentiment est plus communément utilisé, mais dans le monde académique, on utilise fréquemment la fouille d'opinion ou l'analyse des sentiments (B. Liu, 2012). Les termes analysent des sentiments et fouille d'opinion ont été publiés pour la première fois en 2003 respectivement dans (Dave, Lawrence, & Pennock, 2003; Nasukawa & Yi, 2003). Cependant, la recherche sur les sentiments et les opinions est

apparue plus tôt dans (Das & Chen, 2001; Morinaga, Yamanishi, Tateishi, & Fukushima, 2002; Pang, Lee, & Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe & Mihalcea, 2006). Ils représentent essentiellement le même domaine d'étude. Selon l'intitulé de la thèse nous allons adopter le terme 'fouille d'opinion'. Deux définitions ont été attribuées par les premiers chercheurs de la communauté de la fouille d'opinion, la première est celle donnée par Bing (B. Liu, 2012) où il a défini la fouille d'opinion comme étant une discipline récente qui vise à reconnaître les opinions, les jugements, les sentiments, les évaluations, les appréciations, les attitudes et les émotions envers des entités telles que des produits, des services, des organisations, des individus, des événements et des attributs. C'est le domaine de la recherche qui génère des systèmes automatiques permettant d'extraire les opinions humaines à partir des textes écrits en langage naturel. Tandis que, la deuxième définition de (Pang & Lee, 2008), présente le domaine de la fouille d'opinion comme étant le traitement informatique de l'opinion, du sentiment et de la subjectivité dans le texte. La fouille d'opinion est l'intersection de plusieurs spécialités de l'informatique telles que l'apprentissage automatique, le traitement automatique du langage naturel, la fouille de texte et la recherche d'information.

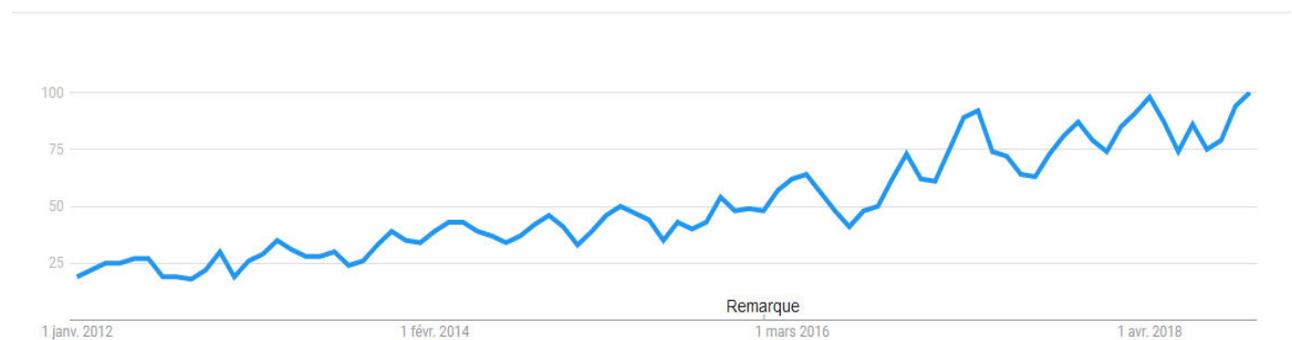
Après l'an 2000, le domaine de la fouille d'opinion a suscité une attention accrue par les chercheurs de sa communauté, notamment en raison de l'augmentation du nombre de publications scientifiques présentées (Cardie, Wiebe, Wilson, & Litman, 2003; Dave et al., 2003; Dini & Mazzini, 2002; Morinaga et al., 2002; Nasukawa & Yi, 2003; Pang et al., 2002) qui avaient un impact important sur la communauté du traitement de la langue.

Parmi les facteurs ayant influencé positivement sur la croissance du secteur de la recherche sur l'opinion publique au cours des dernières années, notons :

- Les progrès du traitement de la langue naturelle ont permis de développer des tâches plus complexes telles que l'analyse et le traitement syntaxique.
- Le progrès des méthodes d'apprentissage automatique, en particulier celles ayant un impact important sur le traitement de la langue naturelle, telles que les travaux effectués dans le domaine de la classification des textes.
- La disponibilité des lexiques créés spécialement pour la fouille d'opinion ainsi que de corpus annotés pour l'apprentissage et le test des algorithmes d'apprentissage automatique.
- Les nombreux sites Web de commerce électronique qui fournissent des évaluations d'utilisateur constituent une source importante de construction automatique de corpus d'apprentissage et de test pour la classification des opinions.

- L'apparition de compétitions scientifiques internationales fournissant un cadre unifié dans lequel les chercheurs peuvent évaluer leurs systèmes de fouille d'opinion. SemEval<sup>11</sup> est l'un des exemples les plus remarquables. Il a introduit les tâches de détection depuis 2007.

Google Trends<sup>12</sup> fournit un aperçu pertinent des mentions d'un terme ou d'une expression à partir de 2004 dans les requêtes de recherche ainsi que dans les articles de presse. La figure 1.1 montre l'intérêt croissant suscité par la fouille d'opinions en dehors du milieu universitaire. Bien que l'expression « fouille d'opinions », principalement utilisée dans le monde universitaire, ne donne aucun résultat dans Google Trends, l'expression « analyse des sentiments » s'avère être très populaire dans les requêtes générales et les mentions d'actualité. Nous pouvons également observer à partir de la figure 1.1 une augmentation constante de l'intérêt apporté à l'analyse des sentiments au cours des dernières années.



**Figure 1.1** : Les données de Google Trends sur les tendances liées à l'analyse des sentiments<sup>13</sup>

### 1.5 Les niveaux d'analyse d'opinion

Comme c'est déjà mentionné, La fouille d'opinion est un domaine de recherche récente, où son but principal est d'introduire des outils informatiques pour l'extraction des opinions à partir des textes écrits en langage naturel. En général, Les chercheurs ont étudié l'analyse ou la fouille d'opinion à trois niveaux principaux de granularité : document, phrase et aspect. Dans ce qui suit nous introduisons ces trois niveaux de granularité :

---

<sup>11</sup> <http://nlp.cs.swarthmore.edu/semeval/>

<sup>12</sup> <http://www.google.ro/trends/>

<sup>13</sup> Google trends : outils en phase de développement fournis par Google, pour analyser les tendances de recherche autour des mots clés.

### 1.5.1 Niveau document

A ce niveau d'analyse, un système de fouille d'opinion est censé de déterminer si le document porteur d'opinion exprime un sentiment globalement positif ou négatif sur un produit (Pang et al., 2002; Turney, 2002). Cette tâche est appelée classification des sentiments au niveau du document, elle suppose que chaque document exprime des opinions sur une seule entité (par exemple, un seul produit). Ainsi, cela ne s'applique pas aux documents qui comparent plusieurs entités.

### 1.5.2 Niveau phrase

Ce niveau d'analyse est étroitement lié à la classification de subjectivité (Wiebe & Mihalcea, 2006) qui distingue les phrases objectives qui expriment des informations factuelles des phrases subjectives qui expriment des points de vues et des opinions subjectives, la tâche à ce niveau concerne les phrases et détermine si chaque phrase exprime une opinion positive, négative ou neutre. Neutre, ne signifie généralement pas d'opinion. A ce niveau d'analyse une phrase est considérée comme étant un court document.

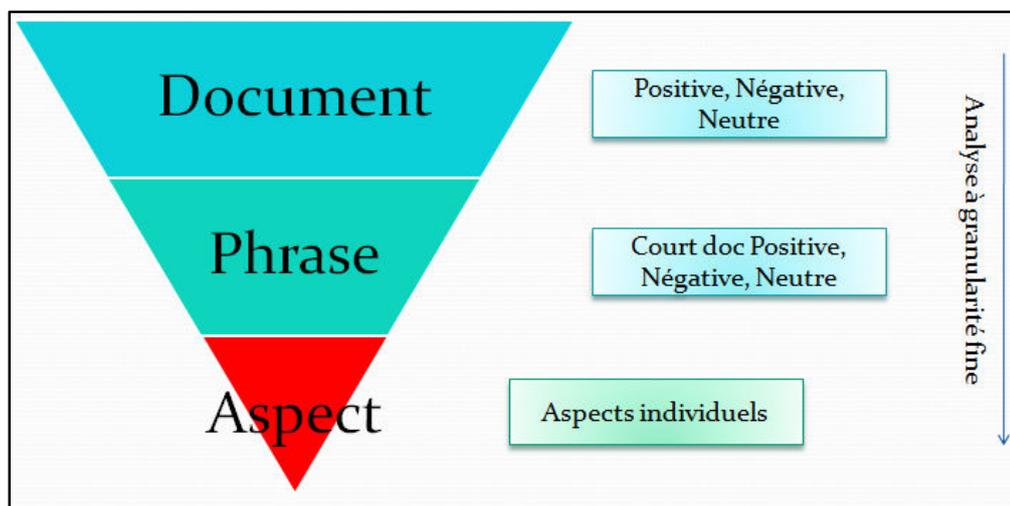


Figure 1.2: Les niveaux d'analyse dans le domaine de la fouille d'opinion (B. Liu, 2012).

### 1.5.3 Niveau aspect

Les analyses au niveau du document et au niveau de la phrase ne permettent pas de découvrir ce que les gens préfèrent exactement et ce qu'ils détestent par rapport au niveau



d'aspect qui effectue une analyse plus fine. Ce niveau était précédemment appelé le niveau de caractéristiques (exploration et synthèse d'opinion basées sur les caractéristiques) (M. Hu & Liu, 2004), ce niveau d'analyse sera présenté d'une manière détaillée dans le chapitre II.

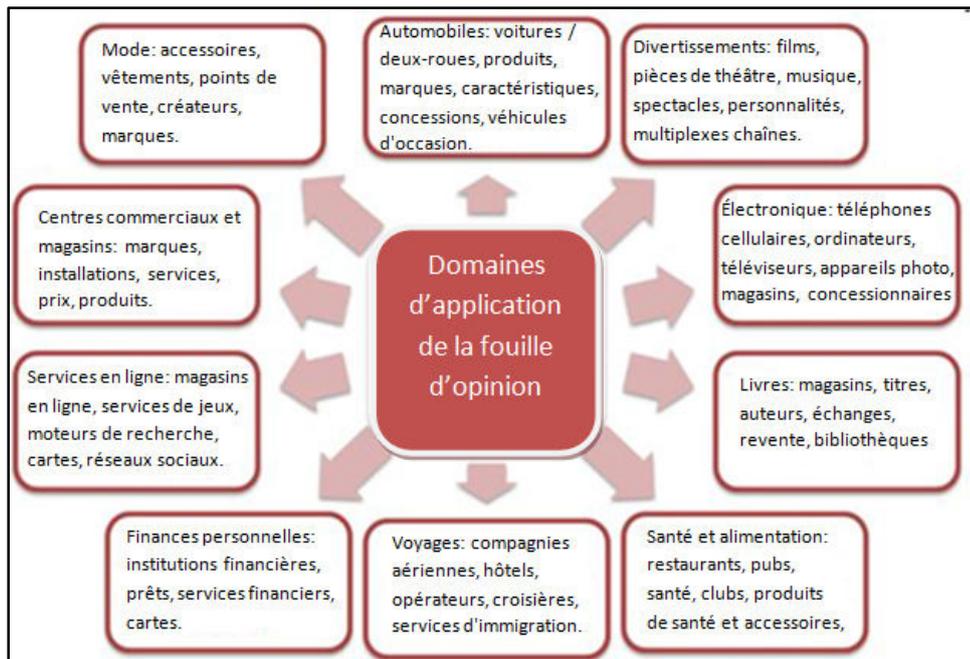
Au lieu de se baser sur les composants de base du langage (documents, paragraphes, phrases, clauses ou expressions), le niveau d'aspect vise directement l'opinion elle-même. Il est basé sur l'idée qu'une opinion est composée d'un sentiment positif ou négatif et d'une cible d'opinion. Pour mieux comprendre le problème de l'analyse des sentiments, les chercheurs à ce niveau de granularité voient qu'une opinion présentée sans que sa cible soit identifiée est d'une utilité limitée. Dans de nombreuses applications, les cibles d'opinion sont décrites par des entités et / ou par leurs différents aspects.

Ainsi, le but de ce niveau d'analyse est de découvrir des opinions ou des sentiments envers des entités et / ou leurs aspects. Exemple : la phrase "La qualité des appels du téléphone est bonne, mais la durée de vie de sa batterie est courte" présente deux sentiments différents exprimés sur deux aspects, la qualité de l'appel et la durée de vie de la batterie, du téléphone (entité générale). L'opinion exprimée sur la qualité d'appel du téléphone est positive, mais celle sur la durée de vie de la batterie est négative. Donc, ici la qualité des appels et la durée de vie de la batterie du téléphone sont les cibles de l'opinion et présentent des aspects de l'entité globale téléphone.

### **1.6 Domaine d'application de la fouille d'opinion**

Dans le monde réel, les opinions des autres sont le noyau de presque toutes les activités humaines, car ce sont elles qui influencent sur nos décisions. Chaque fois que nous devons prendre une décision, nous voulons connaître l'opinion des autres, les entreprises et les organisations veulent toujours connaître l'opinion des consommateurs ou du public sur leurs produits et services. Les consommateurs individuels souhaitent également connaître l'opinion des utilisateurs existants d'un produit avant de l'acheter, ainsi que celle des autres sur les candidats politiques avant de prendre une décision de vote lors d'une élection politique.

Les applications d'analyse des opinions se sont étendues à presque tous les domaines possibles : produits de consommation, services de santé, services financiers, événements sociaux et élections politiques, l'électronique, les services d'émigration, etc.



**Figure 1.3** : Domaine d'application de la fouille d'opinion (B. Liu, 2012).

En outre, de nombreuses recherches ont également été publiées sur des cas réels d'application de la fouille d'opinion. Par exemple, dans (J. Liu, Cao, Lin, Huang, & Zhou, 2007), un modèle basé sur la détection d'opinion était pour prédire la performance des ventes. Dans (Tumasjan, Sprenger, Sandner, & Welpe, 2010), le sentiment de Twitter a également été appliqué pour prédire les résultats des élections. Dans (Asur & Huberman, 2010; Joshi, Das, Gimpel, & Smith, 2010; Sadikov, Parameswaran, & Venetis, 2009), des données Twitter, des critiques de films et des blogs ont été utilisés pour prédire les revenus des salles de cinéma. Dans (S. M. Mohammad & Yang, 2011), les sentiments exprimés dans les courriers ont été utilisés pour déterminer la différence entre féminin et le masculin sur des axes émotionnels.

### 1.7 Les différentes tâches de la fouille d'opinion

De nos jours, les progrès technologiques actuels permettent de stocker et de récupérer d'une manière efficace une grande masse de données à partir des sources brutes sous forme d'expressions en langage naturel qui peuvent être facilement rapportées au moyen de courts messages texte, créant rapidement un contenu unique de grandes dimensions qui doivent être traités en temps réel de manière efficace afin de créer des connaissances exploitables pour la prise de décision. Dans ce sens, les chercheurs de la communauté de la fouille d'opinion ont contribué dans différent axes tel que : la

détection de la polarité, détection de la subjectivité, résumé d'opinion, détection de sarcasme, classification des sentiments basées sur les aspects, etc. La figure 1.4 présente les différentes tâches de la fouille d'opinion, dans cette thèse nous nous intéressons à ce qui est noté en rouge.

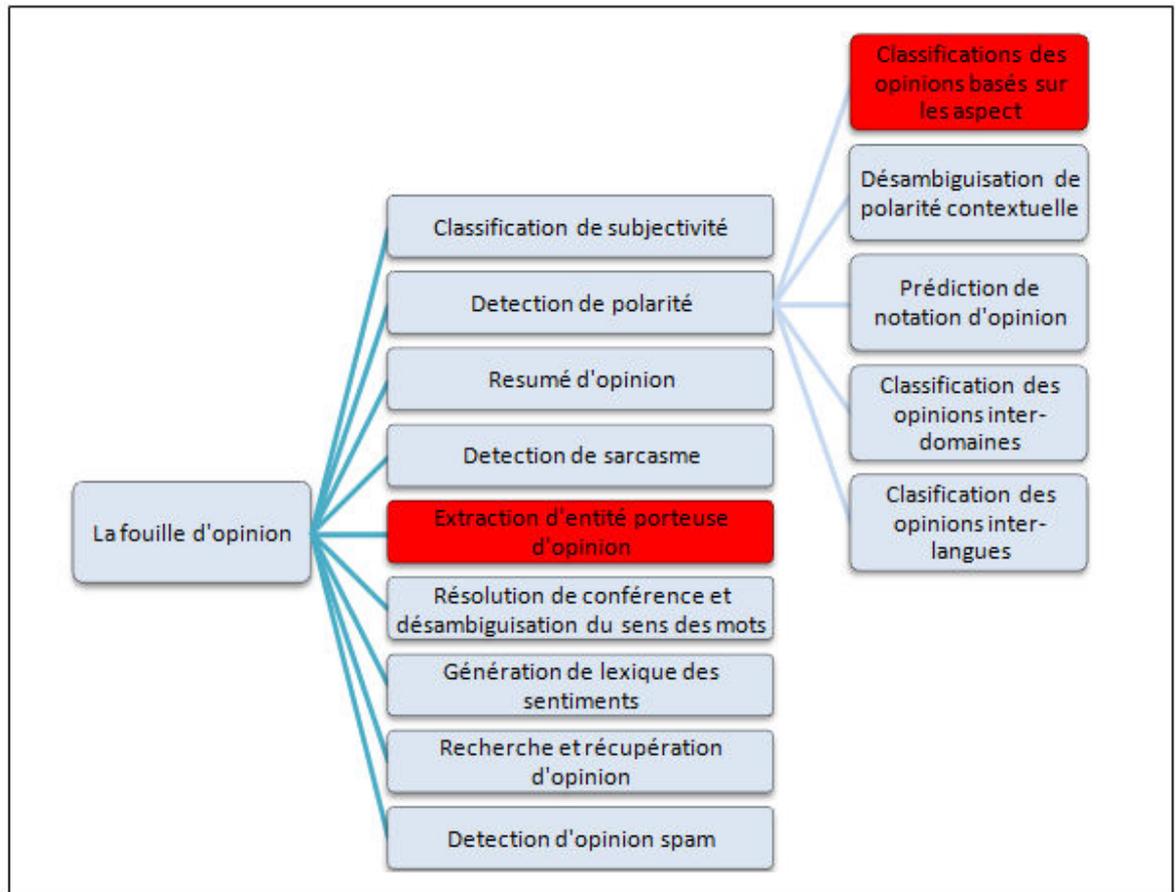
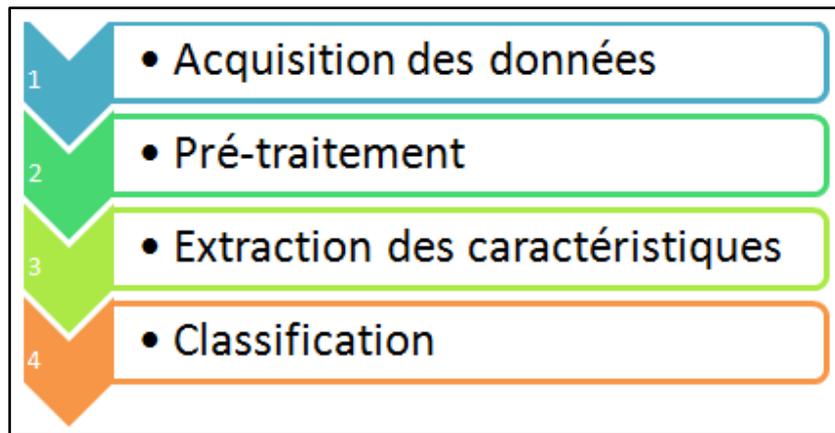


Figure 1.4: Les différentes tâches de la fouille d'opinion (Pozzi, Fersini, Messina, & Liu, 2017).

## 1.8 Processus de classification des opinions

La classification d'opinions peut être considérée comme un processus de plusieurs étapes qui prend en entrée un ensemble de texte portant sur différents sujets et bien sûr porteurs d'opinions sur une entité donnée et fournit en sortie un résumé agrégeant les opinions exprimées dans le texte vis-à-vis de la cible ou éventuellement vis-à-vis des aspects de l'entité (Dave et al., 2003). L'aspect d'une entité comme nous l'avons déjà mentionné est une caractéristique, un attribut ou un élément composant de l'entité.

On peut ainsi découper le processus de fouille d'opinions en étapes principales, illustrées dans la figure 1.5 :



**Figure 1.5** : Le processus général de la fouille d'opinion (B. Liu, 2012).

### 1.8.1 Acquisition des données

La fouille d'opinion est un domaine de recherche qui fait appel à diverses techniques issues du traitement du langage naturel (NLP) et de fouille de données (data mining (DM) en Anglais). Il manipule des données proviennent de plusieurs ressources sous différents formats qu'on les récupère depuis internet (forum et blogs) ou à partir des compagnes d'évaluations qui sont non structurées ce qui nécessitent un prétraitement préalable, donc différentes étapes sont nécessaires pour émettre des avis à partir de ces données, où l'acquisition et le prétraitement de ces données sont les étapes les plus importantes pour l'exploration et l'analyse de ces données (B. Liu, 2012).

### 1.8.2 Prétraitement des données

Le prétraitement des données a deux objectifs principaux. Dans un premier temps, il sert à réduire la taille du vocabulaire, et donc diminuer la taille de l'espace de représentation. Dans un deuxième temps, ce prétraitement permet d'améliorer les performances du système de fouille des opinions. Donc, des techniques de retraitement sont nécessaires pour acquérir une base de données plus propre dans le but d'augmenter les performances du système de classification des opinions. Plusieurs techniques simples et utiles sont incorporées dans la littérature (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011), ces techniques sont spécifiques à la langue citons parmi elles :

- La lemmatisation, qui consiste à remplacer chaque mot du texte par sa forme canonique conventionnelle (lemme).

- La suppression de certains mots ou caractères ne véhiculant pas ou peu d'information peut également permettre d'atteindre les deux objectifs.
- La suppression des mots inutiles comme les mots vides, la ponctuation, les chiffres, les symboles ou encore les articles indéfinis soit retirés du corpus.
- La correction des erreurs de type simple comme les fautes d'orthographe et les lettres répétées on se basant sur un dictionnaire dédié à langue du corpus de données à traiter.
- L'utilisation des anti-dictionnaires (stop-list en anglais). Ces anti-dictionnaires sont des lexiques contenant des mots non nécessaires pour la fouille d'opinion, car ils ne sont pas porteurs de sémantique.
- Le stemming, est appliqué pour diminuer la grande diversité des mots. Elle est une technique pour réduire les conjugaisons des verbes à leur forme racine originale.

D'autres prétraitements peuvent être utiles (Agarwal et al., 2011; B. Liu, 2012), comme l'étiquetage grammatical des mots par leur catégorie, afin d'ajouter de l'information supplémentaire. Ce type de prétraitements enrichit la représentation au lieu de la simplifier.

### 1.8.3 Extraction des caractéristiques

L'analyse des sentiments est considérée comme un problème de classification des opinions exprimées dans des documents textuels qui est généralement représenté sous la forme dite « sac de mots ». Cette représentation consiste à sélectionner des éléments de vocabulaire représentant les dimensions d'un espace vectoriel, et de représenter chaque document par un vecteur dans cet espace. Donc la première étape du problème consiste à extraire et à sélectionner des caractéristiques de texte et de construire le vocabulaire de l'espace vectoriel (Agarwal et al., 2011; B. Liu, 2012). Dans la littérature, différentes techniques ont été implémentées. Dans ce qui suit, nous citons parmi eux celle qui ont été fréquemment utilisées par les chercheurs de la communauté de la fouille d'opinion.

#### A. La représentation

Les documents textuels sont des suites de mots. Afin d'en permettre l'exploitation, il est nécessaire de représenter les textes numériquement. Les mots sont alors considérés comme des objets distincts non ordonnés. Cette représentation consiste tout d'abord à sélectionner des variables représentant les dimensions d'un vecteur puis à représenter chaque document sur ce

vecteur. En classification de texte, la représentation implique donc une étape préliminaire appelée segmentation, qui consiste à découper le texte afin de sélectionner les variables. Une fois les variables acquises, plusieurs documents peuvent être représentés sur le même espace vectoriel. Nous présentons tout d'abord en quoi consiste exactement la segmentation qui consiste à découper l'espace de caractères afin d'obtenir un vocabulaire. On peut considérer les marques de ponctuation, afin de découper le document en phrases, ou encore les espaces associés à la ponctuation pour obtenir les mots. Ce sont d'ailleurs les mots qui sont les variables les plus utilisées en classification d'opinion, mais d'autres choix existent. On peut par exemple limiter la taille des variables à un nombre  $n$  de caractères, on parle alors de  $n$ -grammes de lettres. Les  $n$ -grammes peuvent également être formés de mots selon leur ordre dans la phrase. Une fois le vocabulaire sélectionné, nous pouvons choisir l'une des techniques de présentation les plus courantes suivantes :

### **B. Présence et fréquence des termes**

Ces caractéristiques sont des mots individuels ou des  $n$ -grammes de mots et leur nombre de fréquences. Il attribue soit la pondération binaire aux mots (zéro si le mot apparaît, soit un autre sinon), soit l'utilisation de pondérations de fréquence des termes pour indiquer l'importance relative des caractéristiques (Mejova & Srinivasan, 2011).

### **C. Parties du discours (POS)**

La partie du discours (POS) de chaque mot peut également être importante. Les mots de différentes parties du discours (POS) peuvent être traités différemment. Par exemple, il a été démontré que les adjectifs sont des indicateurs importants des avis. Ainsi, certains chercheurs ont traité les adjectifs comme des caractéristiques spéciales. Cependant, on peut aussi utiliser toutes les étiquettes POS et leurs  $n$ -grammes comme caractéristiques (B. Liu, 2012).

### **D. Mots et expressions d'opinion**

Ce sont des mots couramment utilisés pour exprimer des opinions, bonnes ou mauvaises, similaires ou différentes. Par ailleurs, certaines phrases expriment des opinions sans utiliser de mots d'opinion. Par exemple : me coûte.

### **E. Négations**

L'apparition de mots négatifs peut changer l'orientation de l'opinion, exemple : car pas bon équivaut à mal. Cette caractéristique est considérée comme l'une des caractéristiques

importantes car elle modifie complètement le sens de la phrase (Agarwal et al., 2011; B. Liu, 2012).

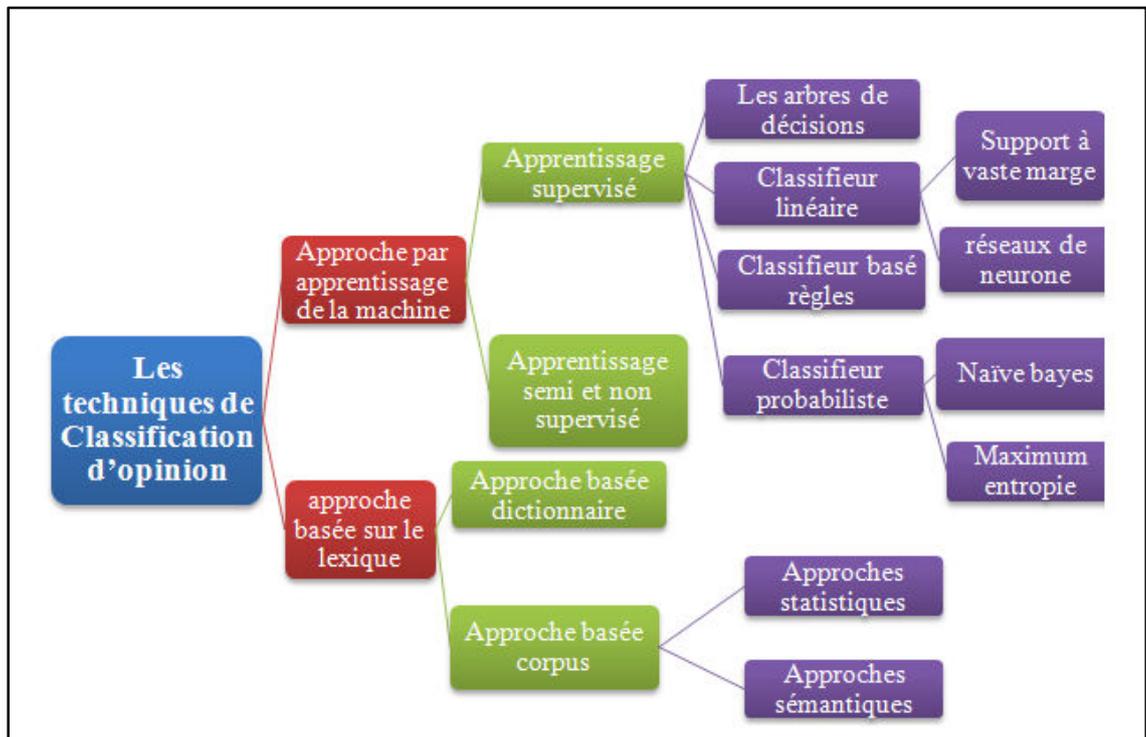
### 1.8.4 Classification d'opinion

La classification des opinions est un processus qui vise à définir la polarité positive ou négative exprimée dans un document donné. Comme nous avons mentionné précédemment, il existe trois niveaux d'analyse principaux : niveau de document, niveau de phrase et niveau d'aspect, où le niveau document vise à classer un document porteur d'opinion et considère l'ensemble du document comme une unité de base ; à l'inverse, le niveau de phrase classe le sentiment exprimé dans chaque phrase et le niveau d'aspect doit identifier le sentiment associé à chaque aspect spécifique de l'entité critiquée. Ces différents niveaux de classification sont mis en œuvre en appelant les célèbres techniques de classification des sentiments qui sont catégorisées selon (Medhat et al., 2014) en approches d'apprentissage automatique, approches basées sur le lexique et approches hybrides. Un schéma explicatif est présenté dans la figure 1.6. Les approches d'apprentissages automatiques (en anglais Machine Learning (ML)) appliquent les différents algorithmes d'apprentissage automatique tels que séparateurs à vastes marges (SVM), les naïfs bayésiens, les arbres de décision etc. L'approche basée sur le lexique est divisée en une approche basée sur un dictionnaire et une approche basée sur un corpus. Toutes ces approches sont utilisées pour trouver la polarité des sentiments. L'approche hybride combine les deux approches (Ravi & Ravi, 2015), ces différentes techniques de classification d'opinion seront détaillées dans la section suivante.

### 1.9 Les techniques de classification d'opinion

Les techniques de classification des sentiments peuvent être grossièrement divisées en une approche d'apprentissage automatique, une approche basée sur le lexique et une approche hybride (Maynard & Funk, 2011). L'approche par apprentissage de la machine applique les célèbres algorithmes apprentissage par la machine et utilise des fonctionnalités linguistiques. L'approche basée sur le lexique repose sur un lexique de sentiments, un ensemble de termes de sentiments connus et précompilés. Il est divisé en une approche basée sur un dictionnaire et une approche basée sur un corpus qui utilisent des méthodes statistiques ou sémantiques pour trouver la polarité des sentiments. L'approche hybride combine les deux approches et est très courante, les lexiques de sentiment jouant un rôle clé dans la plupart des méthodes. Les

différentes approches et les algorithmes les plus populaires pour la classification des opinions sont illustrés dans la Figure 1.6.



**Figure 1.6:** Les techniques de classification des opinions (Medhat, Hassan, & Korashy, 2014).

Les méthodes de classification de texte utilisant l'approche par apprentissage de la machine peuvent être divisées en méthodes d'apprentissage supervisé et non supervisé. Les méthodes supervisées utilisent un grand nombre de documents étiquetés. Les méthodes non supervisées sont utilisées lorsqu'il est difficile de trouver ces documents étiquetés (Medhat et al., 2014).

L'approche basée sur le lexique dépend de la recherche du lexique d'opinion utilisé pour analyser le texte. Il existe deux méthodes dans cette approche. L'approche basée sur le dictionnaire qui consiste à rechercher des mots-clés d'opinion, puis à rechercher dans le dictionnaire leurs synonymes et antonymes. L'approche basée sur un corpus commence par une liste de base (seed list en anglais) de mots d'opinion, puis trouve d'autres mots d'opinion dans un grand corpus pour aider à trouver des mots d'opinion avec des orientations spécifiques au contexte. Cela pourrait être fait en utilisant des méthodes statistiques ou sémantiques. Dans



ce qui suit, une explication approfondie des différentes approches mentionnées dans la figure 1.6

### 1.9.1 Les approches par apprentissages de la machine

Les approches d'apprentissage automatique reposent sur les célèbres algorithmes d'apprentissage automatique pour résoudre le problème de la fouille d'opinion en tant que problème classique de classification de texte utilisant des caractéristiques syntaxiques et / ou linguistiques.

#### 1.9.1.1 Approches par apprentissage supervisé

Les méthodes d'apprentissage supervisé dépendent de l'existence de documents d'apprentissage étiquetés. Il existe de nombreux types de classifieurs supervisés dans la littérature (Medhat et al., 2014). Dans les sous-sections suivantes, nous présentons brièvement les classifieurs les fréquemment utilisés dans la fouille d'opinion.

##### A. Classifieur probabiliste

Les classifieurs probabilistes utilisent des modèles mixtes pour la classification. Ces modèles supposent que chaque classe est un composant mixte. Chaque composant est un modèle génératif qui fournit la probabilité d'échantillonnage d'un terme particulier pour ce composant. Ces types de classifieurs sont également appelés classificateurs génératifs (Medhat et al., 2014). Deux types de classifieurs probabilistes sont abordés dans les sous-sections suivantes.

- *Naïve bayes NB*

Le classifieur Naïve Bayes (Tan, Steinbach, & Kumar) est le classifieur le plus simple et le plus couramment utilisé. Le modèle de classification de Naïve Bayes calcule la probabilité postérieure d'une classe, en fonction de la distribution des mots dans le document. Le modèle fonctionne avec l'extraction de la caractéristique type sac de mot, qui ignore la position du mot dans le document. Il utilise le théorème de Bayes pour prédire la probabilité qu'un

ensemble de caractéristiques donné appartient à une étiquette particulière.  
 Cette probabilité est définie comme suit :

$$P(\text{étiquette}|\text{caractéristiques}) = \frac{P(\text{étiquette}) * P(\text{caractéristiques}/\text{étiquette})}{P(\text{caractéristiques})} \dots \dots \dots (01)$$

Ou :

**P (étiquette) :** est la probabilité antérieure d'une étiquette ou la probabilité qu'une caractéristique aléatoire définisse l'étiquette.

**P (caractéristiques | étiquette) :** est la probabilité a priori qu'un ensemble de caractéristiques donné soit classé comme une étiquette.

**P (caractéristiques) :** est la probabilité antérieure qu'un ensemble de caractéristiques donné se soit produit.

- **Maximum entropie ME**

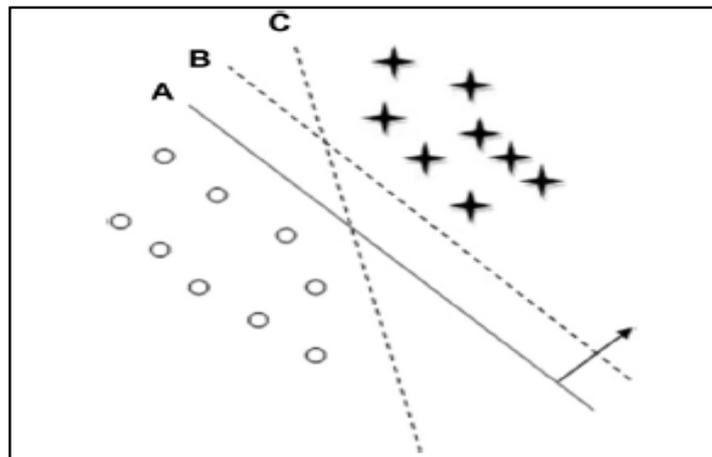
Le classifieur Maximum entropie (Medhat et al., 2014) (connu sous le nom de classifieur exponentiel conditionnel) convertit les ensembles de caractéristiques étiquetés en vecteurs à l'aide du codage. Ce vecteur codé est ensuite utilisé pour calculer des poids pour chaque entité, qui peuvent ensuite être combinés pour déterminer l'étiquette la plus probable pour un ensemble d'entités. Ce classifieur est paramétré par un ensemble de X (poids), qui permet de combiner les entités jointes générées à partir d'un ensemble de caractéristiques par un X (encodage). En particulier, l'encodage mappe chaque paire C (ensemble de caractéristiques, label) sur un vecteur (Tan et al.). La probabilité de chaque étiquette est ensuite calculée à l'aide de l'équation suivante :

$$P(\text{fs}/\text{label}) = \frac{\text{dotprod}(\text{weights}, \text{encode}(\text{fs}, \text{label}))}{\text{sum}(\text{dotprod}(\text{weights}, \text{encode}(\text{fs}, \text{l})) \text{ for } \text{l} \text{ in } \text{labels})} \dots \dots \dots (02)$$

### B. Classifieur linéaire

Étant donné que  $X = \{x_1 \dots \dots \dots x_n\}$  est la fréquence normalisée des mots du document, le vecteur  $A = \{a_1 \dots \dots \dots a_n\}$  est un vecteur de coefficients linéaires ayant la même dimensionnalité que l'espace de fonctions, et  $b$  est un scalaire ; la sortie du prédicateur linéaire est définie par  $p = A.X + b$ , qui est la sortie du classifieur linéaire. Le prédicateur  $p$  est un hyperplan de séparation entre différentes classes. Il existe de nombreux types de classificateurs linéaires ; parmi eux se trouvent les machines à vecteurs de support (SVM) (Cortes & Vapnik, 1995; Vapnik, 1995), une forme de classificateurs qui tentent de déterminer les bons séparateurs linéaires entre différentes classes. Deux des plus célèbres classificateurs linéaires sont décrits dans la suite sous-sections.

- **Support à vaste marge SVM :** Le principe de base des SVM est de déterminer dans l'espace de recherche des séparateurs linéaires capables de séparer au mieux les différentes classes. Sur la figure ci-après, il y a 2 classes  $x$ ,  $o$  et il y a 3 hyperplans A, B et C. L'hyperplan A fournit la meilleure séparation entre les classes, parce que la distance normale de l'un des points de données est la plus grande, elle représente donc la marge de séparation maximale (Cortes & Vapnik, 1995).



**Figure 1.7:** Support à vaste marge dans un problème de classification (Cortes & Vapnik, 1995).

Les données textuelles conviennent parfaitement à la classification SVM en raison de la nature du texte, dans laquelle peu de caractéristiques sont sans importance, mais elles tendent à être corrélées les unes aux autres et généralement organisées en catégories linéairement séparables (Joachims, 1996). SVM peut construire une surface de décision non linéaire dans l'espace de caractéristiques d'origine en mappant les instances de données de manière non linéaire sur un espace de produit interne où les classes peuvent être séparées linéairement avec un hyperplan (Aizerman, 1964).

- **Réseaux de neurones :** Le réseau de neurones est constitué de nombreux neurones, le neurone constituant son unité de base. Les entrées des neurones sont désignées par le vecteur  $X_i$ , qui est la fréquence des mots dans le document. Il existe un ensemble de poids  $A$  qui sont associés à chaque neurone utilisé afin de calculer une fonction de ses entrées  $f()$ . La fonction linéaire du réseau de neurones est la suivante :  $p_i A \cdot \bar{X}_i$ . Dans un problème de classification binaire, on suppose que l'étiquette de classe de  $X_i$  est désignée par  $y_i$  et que le signe de la fonction prédite  $p_i$  donne l'étiquette de classe. Les réseaux de neurones multicouches sont utilisés pour les limites non linéaires. Ces multiples couches sont utilisées pour induire de multiples limites linéaires, utilisées pour approcher des régions fermées appartenant à une classe particulière. Les sorties des neurones dans les premières couches alimentent les neurones dans les dernières couches. Le processus d'apprentissage est plus complexe car les erreurs doivent être propagées en arrière sur différentes couches.

### C. Les arbres de décision

Le classifieur d'arbre de décision fournit une décomposition hiérarchique de l'espace de données d'apprentissage dans lequel une condition sur la valeur d'attribut est utilisée pour diviser les données (Quinlan, 1986). La condition ou prédicat est la présence ou l'absence d'un ou plusieurs mots. La division de l'espace de données est effectuée de manière récursive jusqu'à ce que les nœuds

d'extrémité contiennent un nombre minimum d'enregistrements utilisés aux fins de la classification.

- **Classifieur à base des règles :** Dans les classifieurs basés sur des règles, l'espace de données est modélisé avec un ensemble de règles. Le côté gauche représente une condition sur le jeu de caractéristiques exprimé sous forme normale disjonctive tandis que le côté droit est le libellé de la classe. Les conditions sont sur le terme présence. Il existe un certain nombre de critères pour générer des règles, la phase d'entraînement construit toutes les règles en fonction de ces critères. Les deux critères les plus communs sont le support et la confiance (B. L. W. H. Y. Ma & Liu, 1998). Le support est le nombre absolu d'instances dans le jeu de données d'apprentissage qui sont pertinentes pour la règle. La confiance fait référence à la probabilité conditionnelle que la partie droite de la règle soit satisfaite si la partie gauche est satisfaite. Certains algorithmes de règles combinées ont été proposés dans (Walaa, Ahmed, & Hoda, 2014).

#### 1.9.1.2 Approches semi / non supervisé

L'objectif principal de la classification est de classer les documents dans un certain nombre de catégories prédéfinies. Pour ce faire, un grand nombre de documents d'apprentissage étiquetés sont utilisés pour l'apprentissage supervisé, comme illustré précédemment. Dans la classification de texte, il est parfois difficile de créer ces documents de formation étiquetés, mais il est facile de collecter les documents non étiquetés. Les méthodes d'apprentissage non supervisées surmontent ces difficultés.

#### 1.9.2 Les approches basées sur le lexique

Les mots d'opinion sont utilisés dans de nombreuses tâches de classification des opinions. Les mots d'opinion positive sont utilisés pour exprimer certains états souhaités, tandis que les mots d'opinion négative sont utilisés pour exprimer des états non désirés. Il existe également des expressions d'opinion et des idiomes qui s'appellent ensemble le lexique d'opinion. Pour ce faire trois approches principales pour compiler ou rassembler la liste de mots d'opinion ont été envisagées. L'approche manuelle prend beaucoup de temps et n'est pas utilisée seule. Il est généralement associé aux deux autres approches automatisées lors de la

vérification finale pour éviter les erreurs des méthodes automatisées (B. Liu, 2007). Les deux approches automatisées sont présentées dans les sous-sections suivantes.

### 1.9.2.1 Approches basées sur dictionnaire

La stratégie principale de l'approche par dictionnaire est présentée dans (M. Hu & Liu, 2004; S.-M. Kim & Hovy, 2004). Un petit ensemble de mots d'opinion est collecté manuellement avec des orientations connues. Ensuite, cet ensemble est développé en recherchant dans les corpus bien connus WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) ou dans le dictionnaire des synonymes (S. Mohammad, Dunne, & Dorr, 2009) leurs synonymes et antonymes. Les mots récemment trouvés sont ajoutés à la liste de départ (appelée en anglais 'seed liste'), puis la prochaine itération commence. Le processus itératif s'arrête lorsque aucun nouveau mot n'est trouvé. Une fois le processus terminé, une correction manuelle peut être effectuée pour supprimer ou corriger les erreurs. L'approche basée sur le dictionnaire présente un inconvénient majeur : elle est incapable de trouver des mots d'opinion avec des orientations propres au domaine et au contexte (Qiu et al., 2010).

### 1.9.2.2 Approches basées sur corpus

Une approche basée sur le corpus aide à résoudre le problème de la recherche de mots d'opinion avec des orientations spécifiques au contexte. Ses méthodes dépendent de modèles syntaxiques ou de modèles associés à une liste de départ (seed list) de mots d'opinion pour trouver d'autres mots d'opinion dans un grand corpus. Une de ces méthodes a été représentée par (Hatzivassiloglou & McKeown, 1997).

L'utilisation de l'approche basée sur un corpus seul n'est pas aussi efficace que l'approche basée sur un dictionnaire car il est difficile de préparer un corpus d'énorme taille qui couvre tous les mots de la langue du domaine à étudier, exemple : tous les mots anglais, mais cette approche présente un avantage majeur car elle peut aider à trouver des mots d'opinion spécifiques à un domaine et à un contexte et leurs orientations à l'aide d'un corpus de domaine. L'approche basée sur un corpus est réalisée en utilisant une approche statistique ou une approche sémantique, comme illustré dans les sous-sections suivantes :

### **A- Approches statistiques**

La recherche de modèles de cooccurrence ou de mots d'opinion initiaux peut être effectuée à l'aide de techniques statistiques. Cela pourrait être fait en déduisant des polarités postérieures en utilisant la cooccurrence d'adjectifs dans un corpus (Fahmi & Klenner, 2008). Il est possible d'utiliser l'ensemble des documents indexés sur le Web comme corpus pour la construction du dictionnaire. Cela résout le problème de l'indisponibilité de certains mots si le corpus utilisé n'est pas de taille assez importante (Turney, 2002).

La polarité d'un mot peut être identifiée en étudiant la fréquence d'apparition du mot dans un grand corpus de textes annotés (Read & Carroll, 2009). Si le mot apparaît plus fréquemment parmi les textes positifs, sa polarité est positive. S'il se produit plus fréquemment parmi les textes négatifs, sa polarité est négative. Si les fréquences sont égales, le mot est neutre. Les mots d'opinion similaires apparaissent souvent ensemble dans un corpus, c'est la principale observation sur laquelle reposent les méthodes de l'état de l'art. Par conséquent, si deux mots apparaissent ensemble fréquemment dans le même contexte, ils auront probablement la même polarité. Par conséquent, la polarité d'un mot inconnu peut être déterminée en calculant la fréquence relative de cooccurrence avec un autre mot (Turney, 2002).

Les méthodes statistiques sont utilisées dans de nombreuses applications liées à l'AS. L'un d'eux détecte la manipulation des critiques en effectuant un test statistique du caractère aléatoire (N. Hu, Bose, Koh, & Liu, 2012).

### **B- Approches sémantiques**

L'approche sémantique donne directement des valeurs de sentiment et repose sur des principes différents pour calculer la similarité entre les mots. Ce principe donne des valeurs de sentiment similaires à des mots sémantiquement proches. WordNet, par exemple, fournit différents types de relations sémantiques entre les mots utilisés pour calculer les polarités de sentiment. WordNet pourrait également être utilisé pour obtenir une liste de mots de sentiment en développant de manière itérative l'ensemble initial avec des synonymes et des antonymes, puis en déterminant la polarité de sentiment pour un mot inconnu par le nombre relatif de synonymes positifs et négatifs de ce mot (S.-M. Kim & Hovy, 2004). L'approche sémantique est utilisée dans de

nombreuses applications pour construire un modèle de lexique pour la description des verbes, des noms et des adjectifs à utiliser en fouille d'opinion (Maks & Vossen, 2012).

### **1.10 Conclusion**

Les chercheurs de la communauté de notre domaine de recherche ont étudié la fouille d'opinion selon trois niveaux principaux de granularité : document, phrase et aspect. Le niveau document et celui de la phrase ne peuvent pas fournir des informations détaillées pour la prise de décision. Pour obtenir de telles informations, nous devons atteindre un niveau de granularité plus fin qui est le niveau aspect, il renvoie le sentiment associé aux aspects de l'entité cible pour trouver des paires opinion-aspect dans un texte porteur d'opinion. Cela permet une analyse plus détaillée utilisant plus d'informations fournies dans le texte. Comme la fouille d'opinion au niveau aspect constitue l'axe principal de notre recherche, nous nous concentrons dans le chapitre II sur ce niveau d'analyse en raison de son rôle clé dans notre domaine de recherche.



## 2 Chapitre II : La Fouille d'opinion basée sur les aspects

### 2.1 Introduction

Comme nous avons déjà introduit dans le premier chapitre, les chercheurs de la communauté ont étudié le domaine de la fouille d'opinion principalement selon trois niveaux d'analyse : le niveau du document, le niveau de la phrase, et le niveau de l'aspect. Les chercheurs dans le premier niveau supposent que tout le document exprime un sentiment sur un seul sujet. De toute évidence, ce n'est pas le cas dans de nombreuses situations. L'analyse au niveau phrase est basée sur une hypothèse similaire : une phrase ne devrait contenir que des sentiments sur un sujet. Dans la même phrase, il arrive souvent que plusieurs entités soient comparées ou que certains sentiments exprimant des opinions contrastées. Malgré que, l'extraction des opinions au niveau du document ou de la phrase soit utile dans de nombreux cas, ces niveaux d'information ne sont pas suffisants pour le processus de prise de décision (par exemple, l'achat du produit). Par exemple, une opinion positive sur un élément en particulier ne signifie pas que l'utilisateur aime tous les aspects de celui-ci. De même, une opinion négative ne signifie pas que l'utilisateur n'aime pas tout. Dans une revue typique, l'opinion décrit généralement les aspects positifs et négatifs de l'élément critiqué, bien que son opinion générale sur l'élément puisse être positive ou négative. En fait, les opinions au niveau du document et de la phrase ne peuvent fournir des informations détaillées pour la prise de décision. Pour obtenir de telles informations, nous devons atteindre un niveau de granularité plus fin dit le niveau aspect, qui vise à trouver des paires opinion-aspect dans un texte donné. Dans la fouille d'opinion au niveau de l'aspect, le sentiment général se réfère généralement à l'entité, et il renvoie le sentiment associé aux aspects de l'entité en cours de discussion. Cela permet une analyse plus détaillée utilisant plus d'informations fournies par le texte porteur d'opinion.

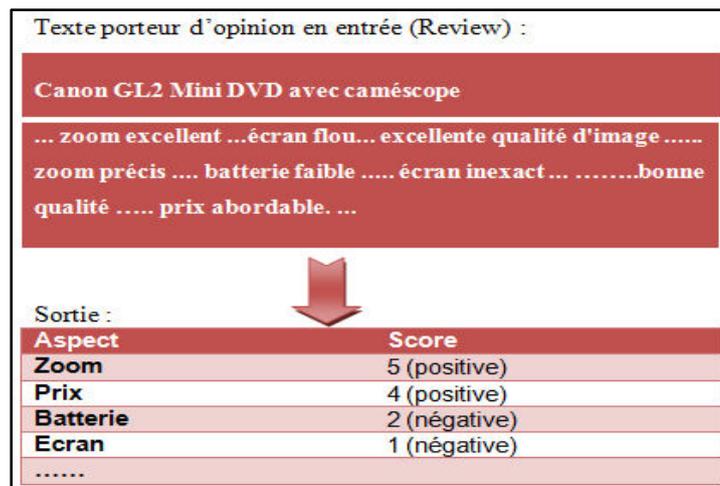
Dans ce chapitre, on va se concentrer sur la fouille d'opinion au niveau d'aspect. Nous présentons une vision éclairée du sous domaine de la fouille d'opinion par aspects. Nous définissons d'abord le problème de la fouille d'opinion basée sur les aspects de manière plus formelle avec quelques concepts de base, et sous tâches du sous domaine, Ensuite, Nous passons en revue les paramètres et mesures d'évaluation utilisés dans la littérature. Puis, nous introduisons les ensembles de jeux de données accessibles aux chercheurs et discutons leurs propriétés. En outre, nous présentons, un état de l'art exhaustif des différentes approches et travaux récents proposés par les chercheurs de la communauté de la fouille d'opinion basée aspect. À la fin de ce chapitre, nous avons abouti à présenter un aperçu général sur les approches les plus récentes pour la fouille d'opinion basée aspects pour ne

pas répéter les connaissances établies que l'on peut trouver dans d'autres travaux, et pour mettre en évidence certaines pistes de recherche prometteuses pour les recherches futures.

## **2.2 Formalisation du problème et concepts de base**

La fouille d'opinion basée sur les aspects répond aux besoins d'informations détaillées. Au cours des dernières années, plusieurs méthodes ont été proposées pour extraire des aspects à partir des textes porteurs d'opinion, ce que nous allons présenter dans ce qui suit. Certaines de ces méthodes ont utilisé des textes porteurs d'opinion d'une manière intégrale qui contient généralement une grande quantité d'informations non pertinentes, tandis que d'autres ont tiré parti des commentaires succincts. Plusieurs algorithmes ont également été présentés pour l'extraction des aspects et la prédiction de l'opinion associée. Les aspects extraits et les opinions estimées fournissent clairement des informations plus détaillées aux utilisateurs pour leur permettre de prendre des décisions et aux fournisseurs de surveiller leurs consommateurs. Alors, d'une manière formelle, le problème de la fouille d'opinion basée sur les aspects est décrit comme suit :

À partir d'un ensemble de texte porteur d'opinion (reviews) sur l'objet P, il s'agit d'identifier les k aspects majeurs de P et de prédire la polarité de chaque. La définition du problème est illustrée à la figure 2.1.



**Figure 2.1** Illustration du problème de la fouille d'opinion basé sur aspects.

Il est nécessaire de clarifier les concepts impliqués dans cette thèse car différents noms sont utilisés de manière interchangeable par de nombreuses références et ressources de données, lesquelles renvoient en réalité aux mêmes terminologies :

### 2.2.1 Le concept d'entité

Une entité est une cible qui fait l'objet d'un commentaire. Ce peut être un lieu, un service ou un produit. Dans notre cas, nous supposons que chaque revue dans l'étude de la fouille d'opinion doit avoir une cible.

### 2.2.2 Le concept d'aspect

Un aspect est un attribut ou une propriété spécifique d'une entité que les réviseurs peuvent mentionner. Par exemple, dans les avis d'hôtels, un aspect pourrait être « emplacement », « service », etc, est un concept de haut niveau et est défini séparément des données de revue. Habituellement, le vocabulaire est plutôt petit.

### 2.2.3 Le concept terme d'aspect

Une séquence de mots (tokens) est une expression linguistique utilisée dans l'examen portant sur certaines caractéristiques de l'entité. Il doit apparaître dans la phrase donnée et avoir généralement un vocabulaire étendu. Cela s'appelle également une expression cible.

Un aspect peut être soit prédéfini, soit extrait des données réelles. Selon Bing Liu (B. Liu, 2012), les chercheurs de la communauté du domaine de la fouille d'opinion ont identifié deux types

d'aspect : explicite et implicite. Un aspect implicite est la cible de l'opinion qui n'est pas explicitement spécifiée dans le texte. Par exemple, dans la revue « Cet appareil photo est élégant et très abordable » donne un avis sur l'aspect et le prix des aspects par les mots explicites « élégant » et « abordable ». Par contre, les aspects explicites correspondent à des mots spécifiques dans le document. Prenant l'exemple suivant : « l'optique de cette caméra est très bonne et la durée de vie de la batterie est excellente ». Les deux aspects optiques et batterie sont présentés d'une manière explicite dans le document.

Une expression cible est un mot court qui doit être explicitement mentionné dans les commentaires. Par exemple, dans une critique de restaurant, l'entité de la critique est le restaurant. Les aspects pourraient inclure le prix, la nourriture et ainsi de suite. Les expressions cibles peuvent être des rouleaux de sushi ou une salade avec une vinaigrette délicieuse.

La mission de la fouille d'opinion basée aspects est d'identifier les aspects des entités et l'opinion ou le sentiment exprimé sur chaque aspect. Pour ce faire, nous devons décomposer la tâche de la fouille d'opinion basée sur les aspects en plusieurs sous-tâches pouvant extraire des aspects, classification en catégories, extraction des termes d'aspect et des polarités d'opinion (Pontiki et al., 2016; Pontiki, Galanis, Papageorgiou, Manandhar, & Androutsopoulos, 2014, 2015). Ces différentes sous tâches sont définies comme suit :

- **L'extraction des aspects (AE)** : le but de cette tâche est d'extraire les aspects de l'élément examiné et de regrouper les synonymes des aspects, car différentes personnes peuvent utiliser différents mots ou expressions pour faire référence au même aspect, par exemple, affichage, écran, LCD réfère au même objet.
- **Classification par catégorie d'aspect (ACC)** : Cette tâche vise à classer le texte donné en fonction d'un ensemble prédéfini de catégories d'aspect. C'est un cas particulier de classification de texte.
- **Extraction des termes d'aspect (ATE)** : Cette tâche est également appelée détection d'expression cible (Pontiki et al., 2014). Il est dédié à la recherche d'un texte intéressant dans une phrase donnée, qui peut être un mot unique ou une phrase. Par exemple « rouleaux de sashimi ».
- **Classification de la polarité des sentiments au niveau des aspects** : La classification de la polarité des opinions peut être effectuée à plusieurs niveaux : niveau du document,

niveau de la phrase et niveau aspect. Si les polarités des opinions sont disponibles au niveau de l'aspect, les lecteurs n'auraient pas besoin d'entrer dans les détails, mais pourraient obtenir plus d'informations sur les opinions à partir des scores de polarité globaux. Le texte donné peut être classé comme positif, négatif ou neutre ; ou ils pourraient être étiquetés en valeurs numériques (de 1 jusqu'à 5).

La stratégie commune entre ces tâches partielles consiste à créer un ensemble de caractéristiques riches et à utiliser des modèles standards. Cependant, la fouille d'opinion basée sur les aspects (ABOM) n'est pas nécessairement un projet séquentiel dans lequel chaque sous-tâche doit être exécutée l'une après l'autre. Chaque sous tâche peut être effectuée individuellement ou peut être combinée avec d'autres tâches. Dans cette thèse, nous explorons des approches multitâches pour résoudre ces sous-tâches, car chaque sous-tâche dans ABOM est étroitement corrélée. Les caractéristiques fournies par une tâche pourraient être utilisées pour enrichir les autres tâches d'apprentissage. Par exemple, les termes d'aspect extraits dans une tâche ATE pourraient constituer de bonnes caractéristiques pour la classification de catégories d'aspect dans une tâche AE. La représentation des caractéristiques communes partagée par la tâche ACC et la tâche ATE peut réduire le bruit généré par chaque tâche, ce qui améliore encore la robustesse des systèmes.

### 2.3 Mesures d'évaluation de la performance des systèmes de la fouille d'opinion

Tout domaine de recherche à une méthodologie d'évaluation quantitative des approches proposées par les chercheurs du domaine, dans la fouille d'opinion en général et particulièrement celle basée sur les aspects, les chercheurs de la communauté ont évalué et prouvé les performances de leurs approches à l'aide des trois mesures suivantes : la précision (P), le rappel (R), et le F-score (F) (Bagheri, Saraee, & de Jong, 2013; M. Hu & Liu, 2004; S.-M. Kim & Hovy, 2006; Moghaddam & Ester, 2010; Popescu & Etzioni, 2005; Wei, Chen, Yang, & Yang, 2010).

Pour obtenir la valeur de précision, nous divisons le nombre total d'exemples positifs correctement classés par le nombre total d'exemples positifs prédits. Le rappel peut être défini comme le rapport entre le nombre total d'exemples positifs correctement classés et le nombre total d'exemples positifs. Ces mesures sont présentées comme suit :

$$Precision = \frac{TP}{TP+FP} \dots\dots\dots(03)$$

$$Recall = \frac{TP}{TP+FN} \dots\dots\dots(04)$$

$$F(score) = \frac{2*P*R}{P+R} \dots\dots\dots(05)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots(06)$$

Où TP (Vrai Positif), TN (Vrai Négatif), FP (Faux Positif) et FN (Faux Négatif) sont les cas correctement classés ou mal classés, et F-score est la moyenne des deux métriques : précision et rappel.

## 2.4 Les datasets d'expérimentation

Les chercheurs de la communauté de la fouille d'opinion manipule généralement des données issues des sites web, des plateformes commerciales, des forums, et des avis des clients sur Amazon, Tripadvisor, Yelp, etc. Une variété de systèmes d'exploration d'opinion basés sur des aspects a été développée pour manipuler ces données, tels que les critiques de films (Thet, Na, & Khoo, 2010), les critiques de produits électroniques (M. Hu & Liu, 2004), les appareils photos numériques (M. Hu & Liu, 2004) et les revues de restaurants, mais le test de ces systèmes était toujours effectué sur des ensembles de données individuels, de sorte que de véritables avancées dans le domaine ne peuvent pas être correctement évaluées. Pour cela, plusieurs ensembles de données de référence ont été rendus publics, tels que les ensembles de données sur les évaluations de produit de Hu et Liu (M. Hu & Liu, 2004) ou sur les évaluations de restaurant de (Ganu, Elhadad, & Marian, 2009).

Certaines parties des jeux de données mentionnés précédemment ont été extraites et réannotées pour l'atelier international de l'évaluation de la sémantique SemEval2014, tâche 4 (Pontiki et al., 2014) et SemEval2015, tâche 12 (Pontiki et al., 2015). En 2016, sept autres langues telles que l'arabe ont également été incluses dans une troisième série de tâches, à savoir SemEval 2016 Task5 (Pontiki et al., 2016), comme indiqué dans le tableau (2.1).

SemEval (évaluation sémantique) est une série continue d'évaluations de systèmes d'analyse sémantique par calcul ; il a évolué à partir de la série d'évaluation du sens des mots. Les évaluations ont pour but d'explorer la nature de la signification du langage. Les exercices SemEval ont pour objectif d'évaluer les systèmes d'analyse sémantique. "Analyse sémantique" désigne une analyse formelle de la signification. Dans cette thèse, nous utilisons l'un des jeux de donnée de l'atelier international sur l'évaluation sémantique (SemEval 2014).

<i>Domaine</i>	<i>Sous-domaine</i>	<i>Langue</i>	<i>#Phrases</i>
<b>Electroniques</b>	Camera	Chinoise	8040
	Pc Portable	Anglais	3308
	Phones	Chinoise	9521
	Phones	Néerlandaise	1697
<b>Hôtels</b>		Arabe	6029
<b>Restaurants</b>		Néerlandaise	2286
		Anglais	2676
		Français	2429
		Russe	4699
		Espagnol	2951
		Turc	1248
<b>Télécoms</b>		Turc	3310

**Tableau 2.1** Différents jeux de données de SemEval (Pontiki et al., 2014)

En général le groupe de SemEval ont fournis aux chercheurs 39 jeux de données pour différentes langues (8 langues) : Anglais (en), arabe (ar), chinois (ch), néerlandais (du), français (fr), russe (ru), espagnol (es) et turc (tu), dont 19 jeux pour la formation et 20 pour les tests.

En outre, Les jeux de données fournis, concernant les domaines : la restauration (repos), les ordinateurs portables (lapt), les téléphones portables (phns), les appareils photos numériques (cam), les hôtels (hote) et les musées (muse) consistent en des commentaires de clients, tandis que les données du domaine des télécommunications (telc) consistent en des tweets. Plusieurs groupes de recherche de la communauté, ont contribué pour garantir le bon déroulement des processus d'annotation dans différentes langues et différents domaines comme indiqué dans le tableau (2.2).

Lang.	Equipe de recherche
Anglais	Institut de traitement du langage et de la parole, Athena RC, Athènes, Grèce Département d'informatique, Université d'économie et de commerce d'Athènes, Grèce
arabe	Département d'informatique, Université jordanienne des sciences et technologies Irbid, Jordanie
chinois	Institut de technologie de Harbin, Harbin, Heilongjiang, République populaire de Chine
néerlandais	LT3, Université de Gand, Gand, Belgique
français	LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Orsay, France
russe	Université d'État de Moscou Lomonossov, Moscou, Fédération de Russie Université d'État de Viatka, Kirov, Fédération de Russie
Espanol	Universitat Pompeu Fabra, Barcelone, Espagne SINAI, Université de Jaén, Espagne
turc	Département de génie informatique, Université technique d'Istanbul, Turquie Turkcell Global Bilgi, Turquie

**Tableau 2.2** Équipes de recherche ayant contribué à la création des jeux de données pour chaque langue. (Pontiki et al., 2015)

## 2.5 Etat de l'art sur la fouille d'opinion basée sur les aspects

Plusieurs travaux ont été réalisés pour effectuer les différentes tâches de la fouille d'opinion, comme indiqué par (Feldman, 2013), plus de 7 000 articles ont été publiés sur différents domaines de la fouille d'opinion. Pour cela, nous incluons le plus récent de l'état de l'art des sept (07) dernières années. En raison du grand nombre de travaux de recherche sur des aspects explicites, nous avons divisé les approches en trois catégories principales : supervisé, semi-supervisé et non supervisé.

Pour chaque catégorie, nous présentons une comparaison et une analyse des différentes techniques utilisées dans les études de la fouille d'opinion basée sur les aspects, en tenant compte l'année de l'étude publiée, le modèle et l'algorithme implémenté, le jeu de donnée utilisé, la langue, et les résultats expérimentaux qui sont extraits à partir des papiers d'origine et présentées par les paramètres suivants : rappel (R), précision (P) et F-mesure (F) comme indiqué dans les tableaux 2.3, 2.4 et 2.5.

### 2.5.1 Approches supervisées

Les, approches supervisées ont été largement utilisées dans un grand nombre d'articles de recherche pour les tâches de la fouille d'opinion basée sur les aspects. Elles ont été appliquées à différents jeux de données dans divers domaines tels que les produits, les mobiles, les ordinateurs portables, les hôtels, etc. dans ce qui suit on va présenter les plus récentes entre ces approches.



(Cruz, Troyano, Enríquez, Ortega, & Vallejo, 2013) ont proposé d'utiliser une taxonomie afin d'extraire des aspects de l'analyse des sentiments. Ils ont introduit une approche orientée domaine où le même mot a différentes significations selon le domaine et le document. (S. Huang, Liu, Peng, & Niu, 2012) ont introduit un modèle d'étiquetage de séquence pour effectuer la sous-tâche d'extraction d'aspect pour les textes porteurs opinion.

(Yang & Cardie, 2013) ont tenté de relier des opinions avec les entités correspondantes en définissant deux relations différentes : IS-ABOUT et IS-FROM. Ils ont également identifié la cible de l'opinion, le détenteur de l'opinion et les expressions de sentiments possibles.

(S. Li, Wang, & Zhou, 2012) ont effectué l'extraction d'aspect sous la forme d'une analyse sémantique. Ils mettent en œuvre leur proposition sur les services universitaires et le Web.

Ces dernières années, les chercheurs ont introduit des modèles d'apprentissage en profondeur dans différentes applications telles que la classification de texte, la classification d'images et la reconnaissance de la parole (Graves, Mohamed, & Hinton, 2013; He, Zhang, Ren, & Sun, 2016; Y. Kim, 2014b; Sun, Luo, & Chen, 2017). En outre, plusieurs chercheurs ont tenté de créer des vecteurs de caractéristiques avec moins d'effort et sans ingénierie d'entités, les chercheurs ayant prouvé que les réseaux de neurones récurrents (RNN) sont des techniques sophistiquées pour la présentation séquentielle des données. Les RNNs ont résolu de nombreuses tâches d'étiquetage de séquences (Jabreel, Hassan, & Moreno, 2018; Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016; Ling et al., 2015).

(Socher et al., 2013) ont proposé une approche basée sur les arbres syntaxiques et les réseaux de neurones récurrents pour présenter leurs vecteurs de caractéristiques de manière à obtenir les sentiments des phrases. Dans (Lakkaraju, Socher, & Manning, 2014), les auteurs implémentent un modèle RNN en utilisant deux types de classificateurs de type softmax capables d'extraire de nombreuses catégories d'aspect avec leurs sentiments correspondants. Ces catégories peuvent être jointes dans un modèle ou séparément.

(Le & Zuidema, 2015) ont étendu les réseaux de neurones récurrents avec la mémoire à court terme (LSTM) pour calculer un vecteur parent dans les arbres d'analyse en combinant les informations des cellules de sortie et des cellules de mémoire LSTM de ses deux enfants.

(K. Liu et al., 2015) ont utilisé l'incorporation de mots en tant qu'extracteur de caractéristiques pour un réseau de neurones récurrent afin d'obtenir les aspects explicites à partir des textes porteurs d'opinion.

(Yin et al., 2016) ont utilisé la présentation des mots prolongée (word embedding) comme une entrée pour le modèle de champ conditionnel aléatoire pour la sous-tâche d'extraction de termes d'aspect.

(Wang, Pan, Dahlmeier, & Xiao, 2016) ont intégré des champs aléatoires conditionnels récurrents et des réseaux de neurones dans la même structure pour les termes d'opinion et les aspects. Leur proposition a été formée à la fois sur les informations à double propagation et sur les caractéristiques discriminantes de haut niveau entre termes et aspects d'opinion.

(P. Liu, Joty, & Meng, 2015) ont présenté une architecture de réseau de neurones récurrents (RNN) avec des intégrations de mots préformés (pré-trained word embedding) sur différents RNN (Elman, 1990; Jordan, 1986) cela, en mettant en œuvre une variante des réseaux de mémoire à court terme (LSTMs).

Une approche d'apprentissage en profondeur mettant en œuvre un CNN en profondeur sur 7 couches a été proposée par (Poria, Cambria, & Gelbukh, 2016) pour marquer chaque mot du texte comme un mot d'aspect ou non. Ils ont également combiné des modèles linguistiques avec le réseau de neurones. Leurs expériences ont été menées sur deux jeux de données incorporant la présentation vectorielle des mots prolongés de Google et Amazon.

Une approche basée sur les mots prolongés (word embedding) et le réseau de neurone récurrent RNN a été proposée par (Jebbara & Cimiano, 2016) pour effectuer la sous-tâche extraction d'aspect / sentiment. Leur modèle pourrait également identifier plus d'une opinion dans une phrase.

(Ruder, Ghaffari, & Breslin, 2016) ont présenté un modèle fondé sur la hiérarchie. Selon leurs propositions, ils ont donc constaté que l'efficacité de leur proposition est meilleure s'ils prennent en compte la structure et le contexte de la revue pour ses prédictions. Une autre approche basée sur les réseaux de neurones récurrents synchronisés bidirectionnels (BGRNN) (Jabreel et al., 2018) a été proposée pour extraire la cible d'un tweet ; elle identifie à la fois les relations entre la cible et son contexte et la polarité de tweet par rapport à la cible. Les différentes approches supervisées que nous avons présentées précédemment sont récapitulées dans le tableau 2.3 comme suit :

Ref/Année	Modèle	Jeux de donnée	Langue	Résultats expérimentaux %		
				P	R	F-M
(S. Huang et al., 2012)	CRF1	Camera	Anglais	80,4	71,4	75,6
(S. Li et al., 2012)	OTE	Web-service University	Anglais	75,26 75,58	59,53 59,81	66,39 66,54
(Cruz et al., 2013)	Tax-based	Headphon Hotel Cars	Anglais	71,15 67,82 71,69	66,17 73,88 72,96	68,57 70,72 72,32
(Yang & Cardie, 2013)	CRF2	Question Answering	Anglais	75,18	57,12	64,92
(Socher et al., 2013)	RNTN	Stanford Sen timent Treebank	Anglais	-	-	80,7
(Lakkaraju et al., 2014)	JMAS + RNTN	beer reviews <sup>14</sup> camera	Anglais	72,02	-	-
(P. Liu et al., 2015)	LSTM-RNN1	Laptop Restaurant	Anglais	85,43 83,64	83,54 81,39	83,18 81,54
(Le & Zuidema, 2015)	LSTM-RNN	Stanford Sentiment Treebank	Anglais	49,9	-	-
(K. Liu et al., 2015)	WAM	CRD <sup>15</sup> COAE 2008 <sup>16</sup>	Anglais Chinoise	92 86	90 86	90 80
(Yin et al., 2016)	WDEmb	Laptop restaurant SemEval 2015 <sup>17</sup>	Anglais	- - -	- - -	75,16 84,97 69,73
(Wang et al., 2016)	RNCRF	Restaurant Laptop <sup>18</sup>	Anglais	- -	- -	84,05 76,83
(Poria et al., 2016)	CNN +LP	Restaurant	Anglais	87,42	85,01	86,20
(Jebbara & Cimiano, 2016)	2SNN	customer reviews	Anglais	66,3	71	68,4
(Ruder et al., 2016)	H-LSTM	SEMEVAL 2016	Anglais Arabe	83 82,8	- -	- -
(Jabreel et al., 2018)	biGRU	Tweets	Anglais	87,39	91,18	89,25

**Tableau 2.3** : Synthèse des approches supervisées (Bouras, Amroune, & Ben Djanna, 2019).

### 2.5.2 Approches semi-supervisées

Diverses études ont été proposées par la communauté des chercheurs pour effectuer la sous-tâche d'extraction d'aspect à l'aide d'approches semi-supervisées comme illustré dans le tableau 2.4,

<sup>14</sup> <http://snap.stanford.edu/data/web-BeerAdvocate.html>.

<sup>15</sup> (CRD): Customer Review Data-sets contient des critiques en anglais sur 5 produits.

<sup>16</sup> Contient des critiques chinoises de quatre types de produits: appareils photo, voitures, ordinateurs portables et téléphones <http://ir-china.org.cn/coae2008.html>

<sup>17</sup> Évaluation sémantique: est une série continue d'évaluations de systèmes d'analyse sémantique.

<sup>18</sup> <http://www.yelp.com/dataset challenge>.

parmi lesquelles celle présentée par (Hai, Chang, & Cong, 2012), dans laquelle les auteurs ont proposé deux approches différentes : l'une est l'analyse sémantique latente appelée approche de bootstrapping (LSABOOT) et l'autre est le rapport de vraisemblance (LRTBOOT) pour extraire un aspect spécifique du texte brut avec opinion. Une autre contribution proposée dans (Zhao, Wang, Lv, & Zhang, 2014), où les auteurs ont introduit une méthode d'amorçage utilisant un ensemble de règles de raffinement, pour capturer les résultats erronés et actualiser les règles afin de rendre le système proposé plus précis.

Une approche basée sur l'analyse de dépendance a été proposée par (Yu, Zha, Wang, & Chua, 2011), dans laquelle les auteurs ont fait référence à une liste d'aspects définie à l'avance pour obtenir les aspects évalués par la majorité des utilisateurs. Dans cette contribution, les auteurs ont classé différents aspects en fonction des opinions de l'utilisateur et de la fréquence des aspects. Ils proposent donc un algorithme probabiliste basé sur la régression, dans lequel leur proposition attribue des pondérations aux aspects habituellement exprimés par les utilisateurs.

Une autre approche de la fouille d'opinion basée sur les aspects (B. Ma, Zhang, Yan, & Kim, 2013), implémentée sur des ensembles de données chinois a été proposée en reliant une liste de synonymes à (LDA) « Latent Dirichlet Allocation ». Les auteurs de cette proposition ont considéré les noms et les expressions nominales comme des aspects. L'incorporation de l'allocation de Dirichlet latent « Latent Dirichlet Allocation » va générer une liste d'aspects candidats, développée à l'aide du lexique des synonymes. Cette liste élargie a comme but : l'enrichissement de la liste finale d'aspects en considérant d'autres termes qui n'étaient pas considérés comme des aspects. Ce processus va éliminer les termes en utilisant des règles de filtrage pour définir la liste finale des aspects.

Dans le but d'identifier l'association (aspect - sentiment), les chercheurs ont présenté deux études récentes, l'une d'entre elles réalisée par (Yan, Xing, Zhang, & Ma, 2015), dans laquelle les auteurs ont développé un algorithme basé à la fois sur le lexique des synonymes et sur un PageRank pour étendre leur liste d'aspect. L'autre est une approche partiellement supervisée basée sur un modèle d'alignement de mots (PSWAM) présenté par (K. Liu, Xu, Liu, & Zhao, 2013). L'idée proposée est basée sur l'utilisation de modèles syntaxiques dans le modèle d'alignement de mots pour trouver les relations entre les sentiments et les aspects mentionnés dans les textes porteurs d'opinion. À cet effet, les auteurs s'inspirent du travail présenté par (K. Liu, Xu, & Zhao, 2012) pour aligner les mots.

(Samha, Li, & Zhang, 2014), ont défini une liste manuelle d'aspects en utilisant à la fois les informations sur les aspects et WorldNet pour définir des aspects similaires. Pour extraire des mots de

sentiment spécifiques, ils ont recherché les termes d'opinion les plus courants en appliquant un ensemble de lexiques d'opinion.

Dans le tableau 2.4, les différentes approches semi-supervisées avec leurs modèles appropriés et leurs résultats fournis par modèle sont listées selon leurs dates de publications, notons que les résultats expérimentaux sont tirés à partir des papiers originaux.

Ref/Année	Modèle	Jeux de donnée	Langue	Résultats expérimentaux %		
				P	R	F-M
(Yu et al., 2011)	DParser	Product	Anglais	-	-	73
(Hai et al., 2012)	LRTBOOT	Cell phone	Chinoise	68.00	79.40	73.25
				57.43	67.13	61.90
	LSABOOT	Hotel		63.25	81.11	71.07
				45.34	69.60	54.91
(B. Ma et al., 2013)	LDA -SL	Product Costumer review	Chinoise	75,17	73,25	73,94
(K. Liu et al., 2013)	PSWAM	COAE2008 <sup>19</sup> Large	Anglais	85 85	75 85	80 85
(Zhao et al., 2014)	BST	Product	Chinoise	80	76	80
(Samha et al., 2014)	FT	Costumer (Hu and Liu's)	Anglais	99	64	77
(Yan et al., 2015)	EXPRS	Product	Chinoise	80,5	70,06	73,96

**Tableau 2.4** : Synthèse des approches semi-supervisées (Bouras et al., 2019).

### 2.5.3 Approches non supervisées

Les techniques non supervisées ont été implémentées dans un grand nombre d'études pour extraire des aspects à partir des commentaires en ligne dans divers domaines et sur différents jeux de données comme illustré dans le tableau 2.5. La plupart des approches se sont concentrées sur les ensembles de données des commentaires des clients, L'utilisation de ces techniques est apparue en premier lieu dans un papier de recherche présenté par (M. Hu & Liu, 2004).

Une approche de synthèse basée sur les caractéristiques pour les critiques des produits a été proposée par (Bafna & Toshniwal, 2013). Dans une première étape, le modèle extrait pour chaque produit une caractéristique correspondante, puis, pour chaque caractéristique, il a sélectionné les revues correspondantes. Ensuite, il extrait des parties des revues contenant des mots d'opinion de la revue

<sup>19</sup><http://ir-china.org.cn/coae2008.html>

globale. Enfin, les parties extraites ont été placées dans leur propre cluster en tant que caractéristique positive ou négative.

(Marrese-Taylor, Velásquez, Bravo-Marquez, & Matsuo, 2013) ont contribué dans le domaine du tourisme (restaurants et hôtels) où ils ont présenté la phrase comme une collection de jetons (tokens) pouvant être un mot ou une ponctuation. Ils ont également considéré deux jetons répétés comme des jetons séparés. Ils ont utilisé la distance entre deux mots pour extraire des aspects à partir des textes porteurs d'opinion. Une extension du même travail est présentée par (Marrese-Taylor, Velásquez, & Bravo-Marquez, 2014) sous forme d'une architecture générique pour générer un modèle d'analyse de sentiment basé sur les aspects pour le domaine du tourisme. Les mêmes chercheurs ont également contribué pour générer un modèle modulaire nommé Opinion Zoom (Marrese-Taylor, Velásquez, & Bravo-Marquez, 2013), pour le même domaine.

Deux algorithmes sont définis par (Eirinaki, Pinal, & Singh, 2012), le premier est utilisé pour extraire à la fois l'aspect et le sentiment dans des textes d'opinion et le second est implémenté pour définir les rangs des aspects extraits. Ces algorithmes ont été proposés pour obtenir le sentiment vis-à-vis l'aspect et son rang en comptant les mots d'opinion ciblant cet aspect.

Un modèle monolingue pour l'alignement de mots a été proposé par (K. Liu et al., 2012) pour la recherche d'associations entre des aspects et des opinions en effectuant la tâche d'association d'aspect.

Une approche basée sur un algorithme de boosting a été proposée par (Bagheri et al., 2013). Cette approche nécessite des ensembles initiaux d'aspects (seed sets). Le POS modèle (part of speech pattern) est utilisé pour définir les aspects, en se basant sur des heuristiques. En outre, les auteurs ont présenté une nouvelle métrique appelée A-score. Ils ont mis en œuvre un algorithme d'amorçage utilisant la métrique définie et les différentes informations sur l'interrelation entre les mots pour obtenir un ensemble principal d'aspects.

Une autre recherche présentée par (Y. Li, Qin, Xu, & Guo, 2015; Y. Li, Wang, Qin, Xu, & Guo, 2014), où ces auteurs ont proposé une approche basée sur un modèle de dépendance à l'amorçage (bootstrapping dependency pattern) pour obtenir à la fois les aspects et les jugements correspondants. La structure grammaticale des phrases est la base des règles utilisées pour identifier certains modèles d'extraction d'aspect. Après le processus d'extraction des aspects, les auteurs rassemblent les mêmes aspects en se basant sur la distance de similarité et sur le dictionnaire WorldNet.

ASPECTATOR est un algorithme basé sur la dépendance syntaxique a été proposé par (Bancken, Alfarone, & Davis, 2014), pour identifier et évaluer les aspects examinés.

Une approche fondée sur des règles a été proposée par (Poria, Cambria, Ku, Gui, & Gelbukh, 2014), dans laquelle ils ont considéré les aspects explicites comme des concepts décrivant une cible dans les commentaires des clients en utilisant la présentation de mot qui correspond à ce concept.

Un nouveau modèle basé sur un langage traduit (TrLM) proposé par (Du, Chan, & Zhou, 2014), pour extraire les aspects du produit critiqué. Les chercheurs ont prouvé que la structure des revues avait un impact à la fois sur les mots d'opinion et sur les aspects.

Un groupe de chercheurs de la communauté de la fouille d'opinion ont contribué dans des travaux sur des bases de données en langue anglaise (Hai, Chang, Kim, & Yang, 2013), où ils ont tenté d'extraire un aspect et un mot d'opinion de la revue chinoise en proposant une approche basée sur des règles. Deux paramètres sont utilisés pour définir l'aspect et l'opinion. Ces paramètres sont IDR (pertinence du domaine intrinsèque) et EDR (pour la pertinence du domaine extrinsèque). IDR et EDR sont des valeurs décrivant la pertinence d'un aspect par rapport à son domaine approprié.

La technique TF-IDF (TermFrequency-Inverse Document Frequency) est implémentée par (Quan & Ren, 2014) pour la recherche d'association entre l'aspect et les mots du domaine (mots de domaine d'aspect). Ils ont prouvé que l'extraction d'opinion basée sur les aspects est une sous-tâche qui dépend de domaine. Deux types de recommandations basées sur des associations d'aspects ont été implémentés dans une autre approche basée sur l'apprentissage tout au long de vie (lifelong learning-based approach) (Q. Liu, Liu, Zhang, Kim, & Gao, 2016).

Dans le tableau 2.5 nous récapitulons les différentes approches non supervisées citées dans la section précédente et proposées par l'ensemble des chercheurs de la communauté de la fouille d'opinion, spécifiquement le groupe du sous domaine de la fouille d'opinion basé sur les aspects.

Ref /Année	Modèle	Jeux de donnée	Langue	Résultats expérimentaux %		
				P	R	F-M
(Bafna & Toshniwal, 2013)	FBS	Product	English	-	-	92
(Marrese-Taylor et al., 2014; Marrese-Taylor, Velásquez, Bravo-Marquez, et al., 2013)	OZ	Tourism	English	38	33	36
(Eirinaki et al., 2012)	HAC	Product	English	-	-	72
(K. Liu et al., 2012)	WTM	COAE2008	Chinese	73	78	75
		Large	English	74	82	78
		Product	English	86	87	86
(Bagheri et al., 2013)	BST	Product	English	84,1	66,2	-

(Y. Li et al., 2015; Y. Li et al., 2014)	CERAB	Product	English	67	94	78
(Bancken et al., 2014)	Aspectator	Movies	English	-	-	72,5
(Poria et al., 2014)	RB	Product (Hu and Lui 2004)	English	89,41	91,42	-
		SEMEVAL 2014 Laptop	English	82,15	85,21	-
		SEMEVAL 2014 Restaurant	English	84,32	88,15	-
(Du et al., 2014)	TrLM	Product	English	-	-	79
(Hai et al., 2013)	IEDR	Cellphone Hotel	Chinese	65,60 50,37	61,71 54,30	63,60 52,26
(Quan & Ren, 2014)	PMI- TFIDF	Product	English	84	91	97
(Q. Liu et al., 2016)	AER	Hu and Liu 2004	English	73,8	78,7	75,9
		Liu et al. 2015	English	72,5	75,5	73,3

**Tableau 2.5** : Synthèse des approches non supervisées (Bouras et al., 2019).

Dans la section précédente nous avons présenté un état de l'art sur la fouille d'opinion basée sur les aspects, dans laquelle nous avons récapitulé plus de 33 techniques durant les sept (07) dernières années (2012-2018) divisées en trois catégories (supervisée, semi-supervisée et non supervisée). Dans le chapitre IV une étude comparative et synthèse sur ces différentes approches sera présentée pour donner une vision générale sur les différents travaux réalisés par les chercheurs de la communauté du domaine de la fouille d'opinion d'une part, et d'une autre part, pour faire face aux défis du domaine et de ressortir de nouveaux axes de recherche afin d'améliorer les performances et les résultats des contributions proposées.

## 2.6 Conclusion

Récemment, l'extraction d'opinion basée sur les aspects a suscité beaucoup d'attention car elle fournit clairement des informations plus détaillées permettant aux utilisateurs de prendre des décisions et aux fournisseurs de surveiller leurs consommateurs. Dans le même but, plusieurs approches ont été proposées pour extraire ces aspects et leurs opinions appropriées. Certaines de ces approches ont utilisé des modèles basés sur apprentissage supervisé et semi-supervisés, tandis que d'autres approches



récentes ont profité du nouveau flux de recherche basé sur l'apprentissage approfondi (Deep Learning) avec des réseaux de neurones à plusieurs couches cachées. Vu l'importance et les résultats fournis par ces modèles tels que les réseaux de neurones convolutionnel (CNN) et les réseaux à long-court terme (LSTM), le chapitre suivant présentera en détail les modèles basés Deep learning pour la sous tâche de la fouille d'opinion basée sur les aspects.

## 3 Chapitre III : Les modèles d'apprentissage approfondi pour la fouille d'opinion basée aspect

### 3.1 Introduction

L'intérêt croissant suscité par l'analyse et la fouille des opinions a conduit les chercheurs à développer de nouvelles technologies basées sur l'apprentissage automatique et l'exploration de données. Ils développent des outils automatiques permettant d'extraire les opinions, de détecter les polarités des sentiments et de résumer le texte porteur d'opinion généré par les utilisateurs sur un même sujet. Les méthodes classiques d'apprentissage automatique dépendent largement de la représentation des données pour découvrir la correspondance entre les caractéristiques et les objectifs. L'extraction manuelle et l'ingénierie de ces caractéristiques sont aussi critiques que la modélisation ou elles nécessitent généralement une grande expertise du domaine et du travail humain. Si on développe par exemple un modèle classique d'apprentissage supervisé basé sur SVM ou sur une classification bayésienne pour prédire la polarité des sentiments d'une critique de film, on doit concevoir un ensemble de caractéristiques discriminantes, on se basant sur des méthodes et des techniques traditionnelles (Pang et al., 2002).

La plupart des algorithmes classiques d'apprentissage automatique sont basés sur des règles ou sur des corpus. Leur efficacité dépend de la qualité des corpus annotés ainsi que de la tâche d'extraction des caractéristiques du processus de la fouille d'opinion. Les caractéristiques doivent être créées manuellement, car elles diffèrent d'un domaine à l'autre et d'un document à l'autre, ce qui la rend moins générique et plus spécifique au texte. La précision des systèmes basés sur le paradigme traditionnel dépend de la manière dont les caractéristiques ont été choisies, de plus, il est très difficile, et souvent impossible, d'adopter un système conçu pour un problème particulier face à de nouveaux problèmes ou dans une langue différente pour le même problème, et pour des textes tels que les tweets, qui ne suivent aucune règle ou grammaire précise.

Récemment, l'apprentissage en profondeur a permis l'apprentissage de la représentation au sein de réseaux de neurones profonds de bout en bout. L'idée principale d'apprentissage en profondeur a été renommée sous de nombreux noms (Goodfellow, Bengio, & Courville, 2016), jusqu'à la récente renaissance de 2006 (Hinton, Osindero, & Teh, 2006). Cette idée a connu plusieurs hauts et bas depuis

les années 1940 (Goodfellow et al., 2016). Les concepts clés, tels que la représentation distribuée, le rétro propagation, la mémoire à court et long terme (LSTM) (Hochreiter & Schmidhuber, 1997) et le réseau de neurone convolutionnel (CNN) (LeCun, Bottou, Bengio, & Haffner, 1998), restent toujours les composants essentiels des modèles d'apprentissage en profondeur. Cependant, ils n'ont pas reçu beaucoup d'attention, comme ils le font aujourd'hui, jusqu'à ce que Geoffrey Hinton ait introduit en 2006 le réseau de croyance profond, un modèle graphique génératif pouvant être efficacement formé par un apprentissage préalable par couches (Hinton et al., 2006). Depuis lors, l'apprentissage en profondeur a présenté sa performance et efficacité dans nombreux domaines de recherche connexes, tels que la vision par ordinateur, le traitement du langage naturel, la reconnaissance de la parole et l'apprentissage par renforcement (Yann LeCun et al., 2015).

La disponibilité d'ensembles de données étiquetées à grande échelle et les progrès de la puissance de calcul contribuent à la renaissance de ce nouveau domaine d'apprentissage en profondeur. La taille du jeu de données augmente de façon exponentielle, de MNIST (LeCun et al., 1998), qui consiste en 70 000 chiffres manuscrits numérisés, à ImageNet (J. Deng et al., 2009) contenant 1,2 million d'images de 1 000 catégories et à la plus récente Youtube-8M (Abu-El-Haija et al., 2016) composée de 8 millions de vidéos annotées (500 000 heures au total).

Il est possible de former des modèles sophistiqués avec une masse de données très importante, mais ils ne peuvent pas fonctionner sans une puissance de calcul énorme, pour cela des unités de traitement graphique (GPU) rapides ont été introduites (Raina, Madhavan, & Ng, 2009) pour accélérer considérablement l'apprentissage automatique du modèle. Les clusters et les serveurs dédiés tels que Nvidia DGX-1 ont été conçus pour l'apprentissage approfondi des modèles de données de grandes échelles. Les unités de traitement nommé Tensor (TPUs), qui sont jusqu'à 70 fois plus rapides que les GPU, elles offrent une efficacité énergétique accrue qui a été utilisée dans les centres de données Google et Google Cloud (Jouppi et al., 2017).

En plus de l'avènement du matériel rapide, la croissance de l'apprentissage en profondeur dépend également d'une meilleure infrastructure logicielle, notamment CUDA (Nickolls, Buck, & Garland, 2008), Theano (Bergstra et al., 2010), Caffè (Jia et al., 2014), Torch (Collobert, Kavukcuoglu, & Farabet, 2011), Tensorflow (Abadi et al., 2016), Pytorch (<http://pytorch.org>), Caffè2 (<http://caffe2.ai>) ect. Grâce à la contribution persistante des communautés "open source", il devient facile de développer des modèles d'apprentissage approfondi sur du matériel informatique avancé sans faire appel à une programmation de bas niveau. La capacité des modèles basés sur les réseaux de neurones augmente considérablement en termes de nombre de couches ou de nombre de neurones. Par

exemple, en vision par ordinateur, le nombre de couches de réseaux de neurones modernes commence à 8 couches (AlexNet (Krizhevsky, Sutskever, & Hinton, 2012)), 16 couches (VGGNet (Simonyan & Zisserman, 2014)), 22 couches (GoogleLeNet (Szegedy et al., 2015)) à un nombre important de 152 couches (ResNet (He et al., 2016)).

Donc, les modèles d'apprentissage approfondi promettent avec suffisamment de données et de temps de formation, la classification des sentiments sur tout type de texte avec un minimum de restrictions et sans ingénierie manuelle des caractéristiques spécifiques à la tâche ou aux données. L'apprentissage en profondeur peut être exploité pour le traitement de données à grande échelle, des modèles puissants et des ressources de calcul accrues. Dans ce chapitre nous introduisant une vue générale sur cette nouvelle discipline de recherche, ses avantages, l'ensemble des méthodes existantes proposées pour l'analyse des sentiments et l'exploration d'opinion en utilisant l'apprentissage en profondeur, une brève introduction aux réseaux de neurones et aux modules essentiels est fournie en premier lieu. Ensuite, nous présentons les architectures des différents réseaux de neurones utilisées pour les tâches importantes de l'analyse des sentiments et de l'extraction d'opinions.

### **3.2 Les techniques traditionnelles pour l'extraction des caractéristiques**

La précision des systèmes d'apprentissage automatique basés sur le paradigme traditionnel dépend de la manière dont les caractéristiques ont été choisies. A cet effet, plusieurs techniques ont été introduites dans la littérature, nous présentons dans ce qui suit celles qui ont été fréquemment utilisées par les chercheurs de la communauté.

#### **3.2.1 La présence et fréquence des termes**

Ces caractéristiques sont des mots individuels ou des n-grammes de mots et leur nombre de fréquences. Il attribue soit la pondération binaire aux mots (zéro si le mot apparaît, soit un autre sinon), soit l'utilisation de pondérations de fréquence des termes pour indiquer l'importance relative des caractéristiques (Mejova & Srinivasan, 2011).

#### **3.2.2 Les parties du discours (POS)**

La partie du discours (POS) de chaque mot peut également être importante. Les mots de différentes parties du discours (POS) peuvent être traités différemment. Par exemple, il a été démontré que les adjectifs sont des indicateurs importants des avis. Ainsi, certains chercheurs ont traité les adjectifs

comme des caractéristiques spéciales. Cependant, on peut aussi utiliser toutes les étiquettes POS et leurs n-grammes comme caractéristiques (B. Liu, 2012).

### 3.2.3 Les mots et expressions d'opinion

Ce sont des mots couramment utilisés pour exprimer des opinions, bonnes ou mauvaises, similaires ou différentes. Par ailleurs, certaines phrases expriment des opinions sans utiliser de mots d'opinion. Par exemple : me coûte.

### 3.2.4 Les négations

L'apparition de mots négatifs peut changer l'orientation de l'opinion, exemple : car « pas bon » équivaut à « mal ». Cette caractéristique est considérée comme l'une des caractéristiques importantes car elle modifie complètement le sens de la phrase.

## 3.3 Le traitement du langage naturel (TLN (NLP))

Selon (liu, 2012), le traitement automatique du langage naturel (NLP<sup>20</sup>) est l'élément qui fait le lien entre la conversation humaine et la compréhension programmée par ordinateur (quels que soient les humains qui parlent, quelle langue est parlée et la façon dont ils parlent, peut-être grammaticalement). Le composant NLP permet à l'ordinateur d'interpréter le langage humain vaste et compliqué, de comprendre ce qui est dit, de tout traiter, de refléter ce qui est demandé et de réagir efficacement, comme le font les humains.

Nous présentons quelques exemples pratiques que l'on rencontre dans notre vie quotidienne pour mieux comprendre :

- **Les spams** : toutes les boîtes mails utilisent un filtre anti-spam et cela fonctionne avec le filtrage bayésien en référence au théorème de Bayes qui est une technique statistique de détection de spam. Ces filtres vont "comprendre" le texte et trouver s'il y a des corrélations de mots qui indiquent un pourriel.
- **Google Traduction** : nous sommes probablement tous utilisés ce système et leur technologie utilise de nombreux algorithmes. Ici, le défi n'est pas de traduire le mot, mais de garder le sens d'une phrase dans une autre langue.

---

<sup>20</sup> NLP : Natural Language Processing

### 3.4 Les problèmes de traitement du langage naturel

Le NLP ne suffit pas pour résoudre tous les problèmes de l'analyse des sentiments ou la fouille d'opinion parce que c'est un domaine multidisciplinaire, elle incorpore de nombreux sous domaines telle que la linguistique informatique, l'apprentissage machine, la sémantique, etc.

Il ne faut pas oublier que l'analyse des sentiments est l'un des problèmes de NLP le plus populaire. Plusieurs problèmes ne peuvent pas être résolus en NLP, et la fouille d'opinion, l'un de ces problèmes extrêmement limités, Résultant en passant à d'autre domaine (technique).

### 3.5 L'apprentissage par la machine

L'apprentissage automatique est un sous-ensemble de l'intelligence artificielle qui offre aux ordinateurs la possibilité d'apprendre sans être explicitement programmé. En apprentissage machine, il n'est pas nécessaire de définir explicitement toutes les étapes ou conditions comme n'importe quelle autre application de programmation. Au contraire, la machine est formée à un jeu de données d'apprentissage suffisamment grand pour créer un modèle, ce qui permet à la machine de prendre des décisions en fonction de son apprentissage.

L'apprentissage automatique est divisé en trois catégories principales :

- **Apprentissage supervisé** : Dans ce paramètre, nous fournissons à un modèle un ensemble de données de formation composé des données et des libellés réels. À partir de ces données, un modèle est appris. Évidemment ici, plus les données sont nombreuses, plus le modèle fonctionnera mieux avec des exemples inconnus. Les facteurs clés ici sont que le modèle devrait généraliser à des exemples invisibles. La classification, la régression et le classement relèvent de ce paradigme d'apprentissage.
- **Apprentissage non supervisé** : dans ce paramètre, le modèle organise automatiquement les données sans aucun exemple étiqueté. Les exemples de ceci incluent le regroupement (clustering) de données, la synthèse, etc. Les données sont organisées en fonction de la similarité des éléments de données.
- **Apprentissage semi-supervisé** : Il s'agit d'un hybride entre non supervisé et supervisé, dans lequel certaines des données sont étiquetées, alors qu'un grand groupe de données est en réalité non étiqueté. Les algorithmes prennent ici la totalité des données étiquetées et utilisent intelligemment les données non étiquetées pour améliorer les performances du modèle.

### 3.6 Définition de l'apprentissage approfondi (en profondeur)

De nos jours les réseaux de neurones sont devenus un sujet de recherche très populaire dans le domaine du traitement du langage naturel, y compris l'analyse des sentiments. Ils s'avèrent utiles pour résoudre presque tout problème de classification d'apprentissage automatique. Le seul ajustement requis est la définition de l'architecture : le nombre de couches cachées à utiliser, le nombre d'unités cachées présentes dans chaque couche, la fonction d'activation de chaque nœud, le seuil d'erreur pour les données, le type d'interconnexion, etc. Une fois qu'une architecture de réseau de neurones appropriée est conçue pour le problème à résoudre, une solution au problème de classification peut être obtenue à l'aide de modèles d'apprentissage en profondeur. La seule exigence pour ces modèles d'apprentissage en profondeur consiste en suffisamment de données de formation, de temps et de ressources pour former le réseau approprié ce qui permet de surmonter les défis de l'extraction de caractéristique rencontrés dans le paradigme traditionnel de l'apprentissage par la machine.

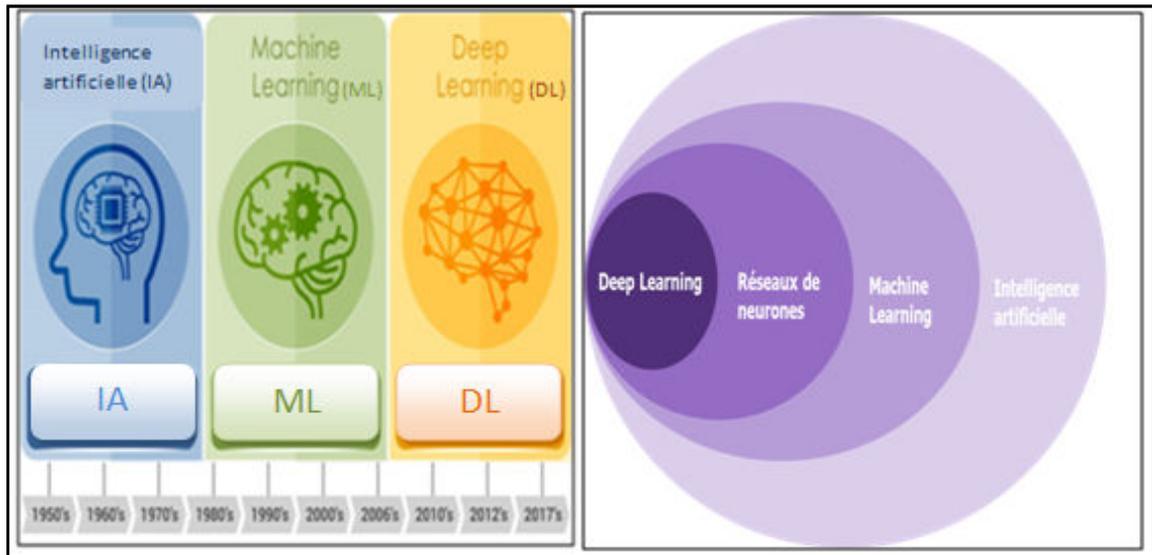
L'apprentissage en profondeur est défini par Deng et Yu (L. Deng & Yu, 2014) et illustré dans la figure 3.1 comme étant une discipline du Machine Learning qui regroupe des méthodes d'apprentissage automatique basées sur des réseaux de neurones profonds avec des algorithmes qui tentent de modéliser des abstractions de haut niveau dans les données en utilisant des architectures des modèles avec des structures complexes, composées de multiples transformations non linéaires.

L'intelligence artificielle (IA) est la principale discipline qui couvre tout ce qui a trait à la fabrication de machines intelligentes. L'apprentissage par la machine (ML<sup>21</sup>) est couramment utilisé aux côtés de l'IA, mais ce n'est pas la même chose. ML est un sous-ensemble de l'IA. ML fait référence à des systèmes pouvant apprendre par eux-mêmes. Des systèmes qui deviennent de plus en plus intelligents au fil du temps sans intervention humaine. Deep Learning (DL<sup>22</sup>) correspond à ML mais s'applique aux grands ensembles de données. La plupart du travail d'intelligence artificielle fait désormais appel à ML, car un comportement intelligent nécessite des connaissances considérables, et l'apprentissage est le moyen le plus simple d'obtenir ces connaissances. L'image ci-dessous illustre la relation entre IA, ML et DL.

---

<sup>21</sup> **ML** : Machine Learning

<sup>22</sup> **DL** : Deep Learning



**Figure 3.1** L'intelligence artificielle (IA), l'apprentissage par la machine (ML) et l'apprentissage en profondeur (DL) (Goodfellow et al., 2016)

### 3.7 Les avantages de l'apprentissage en profondeur

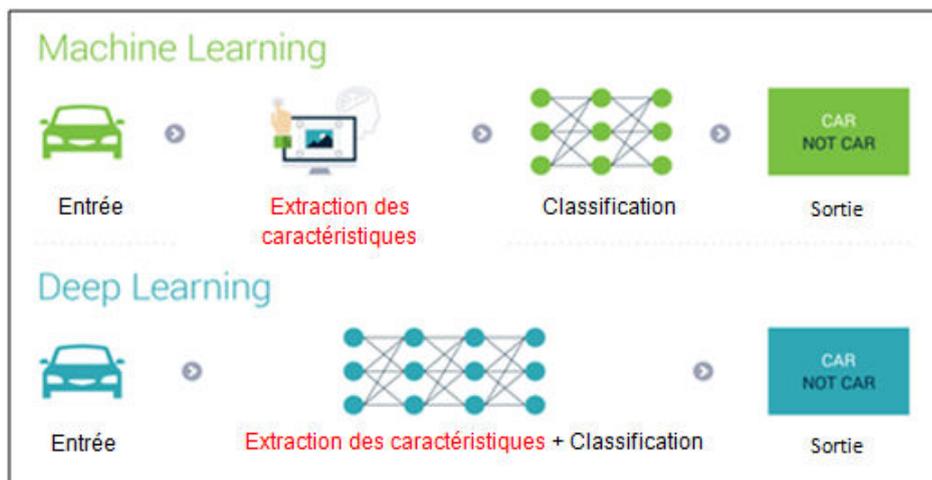
Bien que les algorithmes traditionnels d'apprentissage automatique tels que SVM ont fourni de bonnes performances et résultats dans diverses tâches de la langue naturelle au cours des dernières décennies, ils présentent quelques inconvénients où les modèles d'apprentissage approfondi ont le potentiel de surmonter ces limitations dans une large mesure. Parmi les avantages des réseaux de neurones profonds nous citons (Singhal & Bhattacharyya, 2016; Zhang, Wang, & Liu, 2018) :

- La force des modèles d'apprentissage en profondeur n'exige pas l'extraction des caractéristiques (les vecteurs de caractéristiques) soigneusement optimisées et conçues à la main (figure 3.2). Au lieu des caractéristiques, elles prennent en entrée des mots incorporant des informations de contexte et les couches intermédiaires du réseau de neurones apprennent les caractéristiques au cours de la phase de l'apprentissage. Cela signifie qu'une nécessité, qui est à la base des modèles de classification traditionnels, n'est plus nécessaire pour le fonctionnement des modèles d'apprentissage en profondeur.
- Un apprentissage en profondeur permet un bon apprentissage de la représentation. Alors que les représentations de caractéristiques du texte saisi peuvent être apprises automatiquement à partir des données d'apprentissage pour une tâche particulière



comme illustré dans la figure 3.2. Les représentations des mots contenant des informations de contexte, peuvent être apprises à partir du corpus brut de manière non supervisée. Cela ne tient pas compte de la nécessité de construire manuellement des caractéristiques appropriées ou des informations sur les mots.

- La capacité à s'adapter aux variations de tâches avec de très petits changements dans le système lui-même ajoute une touche supplémentaire au paradigme d'apprentissage en profondeur.



**Figure 3.2** La différence entre apprentissage par la machine (ML) et l'apprentissage en profondeur (DL) (Yann LeCun et al., 2015) (Goodfellow et al., 2016)

Ces dernières années, les modèles d'apprentissage en profondeur ont été largement appliqués dans le domaine de la PNL et montrent de grands potentiels. Dans les sections suivantes, nous décrivons brièvement les principales architectures d'apprentissage en profondeur et les techniques associées qui ont été appliquées aux tâches de traitement de la langue naturelle.

### 3.8 Les modules de base d'apprentissage en profondeur

#### 3.8.1 Les mots plongés (Word Embedding)

Pour de nombreuses applications d'apprentissage en profondeur dans le traitement du langage naturel, les mots plongés sont les éléments de base des réseaux de neurones profonds. Cependant, le concept du plongement de mots (word embedding) n'est pas nouveau. L'idée pourrait remonter aux années 1950 : proposée par Firth (Firth, 1957). On l'appelle aussi un modèle de distribution sémantique (DSM) ou représentation distribuée (Hinton, McClelland, & Rumelhart, 1984), car il est supposé que la signification d'un mot peut être apprise par sa distribution dans le texte.

Aujourd'hui, les mots plongés (words embedding) sont représentés par des vecteurs de grande dimension capables de capturer des informations sémantiques et syntaxiques à partir de texte. Les premiers modèles de distribution sémantique (DSMs), tels que l'analyse sémantique probabiliste latente (Hoffman, 1999) et l'analyse latente de Dirichlet (Blei, Ng, & Jordan, 2003), ont été développés dans les années 1990, suivis par des modèles de plongement (embedding models) basés sur un réseau de neurones tels que l'apprentissage de la représentation distribuée des mots (Bengio, Ducharme, Vincent, & Jauvin, 2003; E. H. Huang, Socher, Manning, & Ng, 2012) et le modèle d'apprentissage plongé (embedding Learning model) des tâches multiples SENNA (Collobert & Weston, 2008). La prédominance de l'incorporation des mots plongés (word embedding) a commencé par CBOW (sac de mots) (Tomas Mikolov, Chen, Corrado, & Dean, 2013; Tomas Mikolov, Sutskever, et al., 2013) et GloVe (Pennington, Socher, & Manning, 2014) en 2013 et 2014.

Le « plongement de mots » ou « plongement lexical » (word embedding en langue anglaise) est une méthode se focalisant sur l'apprentissage d'une représentation de mots. Cette technique permet de représenter chaque mot d'un dictionnaire par un vecteur de nombres réels correspondant. Ceci facilite notamment l'analyse sémantique des mots.

Les plongements de mots constituent notamment une méthode pour traiter un problème récurrent en intelligence artificielle. En effet, sans les plongements de mots, les objets mathématiques utilisés pour représenter les mots ont typiquement un grand nombre de dimensions, tant et si bien que ces objets se retrouvent « isolés » et deviennent épars. La technique des word embeddings diminue le nombre de ces dimensions, facilitant ainsi les tâches d'apprentissage impliquant ces mots. Les “word embeddings” permettent de représenter sous forme d'un vecteur chaque mot (ou terme) d'un corpus, en utilisant les termes qui apparaissent autour du mot, aussi appelés contexte. Cette représentation permet d'identifier les termes qui sont utilisés dans le même contexte. Cette technique est couramment utilisée dans le domaine de traitement automatique du langage naturel. Elle est encore relativement peu employée dans le domaine de la recherche d'information.

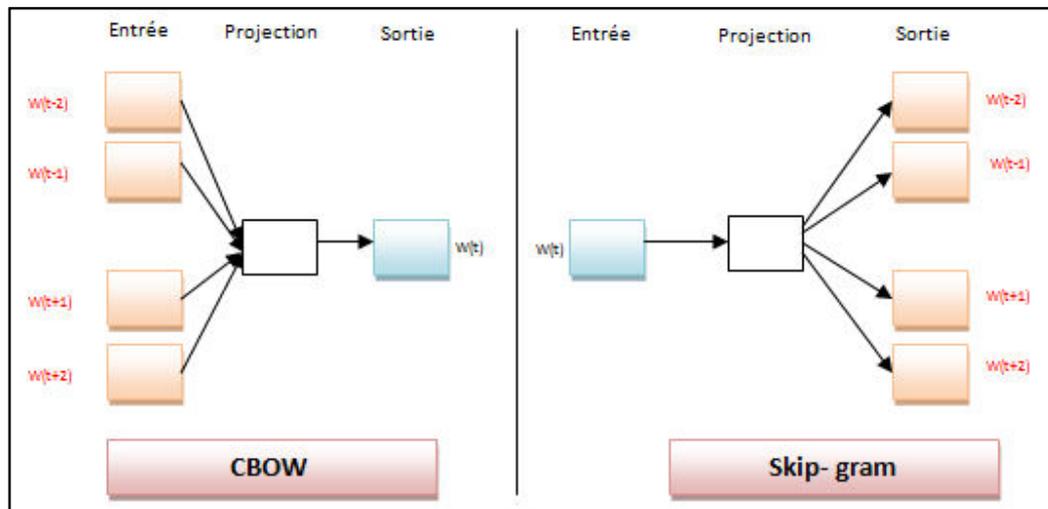
Depuis les travaux de (Tomas Mikolov, Sutskever, et al., 2013) introduisant la technique de représentation vectorielle des termes appelée « word2vec » qui sera présentée en détaille dans les sections suivantes. Plusieurs travaux se sont intéressés à utiliser cette technique pour la recherche d'information (ALMasri, Berrut, & Chevallet, 2016; Kuzi, Shtok, & Kurland, 2016; Zamani & Croft, 2016) ou la classification (Balikas & Amini, 2016; Y. Kim, 2014a, 2014b).

Les réseaux de neurones dans le traitement de la langue naturelle, contrairement à d'autres algorithmes traditionnels, ne prennent pas de mots bruts en entrée, car les réseaux ne peuvent comprendre que des nombres et des fonctions. Par conséquent, les mots doivent être transformés en vecteurs de caractéristiques ou, en d'autres termes, en plongements de mots (Tomas Mikolov, Sutskever, et al., 2013), qui capturent les caractéristiques et la sémantique des mots s'ils sont correctement extraits. Les présentations vectorielles des mots peuvent être apprises en introduisant un grand corpus brut dans un réseau et en l'entraînant suffisamment de temps.

Les chercheurs de la communauté (Tomas Mikolov, Sutskever, et al., 2013) ont contribué dans le but de présenter deux modèles de réseau de neurones différents pour la création des mots prolongés (figure 3.3).

Les modèles étaient de nature log-linéaire, formés sur de grands corpus. L'un d'entre eux est un modèle basé sur un sac de mots appelé CBOW (Continuous Bag Of Word) ; il utilise le contexte du mot afin d'obtenir le mot imbriqué ou plongés (word embedding). L'autre s'appelle le modèle skip-gram ; il prédit les mots imbriqués des mots de contexte étant donné le mot actuel. Ces chercheurs ont rendu public un modèle appelé word2vec qui est sans doute le modèle d'incorporation de mots le plus populaire de nos jours, il comprend les deux architectures (CBOW et skip-gram) montrées dans la figure 3.3. Word2vec est une approche basée sur un apprentissage de représentation vectorielle des termes en utilisant des réseaux de neurones. L'idée principale est de prédire un terme sachant un contexte, ou bien étant donné un terme, prédire son contexte. Plus précisément, les auteurs de word2vec, proposent une architecture de réseau de neurones qui consiste en trois couches : une couche d'entrée, une couche de projection et une couche de sortie pour la prédiction des termes. Chaque représentation vectorielle d'un terme est entraînée afin de maximiser les probabilités que ces termes être proches du corpus.

Le modèle CBOW a pour objectif d'obtenir la représentation appropriée d'un mot, en se basant sur une fenêtre de mots adjacents ou le contexte, afin de retrouver le prochain mot du même contexte. Par contre le modèle Skip-Gram a pour objectif d'obtenir le vecteur du mot permettant d'exprimer le même sens que le mot en cours, i. e. : la représentation qui permet de déduire son contexte.



**Figure 3.3:** Description des modèles CBOW et Skip-gram (Tomas Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)

La fonction objective du modèle skip-gram fait la somme des probabilités des mots du contexte on donnant le mot cible dans une fenêtre prédéfinie :

$$\frac{1}{N} = \sum_i \sum_{j \in n(i)} \log p(w_j/w_i) \dots\dots\dots(07)$$

Le Skip-gram modèle a une fonction softmax définie comme suit :

$$p(w_j/w_i) = \frac{\exp(\mathbb{T}_{w_j}^T \vartheta_{w_i})}{\sum_K \exp(\mathbb{T}_{w_K}^T \vartheta_{w_i})} \dots\dots\dots(08)$$

Où  $w_i$  et  $w_j$  sont les présentations vectorielles des mots plongées (word embedding vectors)

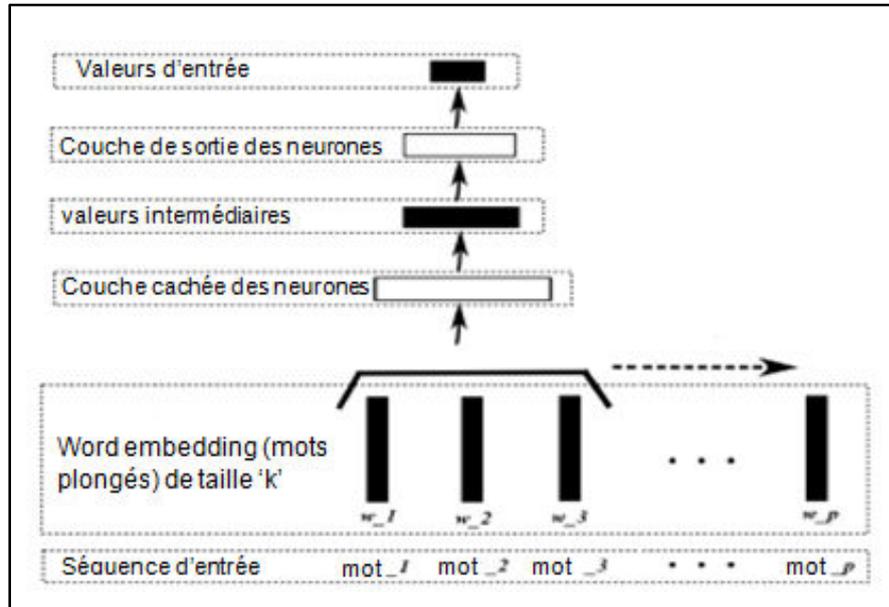
Le modèle CBOW a une fonction softmax définie comme suit :

$$P(w_{y_j/w_1, \dots, w_c}) = y_j = \frac{\exp(u_j)}{\sum_{j=1}^V \exp(u_j)} \dots\dots\dots(09)$$

### 3.8.2 L'architecture du réseau de neurone pour le traitement du langage naturel

Généralement, pour les tâches de traitement de la langue naturelle, nous avons tendance à utiliser l'approche par la fenêtre (Collobert, Weston, et al., 2011). Cette méthode suppose que l'étiquette à attribuer à un mot dans une phrase dépend des mots voisins. Par conséquent, une taille de fenêtre fixe

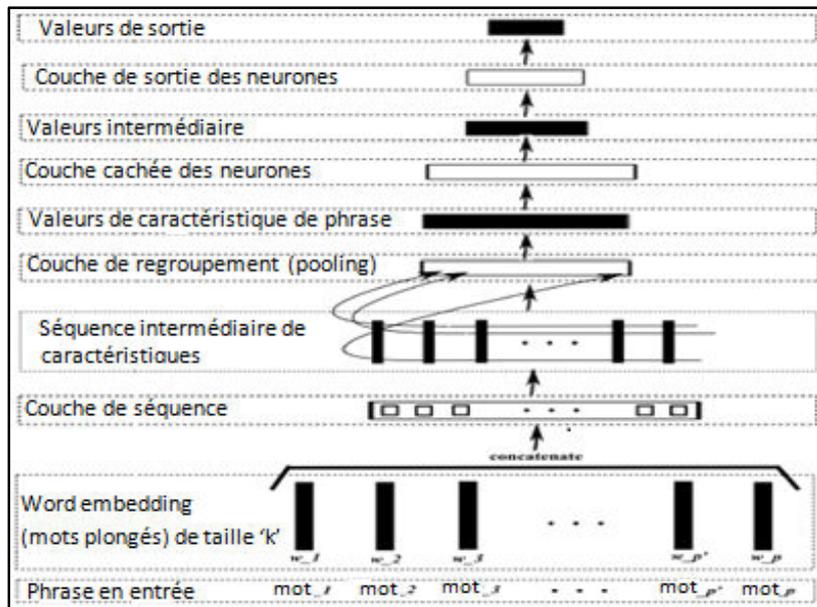
(hyper-paramètre supplémentaire) est choisie et cette quantité de mots est introduite dans le réseau pour marquer le mot du milieu.



**Figure 3.4:** Réseau de neurone pour traitement de la langue naturelle au niveau fenêtre (Collobert, Weston, et al., 2011).

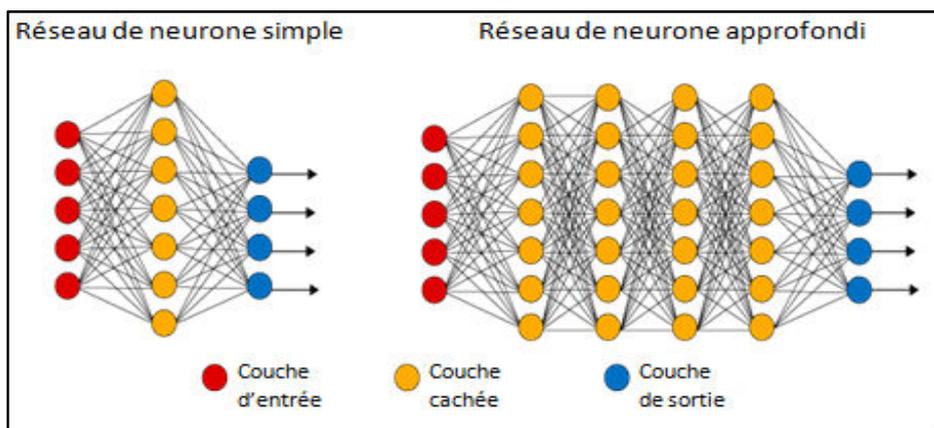
L'approche au niveau de la fenêtre ne peut pas être appliquée à des tâches au niveau macro-texte telles que la classification des sentiments, car cette dernière exige que la phrase entière soit prise en compte, alors que dans l'approche au niveau de la fenêtre, une partie seulement de la phrase est considérée à la fois. En outre, pour d'autres tâches du traitement de la langue naturelle, un mot peut dépendre d'un mot qui n'est pas présent dans la fenêtre prédéterminée. Par conséquent, l'approche au niveau de la phrase est une alternative viable, qui prend les vecteurs des caractéristiques de tous les mots dans le texte d'entrée en entrée (Collobert, Weston, et al., 2011).

La figure 3.4 montre la structure générale du réseau qui prend la phrase entière en entrée. La couche de séquence peut avoir différentes structures pour gérer la séquence de mots dans le texte. Pour la classification des phrases, les phrases pouvant être de taille variable, il existe une couche de regroupement (pooling layer) après la couche de séquence (sequence layer), qui présente une carte de caractéristique de taille fixe. Celui-ci est ensuite introduit dans des couches de neurones entièrement connectées (fully connected) pour produire une seule étiquette pour la phrase entière.



**Figure 3.5:** Réseau de neurone pour traitement de la langue naturelle au niveau phrase (Collobert, Weston, et al., 2011).

Le nombre de couches, le nombre d'unités dans chaque couche, la structure de la couche de séquence, la dimensionnalité des vecteurs de mots, les interconnexions et les fonctions d'activation sont des hyper-paramètres du modèle de réseau de neurone, qui doivent être définis prudemment pour aboutir à des meilleures performances pour une tâche donnée. La différence principale entre un réseau de neurone simple et un réseau profond est bien le nombre des couches cachées cela est illustré dans la figure 3.6.



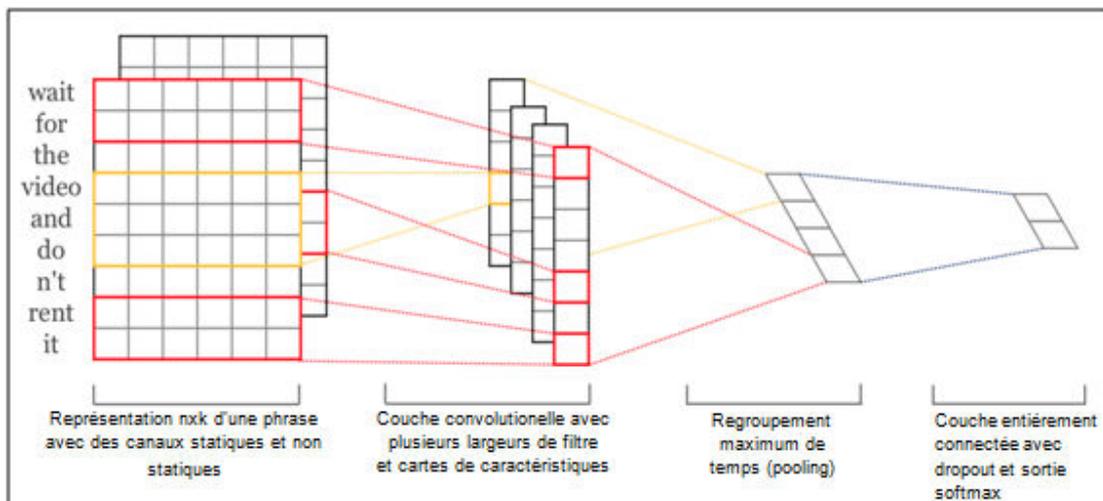
**Figure 3.6:** Architecture d'un réseau de neurone simple Vs réseau de neurone profondi (Yann LeCun, Yoshua Bengio, & Geoffrey Hinton, 2015).

### 3.8.3 Les réseaux de neurone convolutionnels (CNN)

Le réseau convolutionnel (CNN) a été une architecture essentielle pour les chercheurs en apprentissage en profondeur. Il a une longue histoire, qui remonte à l'étude des cellules dans le cortex visuel animal en 1959 (Hubel & Wiesel, 1959). Dans les années 1990, les chercheurs de la communauté ont inventé et amélioré le premier CNN pratique nommé LeNet-5 (LeCun et al., 1990; LeCun et al., 1998), pour la reconnaissance de chiffres manuscrite.

Il n'a pas retenu beaucoup d'attention par rapport à SVM (Cortes & Vapnik, 1995) et aux modèles graphiques probabilistes (Koller & Friedman, 2009) à cette époque. En 2011 (Collobert, Weston, et al., 2011) a proposé une architecture de réseau de neurones unifiée pouvant être appliquée à de nombreuses tâches de traitement du langage naturel, telles que le balisage partiel de la parole, l'analyse syntaxique, le découpage en bloc, le marquage de rôle sémantique et la reconnaissance d'entités nommées.

Cette architecture prend en entrée les vecteurs de mots concaténés du texte et implique des couches de convolution et de pooling maximal pour avoir le modèle général du réseau de neurones. Jusqu'au formidable succès d'AlexNet (Krizhevsky et al., 2012).



**Figure 3.7:** Architecture du réseau de neurone convolutionnel (Y. Kim, 2014b)

Une couche de convolution contient plusieurs noyaux (ou filtres) permettant de calculer des cartes de caractéristiques. Contrairement aux couches entièrement connectées où chaque neurone est connecté à tous les neurones des couches précédentes, un neurone convolutionnel n'est connecté qu'à

quelques neurones dans une petite région appelée champ récepteur (connexions locales). Ensuite, une fonction d'activation par élément est appliquée.

Les modèles basés sur les réseaux de neurone utilisent généralement une fonction d'activation non linéaire telle que Relu, Tanh ou sigmoïde et calculent les pertes (la différence entre la sortie réelle et la sortie calculée) telles que l'erreur moyenne carrée (MSE), Logloss. Cette perte est propagée en arrière pour ajuster les poids et la formation pour minimiser les pertes ou rendre les modèles plus précis.

Supposons que la caractéristique d'entrée soit une matrice X, un neurone convolutionnel 2D calcule une nouvelle caractéristique à l'emplacement (i, j) comme suit :

$$z_{i,j} = \left( \sum_{m=-k/2}^{k/2} \sum_{n=-k/2}^{k/2} X_{i+m,j+n} W_{m+k/2+1,n+k/2+1} + b \right) \dots \dots \dots (10)$$

Il est souvent écrit comme  $Z = X W + b$ , où  $\cdot$  dénote l'opération de convolution,  $b$  est un biais scalaire et  $f$  est une fonction d'activation non linéaire, telle que la fonction Relu ou Tanh.

Une couche de regroupement maximal (max-pooling) prend le maximum de valeurs dans une région.

$$y_{i,j} = \max\{Z_{m,n}\}, (m,n) \in R_{i,j} \dots \dots \dots (11)$$

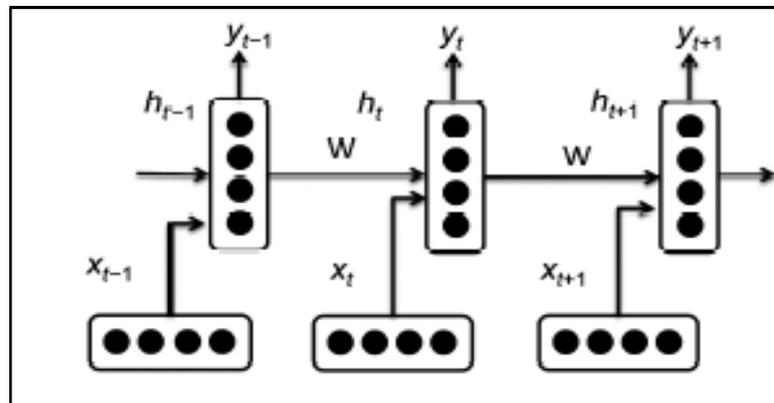
Où  $R$  est une région locale autour de l'emplacement (i, j), qui est généralement une région carrée. Plusieurs couches convolutionnelles et couches de regroupement (pooling layers) peuvent être empilées de manière spéciale. Enfin, il y a plusieurs couches entièrement connectées (fully connected) entre eux.

Les idées clés de CNN sont les connexions locales, les poids partagés, le pooling et l'utilisation de nombreuses couches (Y LeCun, Y Bengio, & G Hinton, 2015). Tous les neurones qui calculent un type de carte de caractéristiques (feature map) partagent un seul ensemble de poids de convolution. Le partage de poids réduit la complexité du modèle ainsi que le nombre de paramètres. Cela correspond à l'intuition qu'un bon détecteur doit être appliqué largement sur une grande surface.



### 3.8.4 Les réseaux de neurone récurrent (RNN)

Les réseaux de neurones récurrents (Tomáš Mikolov et al., 2010) sont une forme de réseaux de neurones qui ne dépendent pas de la taille de la fenêtre pour manipuler les différentes tâches de traitement du langage naturel. RNN est capable de paramétrer le réseau sur toutes les entrées précédemment vues (mots dans le cas d'une phrase). En plus de la dépendance à l'entrée actuelle, la valeur de chaque unité de couche masquée dépend également de son état antérieur, propageant ainsi les effets des mots sur la phrase. Les réseaux de neurones récurrents (RNN), sont spécialement conçus pour l'apprentissage de séquences de données de longueur arbitraire, car ils analysent séquentiellement toute la séquence et en génèrent une forme réduite en sortie.



**Figure 3.8:** Architecture du réseau de neurone récurrent (Tomáš Mikolov, Karafiát, Burget, Černocký, & Khudanpur, 2010)

La présentation vectorielle de chaque mot est introduite dans le réseau une par une et l'effet de chaque mot est appliqué jusqu'à la fin de la phrase, garantissant ainsi que la dépendance de chaque mot sur tous les autres mots est capturée par le biais d'activations de neurones et la rétro-propagation sur les matrices de poids.

Le modèle RNN fonctionne en propageant les matrices de poids sur les pas de temps. Cependant, cela crée des anomalies qui ne sont pas acceptables dans la pratique. Intuitivement, on devrait pouvoir prédire un mot avec plus de précision compte tenu du contexte (c'est-à-dire qu'il y a plus de mots qui précèdent) par rapport à un contexte moins important. Cependant, RNN a tendance à faire le contraire en raison des problèmes de gradient de disparition (vanishing gradient problem) et d'explosion de gradient. Le signal de gradient peut avoir pour résultat de multiplier un grand nombre de fois (selon le nombre de pas de temps) par la matrice de poids associée aux interconnexions entre les

neurones de la couche cachée récurrente. Cela implique que la magnitude des poids peut avoir un impact important sur le processus d'apprentissage.

Comme nous avons déjà introduit au début de la séquence, les modèles fondés sur les réseaux neuronaux d'apprentissage en profondeur, tels que les réseaux de neurones récurrents (RNN), sont spécialement conçus pour l'apprentissage de séquences de données de longueur arbitraire, car ils analysent séquentiellement toute la séquence et en génèrent une forme réduite, où le processus d'apprentissage des RNN se fait au niveau des nœuds cachés en fonction de leurs couches de nœuds précédentes.

En théorie, les modèles RNN sont capables de se souvenir des dépendances à longue distance. Mais, à mesure que la séquence s'allonge, ces modèles souffrent du problème du gradient disparaissant (vanishing gradient problem) (Bengio, Simard, & Frasconi, 1994; Pascanu, Mikolov, & Bengio, 2013), ce qui entrave l'apprentissage de longues séquences de données. Les gradients portent les informations utilisées dans la mise à jour des paramètres RNN. Lorsque le gradient devient de plus en plus petit, les mises à jour des paramètres deviennent insignifiantes, ce qui signifie qu'aucun véritable apprentissage ne sera pas effectué.

Par conséquent, le modèle RNN, en raison de ses limitations inévitables, est inapproprié pour les tâches de traitement du langage naturel. Bien qu'il ait ses mérites, le manque de fiabilité de la rétrodiffusion d'informations le rend moins favori par les chercheurs de la communauté. Donc, la nécessité de modifier le réseau afin de préserver les avantages tout en atténuant les inconvénients. Certaines des versions améliorées est les modèles de mémoire à long court terme (LSTM) (Long short term memory).

### **3.8.5 Les réseaux de mémoire à long-court terme (LSTM)**

Pour surmonter les inconvénients des RNNs, des réseaux de neurones de mémoire à long court terme (LSTM) (Hochreiter & Schmidhuber, 1997) ont été suggérés. Ces réseaux sont une version modifiée des réseaux de neurones récurrents, mais avec des unités d'activation beaucoup plus complexes avec un mécanisme de déclenchement (gating mechanism) en incluant une cellule de mémoire constituée d'une unité avec des unités de grille d'entrée et de sortie multiplicatives.

Des portes d'entrée contre des entrées non pertinentes et des portes de sortie contre des contenus de mémoire non pertinents actuels pour propager de nouvelles informations et oublier certaines informations précédentes permettant aux gradients de se propager.

Ces modèles ont été suggérés comme solution au problème précité et ont démontré leur efficacité dans de nombreux problèmes liés au traitement du langage naturel (NLP).

Au lieu d'avoir une seule couche réseau dans un RNN, un réseau de type LSTM a quatre couches réseau en interaction comme décrit dans la figure 3.9. Sur cette figure, les zones grises représentent une couche de réseau et les cercles bleus sont des opérations ponctuelles telles que l'addition de vecteurs. La ligne horizontale à droite du schéma représente l'état de la cellule. Cet état parcourt toute la chaîne avec seulement des interactions linéaires simples. L'état de la cellule peut également être interprété comme la mémoire du réseau et peut supprimer ou ajouter des informations à l'aide de structures appelées portes.

À chaque pas de temps  $t$ , il y a un ensemble de vecteurs, comprenant une porte d'entrée, une porte oubliée  $f_t$ , une porte de sortie  $o_t$  et une cellule de mémoire  $C_t$ . Tous ces éléments sont utilisés ensemble pour calculer la sortie de la couche cachée  $h_t$  comme suit :

$$f_t = \sigma (W_f \times x_t + U_f \times h_{t-1} + b_f) \dots\dots\dots(12)$$

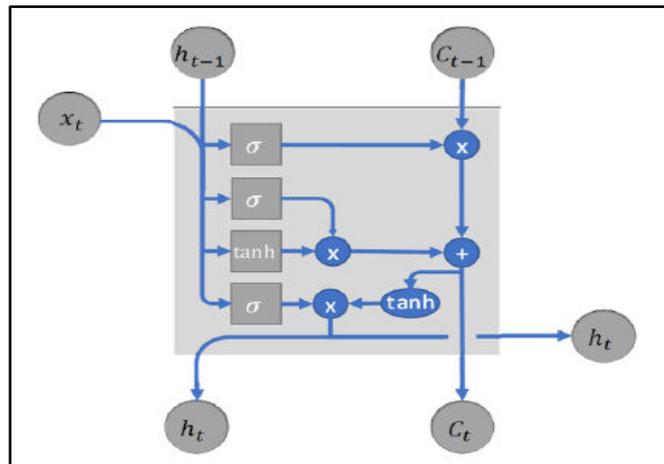
$$i_t = \sigma (W_i \times x_t + U_i \times h_{t-1} + b_i) \dots\dots\dots (13)$$

$$\check{C}_t = \tanh (W_c \times x_t + U_c \times h_{t-1} + b_c) \dots\dots\dots (14)$$

$$C_t = i_t \times \check{C}_t + f_t \times C_{t-1} \dots\dots\dots (15)$$

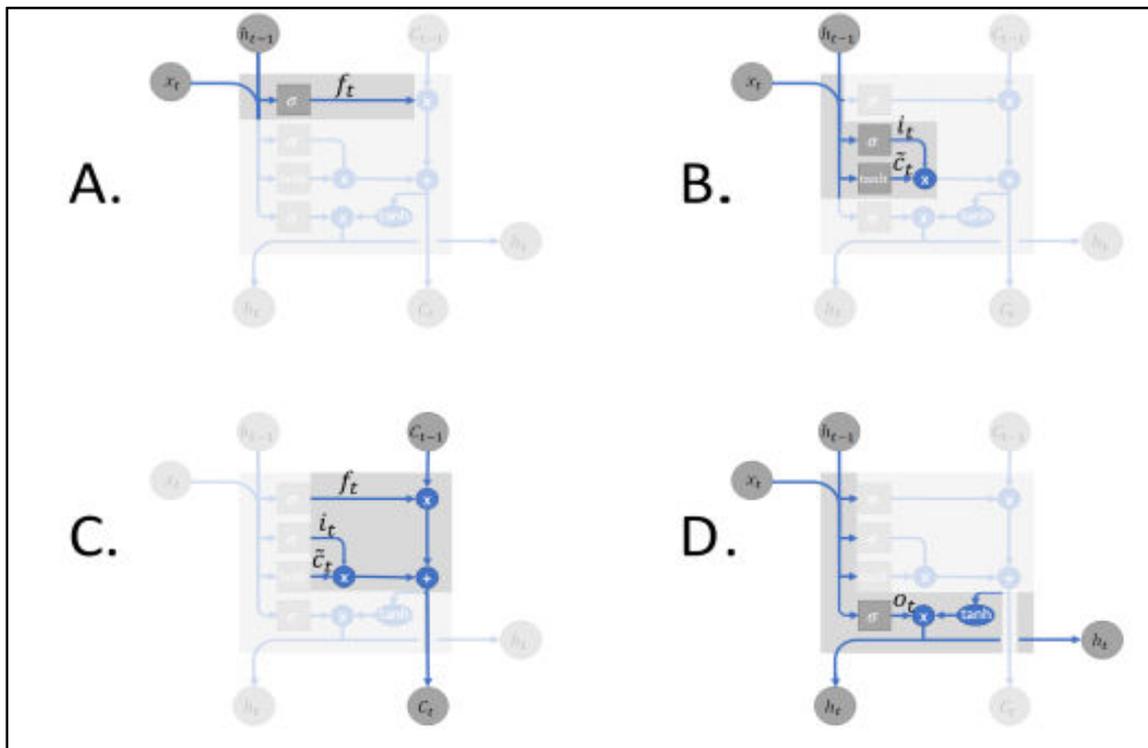
$$o_t = \sigma (W_o \times x_t + U_o \times h_{t-1} + b_o) \dots\dots\dots(16)$$

$$h_t = O_t \times \tanh(C_t) \dots\dots\dots (17)$$



**Figure 3.9:** Architecture d'un réseau de neurone de type LSTM (Hochreiter & Schmidhuber, 1997).

Les différentes étapes du comportement d'un réseau de neurone de type LSTM sont synthétisées dans la figure 3.10. Chaque étape du réseau LSTM sera discutée ci-après.

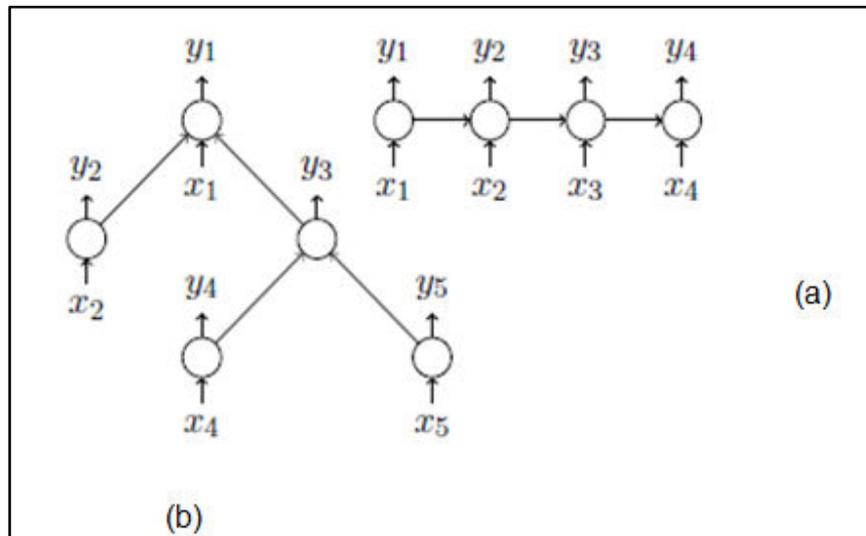


**Figure 3.10:** Comportement d'un réseau de neurone type LSTM (Hochreiter & Schmidhuber, 1997).

- **Etape 01 :** Tout d'abord, le modèle doit déterminer l'information il faut oublier à partir de l'état de cellule, cette partie du réseau est illustrée à la figure 3.10 (A) et à l'équation (3). C'est ce que l'on appelle les valeurs de la porte oubliée  $f_t$ . L'entrée à cette étape est la sortie de l'étape précédente  $h_{t-1}$  et l'entrée  $x_t$ . Une fonction d'activation sigmoïde est utilisée pour donner des valeurs de sortie comprises entre 0 et 1, 0 correspondant à « ne rien laisser passer » et 1 à « tout mémoriser ».
- **Etape 02 :** L'étape suivante consiste à déterminer quelles informations vont être ajoutées à l'état de la cellule, comme illustré à la figure 3.10 (B) et aux équations (4) et (5). Dans cette étape, les entrées sont  $h_{t-1}$  et  $x_t$ . La porte de la couche d'entrée applique d'abord une fonction sigmoïde sur l'entrée pour déterminer quelles parties de l'état de la cellule seront mises à jour. Ensuite, une fonction Tanh est utilisée pour créer de nouvelles valeurs candidates  $\check{C}_t$ . Dans l'étape suivante, ces deux éléments seront combinés pour mettre à jour l'état de la cellule,  $C_t$ .
- **Etape 03 :** A ce niveau-là, l'ancien état de la cellule est multiplié par  $f_t$ , pour oublier les informations qui ne sont plus nécessaires et les nouvelles informations sont ajoutées à la mémoire d'état de la cellule. Cette partie du réseau est illustrée dans la figure 3.10 (C) et à l'équation (6).
- **Etape 04 :** Dans la dernière étape, la sortie  $h_t$  sera calculée, cela est illustré dans la figure 3.10 (D) et aux équations (7) et (8). Cette sortie est basée sur l'état de la cellule mais dans une version filtrée. Tout d'abord, une fonction sigmoïde est appliquée à la sortie précédente  $h_{t-1}$  et à l'entrée  $x_t$ , afin de déterminer les valeurs de la porte de sortie  $o_t$ . C'est une valeur comprise entre 0 et 1 indiquant quelles parties de l'état de la cellule vont être sorties. Ensuite, l'état de la cellule  $C_t$  est transformé par une fonction Tanh pour obtenir des valeurs comprises entre -1 et 1. Ces valeurs transformées de l'état de la cellule sont ensuite multipliées par les valeurs de la porte de sortie pour aboutir à la sortie  $h_t$ . Cette sortie sera l'entrée de la prochaine étape du réseau. Dans ce qui suit nous allons présenter les différents types de réseau de neurone de type LSTM.

### 3.8.5.1 Les différents types du réseau LSTM

Selon la structure des données en entrée les chercheurs de la communauté ont catégorisé les réseaux de neurone de type LSTM en de groupe comme suit :



**Figure 3.11:** Un réseau LSTM standard (a) & Un réseau LSTM arborescent avec facteur de branchement arbitraire  $N = 2$  (b) (Tai, Socher, & Manning, 2015).

#### 3.8.5.1.1 LSTM sur des données linéaires :

Les deux variantes habituellement utilisées de l'architecture LSTM de base sont le LSTM bidirectionnel et le LSTM multicouche (également appelé LSTM empilé ou profond).

- **LSTM bidirectionnel :** Un LSTM bidirectionnel (Graves, Jaitly, & Mohamed, 2013) est constitué de deux LSTM exécutés en parallèle : l'un sur la séquence d'entrée et l'autre au verso de la séquence d'entrée. À chaque pas de temps, l'état caché du LSTM bidirectionnel est la concaténation des états cachés en avant et en arrière. Cette configuration permet à l'état masqué de capturer des informations passées et futures.
- **LSTM multicouche :** Dans les architectures LSTM multicouches, l'état caché d'une unité LSTM dans la couche (c) est utilisé comme entrée de l'unité LSTM dans la couche (c + 1) au même pas de temps (Graves, Jaitly, et al., 2013; Sutskever, Vinyals, & Le, 2014; Zaremba & Sutskever, 2014). Ici, l'idée est de laisser les couches supérieures capturer les dépendances à long terme de la séquence en entrée.

Ces deux variantes peuvent être combinées en un LSTM bidirectionnel multicouche (Graves, Jaitly, et al., 2013).

#### 3.8.5.1.2 LSTM sur des données arborescentes

Bien que le problème du gradient dans les réseaux RNNs soit résolu par le mécanisme de déclenchement du LSTM (gating mechanism), il reste limité pour le traitement de données linéaire, car un LSTM standard traite les informations de manière séquentielle alors que les architectures arborescentes (Tree-LSTM) de LSTM permettent la représentation de phrases par le biais d'une structure syntaxique.

En outre, une phrase en langue naturelle code des informations supplémentaires qui sont plus qu'une séquence de mots et régulièrement représentées dans des données de structure arborescente. Ce dernier soulève le problème de l'incompatibilité des données entre le LSTM standard et le LSTM en arborescence (Tree-LSTM). Pour faire face à ce type de problème les chercheurs de la communauté (Tai et al., 2015) en présenter deux nouvelles architectures basées sur le fonctionnement de base d'un LSTM standard mais sur des données sous forme syntaxique arborescente qui sont : le Child-Sum Tree-LSTM et le N-ary Tree-LSTM. Les deux variantes permettent de créer des topologies de réseau plus riches, dans lesquelles chaque unité LSTM peut intégrer des informations provenant de plusieurs unités fils (child).

Chaque unité Tree-LSTM (indexée par  $j$ ) contient des portes d'entrée et de sortie  $i_j$  et  $o_j$ , une cellule de mémoire  $c_j$  et un état masqué  $h_j$ . La différence entre les unités LSTM standards et les unités Tree-LSTM réside dans le fait que les mises à jour des vecteurs de déclenchement et des cellules mémoire dépendent de l'état de nombreuses unités fils. De plus, au lieu d'une seule porte oubliée, l'unité Tree-LSTM contient une porte oubliée  $f_{jk}$  pour chaque fils  $k$ . Cela permet à l'unité Tree-LSTM d'incorporer de manière sélective les informations de chaque fils.

Comme avec le LSTM standard, chaque unité Tree-LSTM prend un vecteur d'entrée  $x_j$  qui est la représentation vectorielle d'un mot dans une phrase. Le mot d'entrée à chaque nœud dépend de la structure arborescente utilisée pour le réseau.

- **Child-Sum Tree-LSTM :**

Soit un arbre, soit  $C(j)$ , l'ensemble des fils du nœud  $j$ . Les équations de transition du réseau de type Child-Sum Tree LSTM sont les suivantes:

$$\check{h}_j = \sum_{k \in c(j)} h_k \dots\dots\dots (18)$$

$$i_j = \sigma(W^{(i)} x_j + u^{(i)} \check{h}_j + b^{(i)}) \dots\dots\dots (19)$$

$$f_{jk} = \sigma(W^{(f)} x_j + u^{(f)} h_k + b^{(f)}) \dots\dots\dots (20)$$

$$o_j = \sigma(W^{(o)} x_j + u^{(o)} \check{h}_j + b^{(o)}) \dots\dots\dots (21)$$

$$u_j = \tanh(W^{(u)} x_j + U^{(u)} \check{h}_j + b^{(u)}) \dots\dots\dots (22)$$

$$C_j = i_j \odot u_j + \sum_{k \in c(j)} f_{jk} \odot C_k \dots\dots\dots (23)$$

$$h_j = O_j \odot \tanh(C_j) \dots\dots\dots (24)$$

Le mot d'entrée à chaque nœud dépend de la structure arborescente utilisée par le réseau. Dans le cas du réseau de type Child Sum Tree-LSTM le modèle manipule des arbres de dépendance où chaque nœud de l'arbre prend en entrée le vecteur correspondant au mot racine, tandis que dans un réseau de type N-ary Tree-LSTM le modèle parcourt des arbres de circonscription où les nœuds feuilles prennent en entrée les présentations vectorielles des mots.

▪ **N-ary Tree-LSTM**

Le N-ary Tree-LSTM peut être implémenté sur des structures arborescentes où le facteur de branchement est au plus N et où les fils sont ordonnés, c'est-à-dire qu'ils peuvent être indexés de 1 à N. Pour tout nœud j, l'état caché et la cellule mémoire de son k<sup>th</sup> fils comme h<sub>jk</sub> et c<sub>jk</sub> respectivement. Les équations de transition du réseau N-ary Tree-LSTM sont les suivantes:

$$h_t = O_t \odot \tanh(C_t) \dots\dots\dots (24)$$

$$O_t = \sigma(W_o X_t + U_o^L h_{t-1}^L + U_o^R h_{t-1}^R + b_o) \dots\dots\dots (25)$$

$$C_t = f_t^L \odot C_{t-1}^L + F_t^R \odot C_{t-1}^R + i_t \odot u_t + b_c \dots\dots\dots (26)$$

$$f_t^L = \sigma(W_f X_t + u_f^{LL} h_{t-1}^L + u_f^{LR} h_{t-1}^R + b_f) \dots\dots\dots (27)$$

$$f_t^R = \sigma(W_f X_t + u_f^{RL} h_{t-1}^L + u_f^{RR} h_{t-1}^R + b_f) \dots\dots\dots (28)$$

$$i_t = \sigma(W_i X_t + u_i^L h_{t-1}^L + u_i^R h_{t-1}^R + b_i) \dots\dots\dots (29)$$

$$u_t = \tanh(W_c X_t + u_c^L h_{t-1}^L + u_c^R h_{t-1}^R) \dots\dots\dots (30)$$



Dans le cas où  $N=2$ , le réseau N-ary Tree-LSTM va manipuler des arbres de circonscription binaires, dont les nœuds fils gauche et droit sont distingués. Notant que dans ce type de réseau, un nœud  $j$  reçoit un vecteur en entrée  $x_j$  uniquement s'il s'agit d'un nœud feuille.

### 3.9 La stratégie d'apprentissage des réseaux en profondeur

La rétropropagation (Backpropagation) est un algorithme critique utilisé pour la phase d'apprentissage des modèles basés sur les réseaux de neurone en profondeur. Généralement, un modèle d'apprentissage en profondeur a une fonction de perte  $L$  qui dépend des sorties du modèle avec un paramètre  $w$ . Nous avons besoin des dérivées partielles pour tout paramètre  $w$  afin de les optimiser avec la méthode de descente de gradient (gradient descent). La rétro-propagation accélère le processus de calcul des dérivées partielles des paramètres des millions de fois plus rapide.

Ce processus décrit comment :

- Les signaux d'erreur de la fonction de perte changent et se propagent à chaque paramètre ;
- Calculer les dérivées partielles pour les paramètres actuels.

Pour plus de détails, on pourrait se référer aux livres sur les réseaux de neurones (Goodfellow et al., 2016).

Une fois que nous obtenons les dérivées partielles de paramètres, nous pouvons mettre à jour les paramètres par descente de gradient stochastique (SGD).

Au début de l'apprentissage en profondeur, en raison de l'insuffisance des données étiquetées et des ressources informatiques, les réseaux de neurones ne fonctionnaient pas bien on se basant sur l'apprentissage par la rétro-propagation et l'initialisation aléatoire. (Bengio, Lamblin, Popovici, & Larochelle, 2007; Hinton et al., 2006) ont mis au point une stratégie d'apprentissage efficace consistant en une phase de pré- apprentissage avec des données non supervisées et une phase de mise au point globale (fine tuning stage) pour l'apprentissage de la machine de Boltzmann restreinte.

Le SGD et d'autres variantes sont couramment utilisés pour l'apprentissage de la plupart des réseaux de neurones. Le SGD effectue une mise à jour des paramètres pour chaque instance d'apprentissage  $x$  et  $y$ .

$$w_{t+1} = w_t - \alpha \nabla f(w; x_t, y_t) \dots \dots \dots (31)$$

Où  $w_t$  est la  $t^{\text{ième}}$  valeur du paramètre  $w$ ,  $f$  est la fonction concernant les paramètres du réseau, la taille des pas ou la vitesse d'apprentissage (Learning rate).

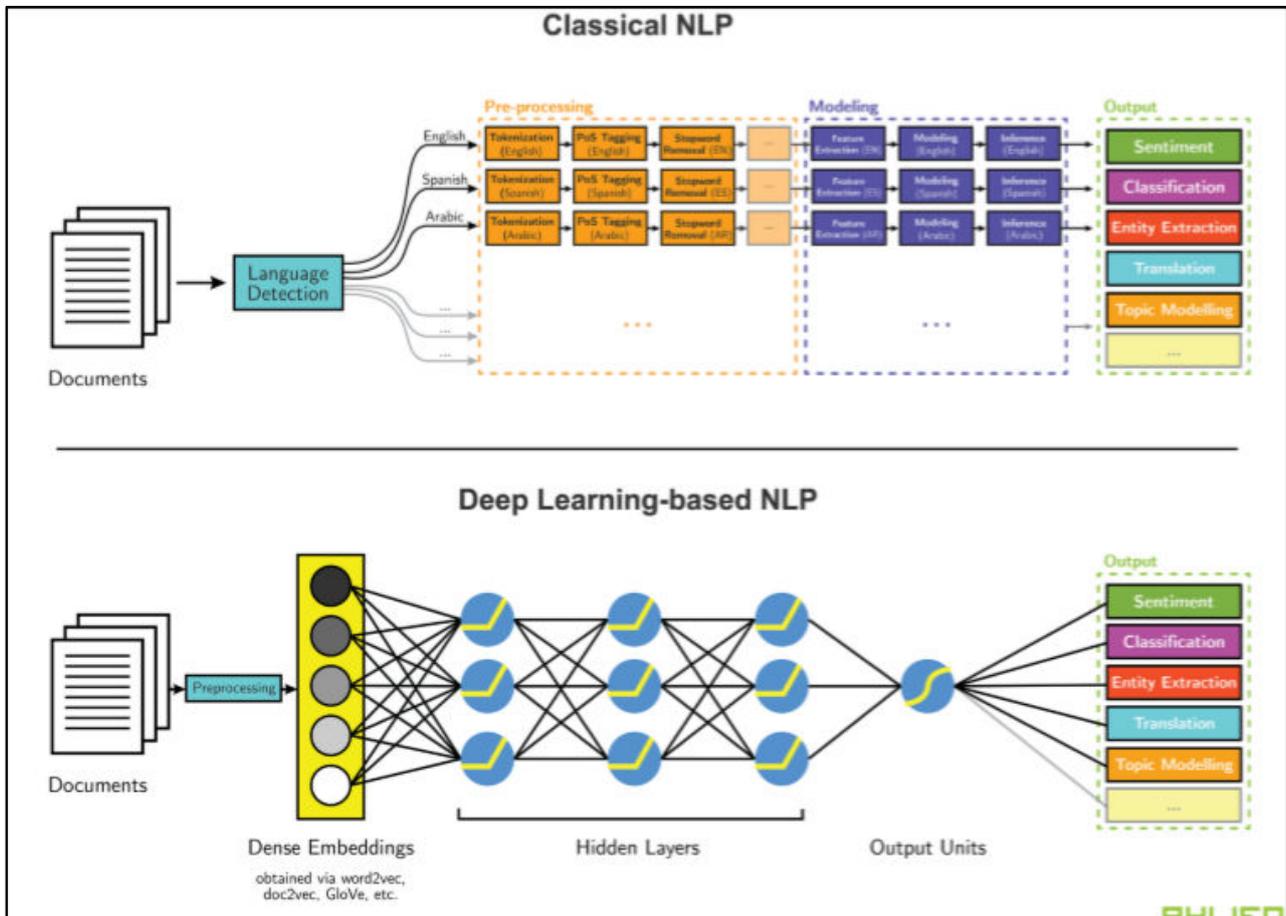


Figure 3.12: Stratégie d'apprentissage classique & d'apprentissage en profondeur (Thanaki, 2017).

### 3.10 Quelques travaux récents basés Deep Learning pour la fouille d'opinion basée aspects

Additionnant à ce que nous avons présenté précédemment dans le chapitre (II) et dans le but d'améliorer les performances des anciennes approches proposées dans le domaine de la fouille d'opinion et plus précisément dans au niveau aspect, et pour obtenir de nouveaux résultats à la pointe de la technologie, les chercheurs de la communauté de la fouille d'opinion ont contribué pour montrer des expériences et les analyses démontrent l'efficacité et la légèreté des modèles proposée. Dans cette section nous présentons quelques approches très récentes introduites par les chercheurs pour résoudre les problèmes de la fouille d'opinion basée aspect en se basant sur les réseaux de neurones approfondis.

### 3.10.1 Réseau de codeurs attentionnels pour la classification de sentiments ciblés

(Song, Wang, Jiang, Liu, & Rao, 2019) Ont proposé un réseau de codeur attentionnel (AEN) qui évite la récurrence et utilise des codeurs basés sur l'attention pour la modélisation entre contexte et cible. Ils ont soulevé le problème de manque de fiabilité des étiquettes et ils ont introduit une régularisation du lissage des étiquettes et ont obtenu de nouveaux résultats à la pointe de la technologie. Les expériences et les analyses démontrent l'efficacité et la légèreté de leur modèle.

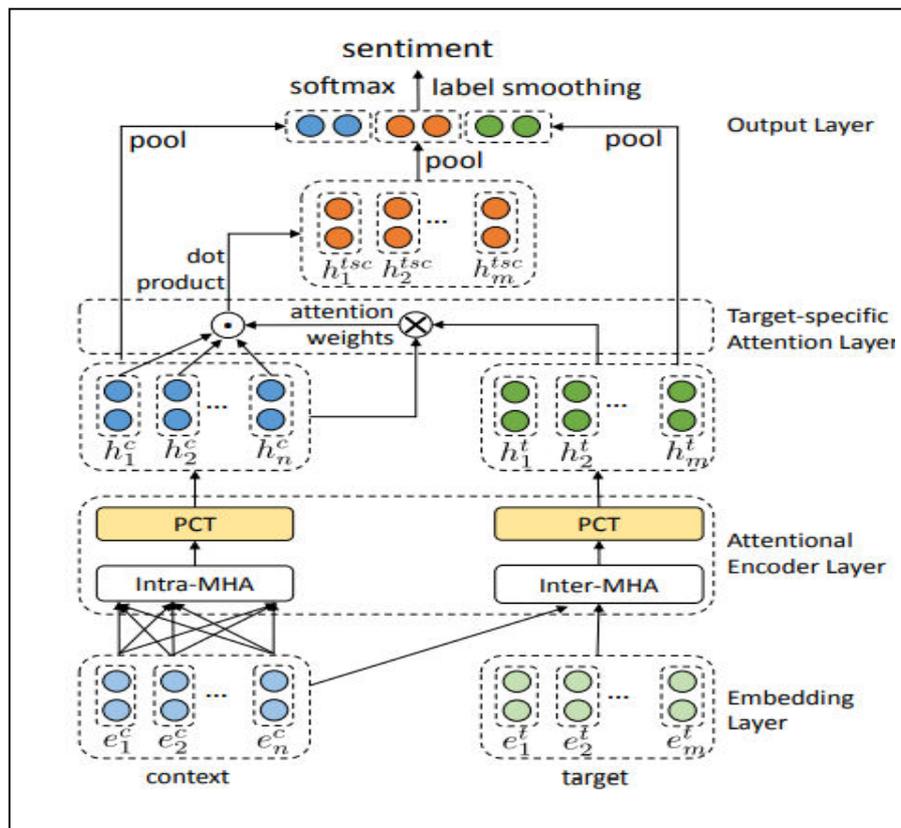


Figure 3.13 : Architecture globale de l'AEN (Song et al., 2019)

### 3.10.2 Classification des opinions avec des réseaux de neurone Attention-Sur-Attention

(B. Huang, Ou, & Carley, 2018) Ont introduit un réseau de neurones appelé Attention sur Attention (Attention-over-Attention (AOA)) pour la classification des sentiments au niveau de l'aspect. L'idée de leur approche modélise les aspects et les phrases de manière conjointe et capture explicitement l'interaction entre les aspects et les phrases contextuelles. Avec le module AOA, le modèle apprend conjointement les représentations des aspects et des phrases et se concentre automatiquement sur les parties importantes des phrases.

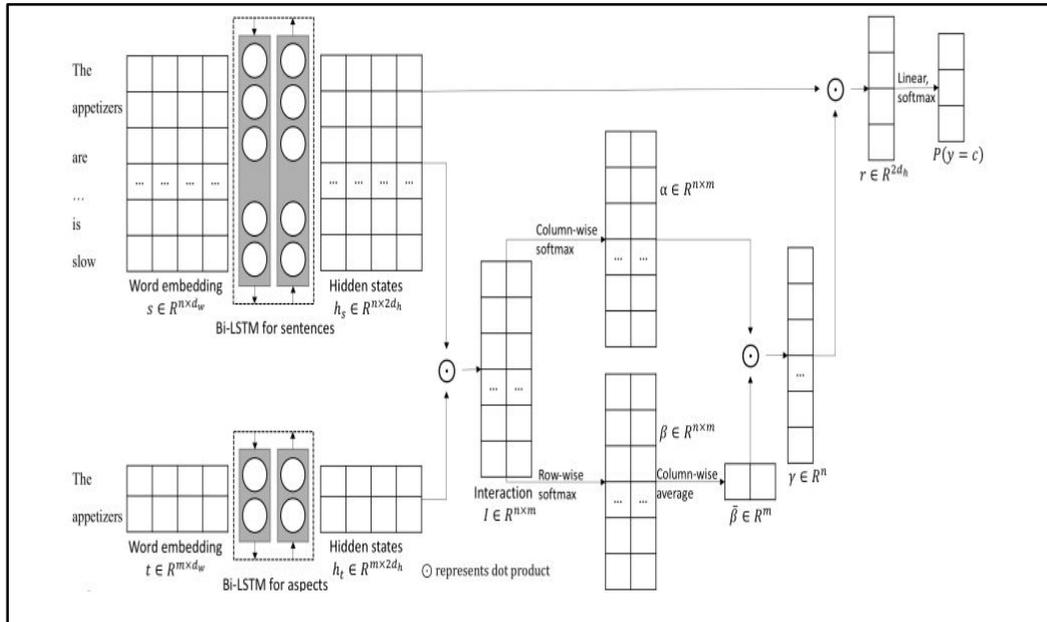


Figure 3.14 : Modèle de classification des sentiments par niveau d'aspect (B. Huang et al., 2018)

### 3.10.3 Réseau d'attention récurrent en mémoire pour l'analyse des sentiments

(Chen et al., 2017) ont adopté un mécanisme d'attention multiple pour capturer les caractéristiques de sentiment séparées par une longue distance, de manière à ce qu'il soit plus résistant contre les informations non pertinentes. Les résultats des attentions multiples sont combinés de manière non linéaire avec un réseau de neurones récurrent, ce qui renforce le pouvoir expressif de modèle pour traiter davantage de complications.

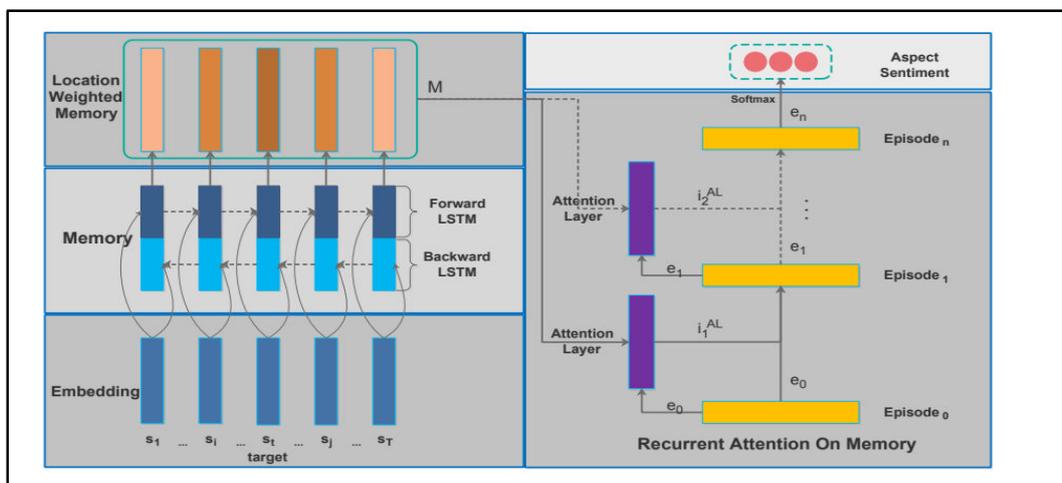


Figure 3.15: Architecture du modèle de (Chen, Sun, Bing, & Yang, 2017).

### 3.10.4 Classification de sentiment au niveau Aspect avec Deep Memory Network

(Tang et al., 2016) Ont proposé une approche pour capturer explicitement l'importance de chaque mot de contexte pour déduire la polarité de sentiment d'un aspect. Ce degré d'importance et cette représentation du texte sont calculés avec plusieurs couches de calcul, chacune d'entre elles étant un modèle d'attention neurale sur une mémoire externe. Des expériences sur des jeux de données d'ordinateurs portables et de restaurants montrent que leur approche est comparable à celle d'un système SVM basé sur des fonctionnalités de pointe et qu'elle est nettement supérieure à celle des architectures LSTM et des architectures LSTM basées sur l'attention. Sur les deux jeux de données

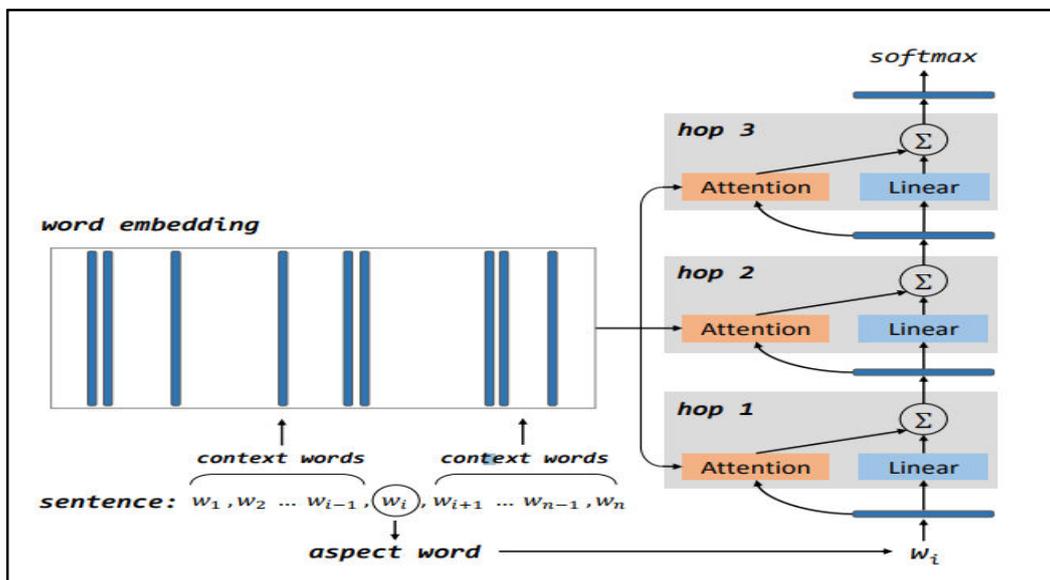


Figure 3.16: la classification des sentiments au niveau aspect (Tang, Qin, & Liu, 2016).

### 3.10.5 Double embeddings et étiquetage de séquence basée sur CNN pour l'extraction d'aspect

(Xu et al., 2018) Ont proposé une approche basée sur l'extraction d'aspect supervisé en utilisant l'apprentissage en profondeur. Contrairement à d'autres modèles d'apprentissage en profondeur supervisés très sophistiqués, ils ont proposé un modèle CNN innovant et simple, qui utilise deux types d'embeddings avec un pré-apprentissage pour l'extraction d'aspects : les incorporations à usage général et les intégrations spécifiques à un domaine. Sans recourir à une supervision supplémentaire, ce modèle obtient des résultats étonnamment bons, surpassant les méthodes existantes les plus sophistiquées.

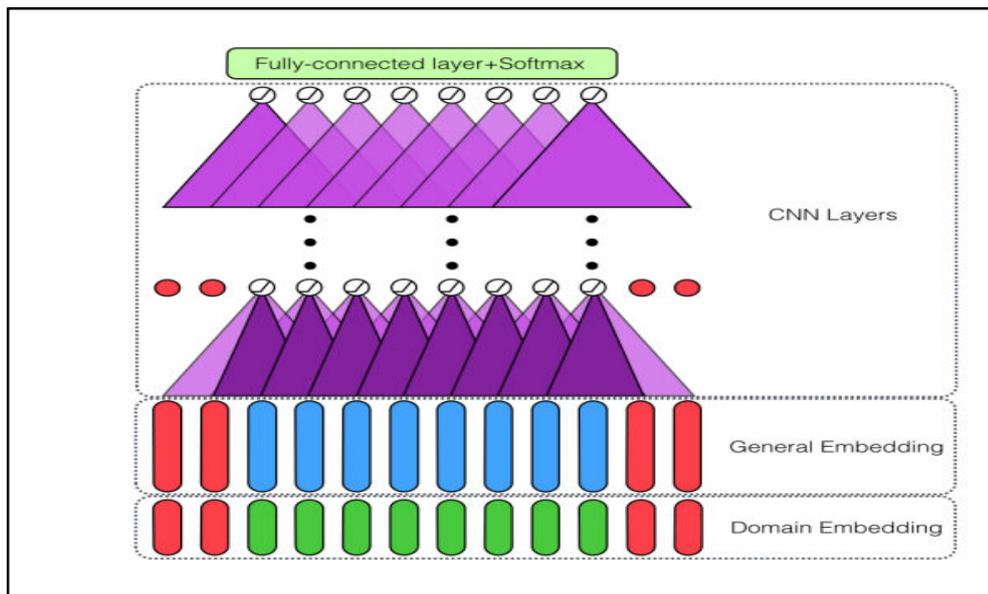


Figure 3.17: Modèle général pour la fouille d'opinion basée aspect (Xu, Liu, Shu, & Yu, 2018)

### 3.10.6 Analyse du sentiment basée sur les aspects avec des réseaux convolutionnels à porte

(Xue & Li, 2018) Ont proposé un modèle plus précis et plus efficace basé sur les réseaux de neurones convolutionnels et les mécanismes de déclenchement. Tout d'abord, les nouvelles unités Gated Tanh-ReLU peuvent générer en sortie les caractéristiques de sentiment en fonction de l'aspect ou de l'entité donnés. Ils ont construit une architecture beaucoup plus simple que la couche d'attention utilisée dans les modèles existants. Deuxièmement, les calculs de modèle pourraient être facilement parallélisés au cours de la formation, car les couches convolutives ne sont pas dépendantes du temps, contrairement aux couches LSTM, et les unités de déclenchement fonctionnent également de manière indépendante.

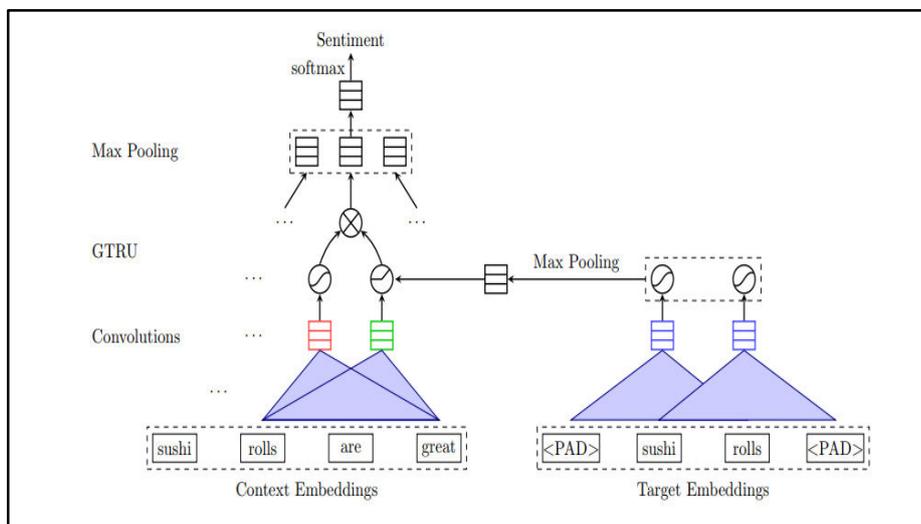


Figure 3.18: Illustration de modèle GCAE pour Tâche ACSA (Xue & Li, 2018).

Une paire de neurones de convolution calcule les caractéristiques d'une paire de porte Tanh et Porte ReLU. La porte ReLU reçoit le donné : informations d'aspect pour contrôler la propagation de caractéristiques de sentiment. Les sorties de deux portes son élément sagement multiplié pour le pooling maximum couche.

### **3.11 Conclusion**

Dans ce chapitre, nous passons brièvement en revue l'histoire des réseaux de neurones et de l'apprentissage en profondeur, puis nous nous concentrons sur les composants de base des modèles à base de réseaux de neurones : réseau neuronal convolutionnel (CNN), réseau neuronal récurrent (RNN), Le réseau de neurone de mémoire à long-court terme (LSTM), modèles de prolongement des mots, la stratégie d'apprentissage. Additionnellement à ce que nous avons présenté dans le chapitre précédent, nous nous limitons à quelques études récentes intéressantes présentées par les chercheurs de la communauté liées à l'extraction d'opinions basée sur les aspects avec des modèles d'apprentissage en profondeur. Ces chercheurs ont constaté que les modèles d'apprentissage en profondeur par rapport aux méthodes traditionnelles pourraient économiser beaucoup d'effort pour la tâche d'extraction des caractéristiques et peuvent fournir des résultats satisfaisants dans le domaine. Pour cela le chapitre suivant présentera en détail notre contribution qui consiste en un modèle basé deep learning et précisément sur les réseaux de neurones récurrents de type LSTM.

## 4 Chapitre IV : une approche basée Tree-LSTM pour la fouille d'opinion basée sur les aspects

### 4.1 Introduction

Dans la première partie de cette thèse, nous avons présenté le domaine de la fouille d'opinion ou l'analyse des sentiments. Nous nous sommes particulièrement intéressés à la fouille d'opinion au niveau fine de granularité c'est-à-dire au niveau d'aspect. La fouille d'opinions combine les techniques d'exploration de données et de traitement du langage naturel (TLN (NLP)) afin de détecter la subjectivité ou la polarité exprimée sur des aspects d'entité dans des textes porteurs d'opinion.

La fouille d'opinion basée aspect, reste toujours un domaine émergent qui a attiré et qu'il attire jusqu'à présent l'attention de beaucoup de chercheurs et nécessite plus d'effort par la communauté des chercheurs du domaine. Dans la littérature, un nombre important de travaux qui ont été réalisés par les chercheurs pour développer des modèles et des approches robustes faisant face aux différents défis du domaine. A ce propos, dans ce chapitre nous allons présenter nos contributions. La première contribution consiste à présenter une étude comparative et un état synthèse des différents travaux récemment présentés dans le chapitre (II). La deuxième contribution est une proposition d'une approche Deep learning basée sur les réseaux de neurone de type LSTM arborésant (Tree LSTM) pour la fouille d'opinion au niveau aspect.

### 4.2 Contribution 1 : Etude comparative et synthèse sur des approches existantes de la fouille d'opinion basée aspects

Vu le nombre important des travaux réalisés et publiés par les chercheurs de la communauté de la fouille d'opinion que ce soit basée sur des approches classiques ou sur des modèles d'apprentissage profondi Deep learning, comme première partie de notre travail , nous allons présenter une étude comparative et synthèse sur ces différentes approches proposées et qui sont déjà présentées dans le chapitre (II) sous forme d'un état de l'art sur la fouille d'opinion basée sur les aspects où une attention particulière sur les approches basées Deep Learning a été donnée à cause de l'importance des résultats fournis par ces approches en utilisant des réseaux CNN, LSTM et autres modèles.

Dans l'état de l'art qui a été présenté dans le chapitre (II), nous avons résumé plus de 33 techniques récentes de la période (2012-2018) présentées pour résoudre les différents problèmes et défis de la sous-tâche d'extraction d'aspect explicite dans le domaine de la fouille d'opinion liée aux trois catégories : supervisée, semi-supervisée et non supervisée.



En raison du grand nombre d'approches, de la diversité des jeux de données et de la diversité des modèles proposés, il serait pratiquement impossible de tracer une comparaison précise entre toutes les approches. Par conséquent, et afin de mieux comprendre l'efficacité de chaque approche, nous avons analysé et résumé séparément les différentes approches pour une catégorie bien précise. Bien qu'il soit injustifié de comparer les résultats des différents domaines, cela peut aider à comprendre la contribution de chaque approche. Nous avons comparé ces méthodes en se basant sur les trois métriques les plus connues à savoir la Précision, Recall et F-mesure comme le montrent les figures 4.1, 4.2, et 4.3.

Les chercheurs ont rapporté les expériences dans différentes langues, mais la plupart d'entre elles suivaient des ensembles de données en anglais. Par conséquent, nous avons marqué sur les graphes les approches utilisant des ensembles de données en langue chinoise par le signe "\*".

Les résultats expérimentaux ont été recueillis à partir des références (tableaux 2.3, 2.4, 2.5) du chapitre II, qui se sont présentés sous la forme de Précision, de Recall et de F-mesure. Dans certains cas, les auteurs n'ont pas fourni tous les résultats. Seuls les résultats fournis ont été mentionnés, par exemple : pour certaines approches, seule la précision est disponible et F-mesure pour certaines.

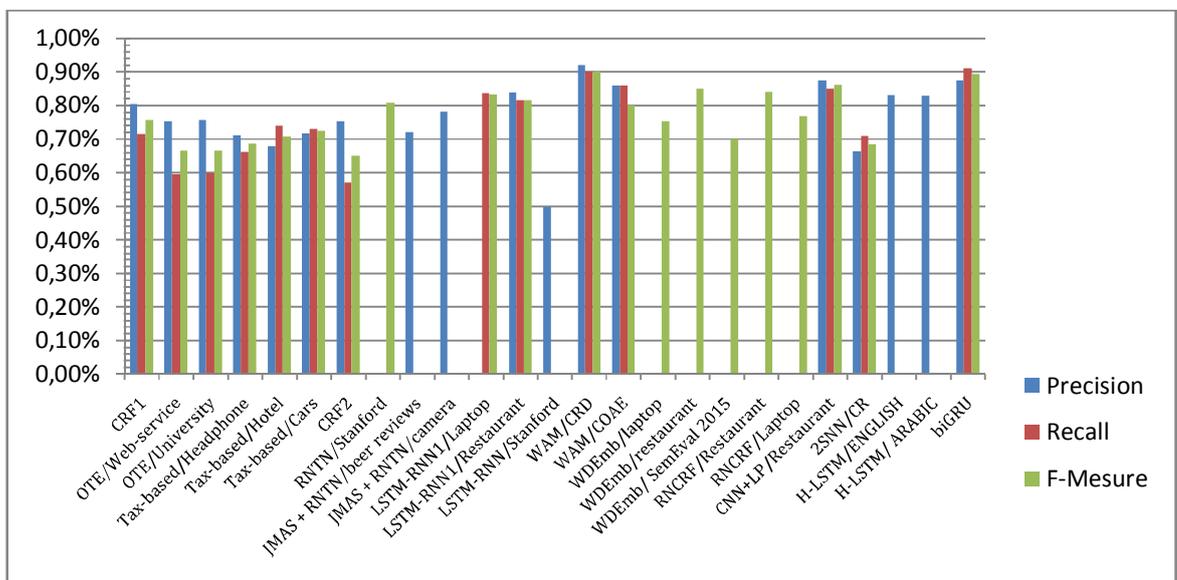


Figure 4.1 Approches supervisées pour la fouille d'opinion basée aspects (Bouras et al., 2019)

Il est remarquable à partir du tableau 2.3 et de la figure 4.1 que la plupart des études antérieures à 2013 étaient basées sur le champ aléatoire conditionnel (CRF, CRF1) (S. Huang et al., 2012) (Yang & Cardie, 2013). Le modèle CRF1 présente une valeur de précision considérable par rapport au modèle basé sur la taxonomie (Cruz et al., 2013) dans divers domaines. Le modèle LSTM-RNN (Le & Zuidema, 2015) sur le jeu de données de Stanford Treebank présente la précision la plus faible (inférieure à 50%) respectivement par rapport à WAM / CRD (K. Liu et al., 2015) et biGRU / tweets (Jabreel et al., 2018).

Avant l'année 2013, la plupart des chercheurs se sont concentrés sur l'apprentissage non supervisé et semi-supervisé, alors que peu de chercheurs ont exploré les approches supervisées. Cela est dû à la tâche difficile qui prend beaucoup de temps pour l'apprentissage de l'ensemble de données. D'autre part, les approches non supervisées ne nécessitent aucun ensemble d'apprentissage, tandis que les méthodes semi-supervisées ne nécessitent que quelques mots-clés pour démarrer le système.

Les approches supervisées nécessitent plus d'efforts pour l'étape d'apprentissage. Il est évidemment claire à partir du tableau 2.3 qu'elles ont été testées sur plusieurs jeux de données et qu'elles ont fourni des résultats plus significatifs que les techniques non supervisées et semi-supervisées où leurs graphes comparatifs seront représentés au fur et à mesure dans les sections suivantes.

Après l'année 2013, les chercheurs ont fourni plus d'efforts pour introduire dans leurs propositions des modèles basés sur les réseaux de neurones récurrents. Ces modèles ont présenté de bons résultats dans de nombreuses tâches de la PNL, tel que présenté par (Lakkaraju et al., 2014; Le & Zuidema, 2015; K. Liu et al., 2015; P. Liu et al., 2015; Socher et al., 2013; Wang et al., 2016). Les chercheurs ont introduit l'apprentissage en profondeur dans leurs contributions pour améliorer la recherche dans le domaine de la vision par ordinateur et de la reconnaissance de la parole. Pour cela, les puissants modèles basés sur les réseaux RNNS ont été proposés pour résoudre les problèmes de sous-tâches d'extraction d'aspect. Les chercheurs introduisent des présentations de mots prolongées (Word embeddings) et d'apprentissage en profondeur dans leurs études.

A partir du graphique de la figure 4.1, on peut établir les remarques principales suivantes :

- L'implémentation des algorithmes d'apprentissage en profondeur présente les taux les plus élevés en termes de jeux de données anglais. Les chercheurs sont donc, invités à implémenter ces algorithmes avec d'autres jeux de données liés à d'autres langues pour les tester.

- La comparaison donnée dans le tableau 2.3 et la figure 4.2 reste subjective car les résultats obtenus dépendent fortement de la langue et les jeux de données utilisés.

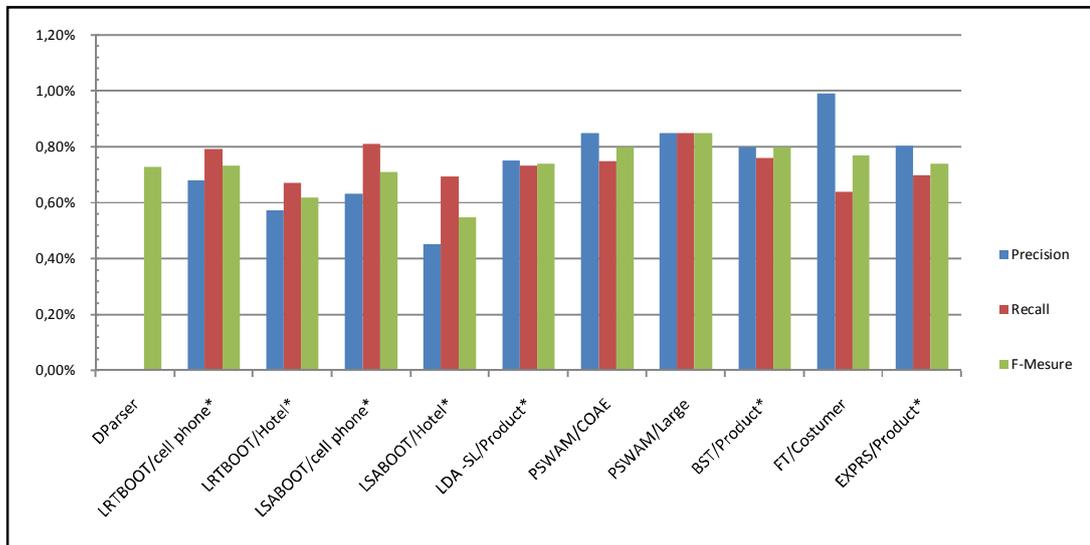


Figure 4.2 Approches semi supervisées pour la fouille d'opinion basée aspects (Bouras et al., 2019).

D'une part, et selon le tableau 2.4, les techniques semi-supervisées ont amélioré la précision de leurs systèmes en implémentant plusieurs types d'algorithmes. Le tableau présente 11 approches, sur lesquelles 7 ont été appliquées sur des jeux de données chinois (marquées par (\*) dans la figure 4.2).

D'une autre part, et selon la figure 4.2, l'approche FT (Samha et al., 2014) a la meilleure précision et une valeur moyenne pour les autres résultats expérimentaux par rapport à PSWAM (K. Liu et al., 2013) avec les meilleurs rappels et F-mesure sur le jeu de données anglais COAE. Les auteurs de DParser présentent leurs résultats uniquement avec F-Mesure. L'approche FT présente la meilleure précision par rapport à l'état de l'art présenté par rapport à la catégorie semi-supervisée.

La lecture de la figure 4.2, dégage les remarques suivantes :

- Des études sur la langue chinoise (mentionnées dans la figure 3 avec \*) sont présentes.
- La fouille d'opinion est une tâche qui dépend du domaine, cela signifie que le même algorithme renvoie des résultats différents avec des domaines et des langues différents. C'est le cas du réseau LRT-BOOT (Hai et al., 2012) et du réseau LSA-BOOT (Hai et al.,

2012), qui produisent tous les deux de meilleurs résultats sur les ensembles de données « CellPhone » ce qui n'est pas le cas sur les commentaires d'hôtels « Hotel ».

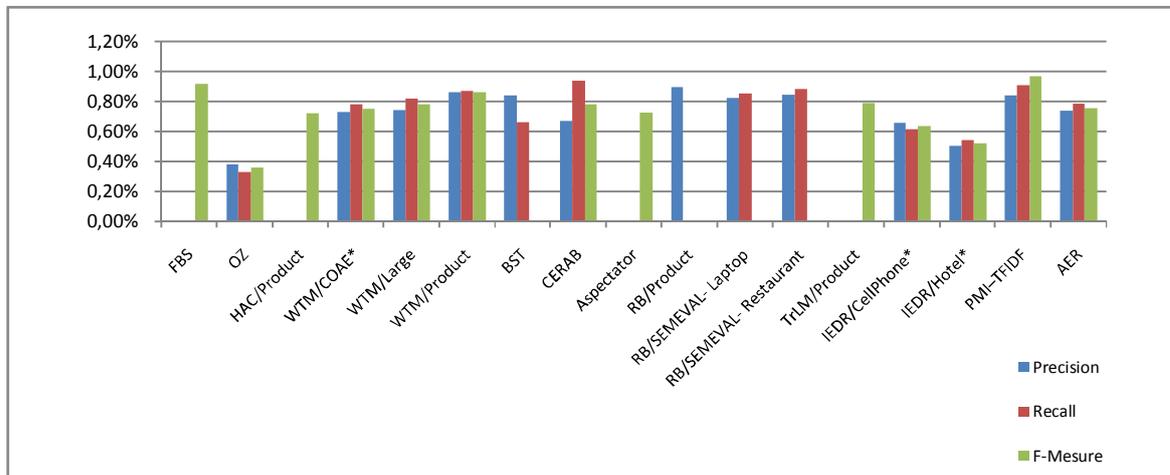


Figure 4.3 Approches non supervisées pour la fouille d'opinion basée aspects (Bouras et al., 2019).

La figure 4.3 présente les résultats de dix-sept (17) approches non supervisées dans différents domaines. Le modèle PMI-TFIDF (Quan & Ren, 2014) présente les meilleurs résultats expérimentaux.

Le modèle FBS (Bafna & Toshniwal, 2013) a fourni une F-mesure remarquable, mais l'étude n'a fourni aucun calcul de Recall ni de précision ce qui n'est pas justifié dans leur papier. Le modèle RB (Poria et al., 2014) a fourni la haute Recall avec plusieurs jeux de données anglais après le modèle CERAB (Y. Li et al., 2015; Y. Li et al., 2014).

La figure 4.3 montre clairement que presque toutes les approches ont produit des résultats intéressants dans des domaines différents. Cependant, Opinion Zoom (OZ) (Marrese-Taylor et al., 2014; Marrese-Taylor, Velásquez, Bravo-Marquez, et al., 2013) a signalé une très faible précision par rapport à toutes les autres approches. Ceci est dû aux domaines du tourisme, où les gens aiment raconter des histoires sur leurs expériences. Dans de tels cas, le nombre de phrases objectives est beaucoup plus élevé que les phrases subjectives et donne donc une performance médiocre. Il est donc recommandé de l'appliquer à d'autres domaines plutôt qu'au tourisme.

En outre, le modèle IEDR (Hai et al., 2013) n'a pas produit de bons résultats pour des critiques des hôtels en langue chinoise, où il a produit des résultats comparables dans le domaine des produits.

### 4.2.1 Synthèse

Après l'analyse et la comparaison des différentes méthodes et modèles que nous avons présentés dans les sections précédentes où les chercheurs de la communauté de la fouille d'opinion ont présenté des approches dans le but de contribuer et faire face aux défis et problèmes rencontrés dans le domaine de la fouille d'opinion, dans ce qui suit nous synthétisons les points majeurs et importants obtenus à partir de cette étude comparative, que nous avons effectué et qui nécessitent plus d'efforts par les chercheurs.

- Les approches comparées précédemment ont utilisé différents jeux de données, qui appartiennent à différents types de domaines et de différentes langues. La plupart des travaux portaient sur des ensembles de données en anglais, et quelques chercheurs ont mené des expériences sur des domaines appartenant à la langue chinoise.
- Les résultats d'un domaine ne peuvent pas être comparés aux résultats d'un autre domaine et cela est similaire pour le langage. Par conséquent, il est nécessaire d'identifier les domaines et les langues de toutes les approches pour effectuer une comparaison complète et justifiable.
- Malgré tous les efforts déployés, par les chercheurs dans le domaine de la fouille d'opinion, et précisément dans les sous tâches de la fouille d'opinion basée sur les aspects, ce domaine est toujours considéré comme un domaine de recherche de premier plan qui nécessite plus de recherche. Pour cela, les chercheurs devraient fournir plus efforts pour relever un certain nombre de défis. Parmi ces derniers nous citons : les corpus annotés, les erreurs cumulatives de prétraitement, la dépendance du domaine et l'apprentissage en profondeur qui n'est pas encore montré ces points forts.
- L'insuffisance des corpus annotés a incité les chercheurs de la communauté à faire appel aux approches semi-supervisées et non supervisées pour faire face au problème des données non annotées du domaine. Malgré tout, l'utilisation de corpus annotés reste un besoin indispensable pour estimer un taux de satisfaction lors du processus de la fouille d'opinion.
- Il est à noter que la plupart des corpus annotés sont disponibles en langue anglaise, ce qui pose un problème pour le processus de la fouille d'opinion dans d'autres langues. De plus, les méthodes basées sur l'apprentissage par transfert et la traduction automatique ont leurs inconvénients respectifs dans la fouille d'opinion multilingue. Donc, des

efforts supplémentaires en termes de corpus annotés sont attendus par la communauté de la fouille d'opinion.

- Le processus de la fouille d'opinion est basé sur des sous-tâches telles que la segmentation par mots, le marquage POS<sup>23</sup> (parties de discours). Les utilisateurs de ces techniques supposent que les textes d'opinion sont écrits selon les règles grammaticales et sémantiques du langage. Mais, en réalité, les opinions des utilisateurs sont exprimées sous formes qui ne respectent pas littéralement les règles de la langue. Pour prendre en compte le problème des documents de forme libre, des approches plus précises sont nécessaires.
- Il ressort clairement, à partir de notre étude, qu'une méthode qui peut bien fonctionner dans un domaine, rien ne garantit des performances similaires dans d'autres domaines et les paramètres d'apprentissage dans un domaine ne peuvent pas être appliqués à un autre domaine. Ainsi, les performances des modèles varient de manière significative entre les domaines et dépendent principalement des paramètres d'apprentissage. Donc, les approches d'adaptation et de grands corpus similaires sont nécessaires pour surmonter le problème de la dépendance de domaine.
- Les modèles d'apprentissage en profondeur qui constituent une branche de l'apprentissage automatique basé sur des réseaux de neurones, récemment utilisés dans de nombreuses tâches de traitement de la langue naturelle ont présenté les meilleurs résultats, sans ingénierie manuelle des caractéristiques et avec le moins de restrictions possibles, mais avec une quantité de données et une durée d'apprentissage suffisante.
- D'après notre comparaison nous avons constaté que la majorité des travaux de recherche récente dans le domaine de la fouille d'opinion, particulièrement au niveau aspect sont orientés vers les approches basées sur l'apprentissage en profondeur.

---

<sup>23</sup> POS : Part of Speech en anglais, partie du discours: En grammaire, une partie du discours (également une classe de mots, une classe lexicale ou une catégorie lexicale) est une catégorie linguistique de mots (ou plus précisément "éléments lexicaux"), qui est généralement définie par le comportement morphologique de l'élément lexical en question.

### **4.3 Contribution 2 : Une approche basée Tree-LSTM pour la fouille d'opinion basée aspect**

L'objectif principal de cette thèse est de proposer une nouvelle approche, pour la fouille d'opinion utilisant les différents aspects pour chaque opinion. La plupart des anciennes approches de la fouille d'opinion sont basées sur les algorithmes d'apprentissage par machine classique tel que SVM (support vector machine) et des classificateurs de champ aléatoires conditionnels (CRF). Bien que ces méthodes permettent une amélioration remarquable des performances des systèmes de la fouille d'opinion, leurs résultats dépendent énormément de la qualité des caractéristiques extraite manuellement et ne peuvent pas être extraite automatiquement, ce qui rend impossible l'obtention de résultats significatifs sur une grande quantité de données. Parmi les points essentiels que nous avons dégagé de notre première contribution et bien l'utilisation des modèles basés sur l'apprentissage en profondeur dans de nombreuses tâches de traitement de la langue naturelle qui présentent les meilleurs résultats, sans ingénierie manuelle des caractéristiques et avec le moins de restrictions possibles, mais avec une quantité de données et une durée d'apprentissage suffisante. Pour cela, et comme deuxième contribution, nous allons développer en détail un modèle Deep Learning basé sur les réseaux de neurone de type LSTM arborescent ( Tree-LSTM) pour l'extraction et la classification des opinions au niveau aspect. Les sections suivantes présentent tous les aspects et facettes de l'approche proposée.

#### **4.3.1 Présentation de l'approche proposée**

Dans cette section, nous détaillons notre approche. Nous avons, tout d'abord, mis en lumière deux problèmes majeurs liés au mécanisme des réseaux de neurones récurrents (RNN), le premier étant: les problèmes de gradient en voie de disparition ( vanishing gradient problem) et le second: le problème d'incompatibilité de données entre le réseau de neurones standard LSTM et le Tree-LSTM. Ensuite, nous présentons notre approche basée tree-LSTM pour les tâches d'extraction ou fouille d'opinion basées sur les aspects.

##### **4.3.1.1 Le problème de gradient (vanishing gradient problem)**

Les modèles fondés sur les réseaux neuronaux d'apprentissage en profondeur, tels que les réseaux de neurones récurrents (RNN), sont spécialement conçus pour l'apprentissage de séquences de données de longueur arbitraire, car ils analysent séquentiellement toute la séquence et en génèrent une forme comprimée. Le processus d'apprentissage des RNN se fait au niveau des nœuds cachés en fonction de leurs couches de nœuds précédentes. En théorie, les modèles RNN sont compétents pour se

souvenir des dépendances à longue distance. Mais, à mesure que la séquence s'allonge, ces modèles souffrent du problème du gradient disparaissant, ce qui entrave l'apprentissage de longues séquences de données. Les gradients portent les informations utilisées dans la mise à jour des paramètres RNN. Lorsque le gradient devient de plus en plus petit, les mises à jour des paramètres deviennent insignifiantes, ce qui signifie qu'aucun véritable apprentissage n'est effectué.

Pour surmonter cet inconvénient, nous allons présenter un modèle Deep learning basé sur les réseaux de neurone de type LSTMs, qui ont été suggérés avec un mécanisme de déclenchement en incluant une cellule mémoire constituée d'une unité avec des unités de portes d'entrée et de sortie multiplicatives. Des portes d'entrée contre des entrées non pertinentes et des portes de sortie contre des contenus en mémoire non pertinents actuels pour propager de nouvelles informations et oublier certaines informations précédentes permettant aux gradients de se propager davantage. Ces modèles ont été suggérés comme une solution au problème précité et ont démontré leur efficacité dans de nombreux problèmes liés au traitement du langage naturel.

#### **4.3.1.2 Le problème incompatibilité de données entre LSTM et le Tree-LSTM**

Bien que nous avons résolu le problème du gradient de disparition de RNN par le mécanisme de déclenchement du LSTM, il reste limité pour le traitement de données linéaire, car un LSTM typique traite les informations de manière séquentielle alors que les architectures Tree-LSTM permettent la représentation de phrases par le biais d'une structure syntaxique.

En outre, une phrase en langage naturel, code des informations supplémentaires qui sont plus qu'une séquence de mots et régulièrement représentées dans des données de structure arborescente. Ce dernier soulève le problème de l'incompatibilité des données entre le LSTM standard et le LSTM arborescent ( Tree-LSTM).

Après avoir traité les deux problèmes que se soit lié au type de réseau ou à la structure de données en entrée, à ce niveau-là, nous allons présenté le schéma général de notre approche.

#### **4.3.1.3 Schéma général de notre approche**

Une phrase peut être présentée sous deux arborescences possibles: l'arbre de dépendance et l'arbre de circonscription. Comme nous l'avons présenté dans la section précédente, les données structurées en arborescence ne peuvent pas être analysées correctement à l'aide du LSTM standard.

A cet effet, pour traiter ces données structurées, dans notre approche, nous proposons un modèle d'arborescence de circonscription-LSTM pour l'exploration d'opinion basée sur les aspects.



Pour cela nous avons fait appel aux modèles proposés par Tai et al (Tai et al., 2015), Les modèles de référence sont de type LSTM. Le premier est la somme des fils Tree-LSTM « somme childs-LSTM » et le second est le « N-array Tree-LSTM », comme les montres dans la figure (4.4). Dans notre approche, nous adoptons le modèle de réseau neuronal de type «N-array Tree-LSTM», qui convient au traitement des arbres de circonscription car il gère des arbres avec une ramification binaire, en particulier avec l'arbre de circonscription ( $b = 2$ ). Notre approche analyse les données structurées en arborescence en préservant chaque propriété du mécanisme de synchronisation LSTM standard pour calculer deux sous-tâches de la fouille d'opinion basée sur les aspects, l'extraction d'aspect et la détection de polarité. Nous utilisons le réseau tree-LSTM pour obtenir des états cachés de mots au niveau des mots utilisés pour un aspect et sa polarité, respectivement. Dans la sous-section ci-dessous, nous détaillons le modèle de circonscription Tree-LSTM, ainsi que ses différentes couches, depuis la couche d'entrée jusqu'à la couche entièrement connectée, comme le montre la figure (4.5).

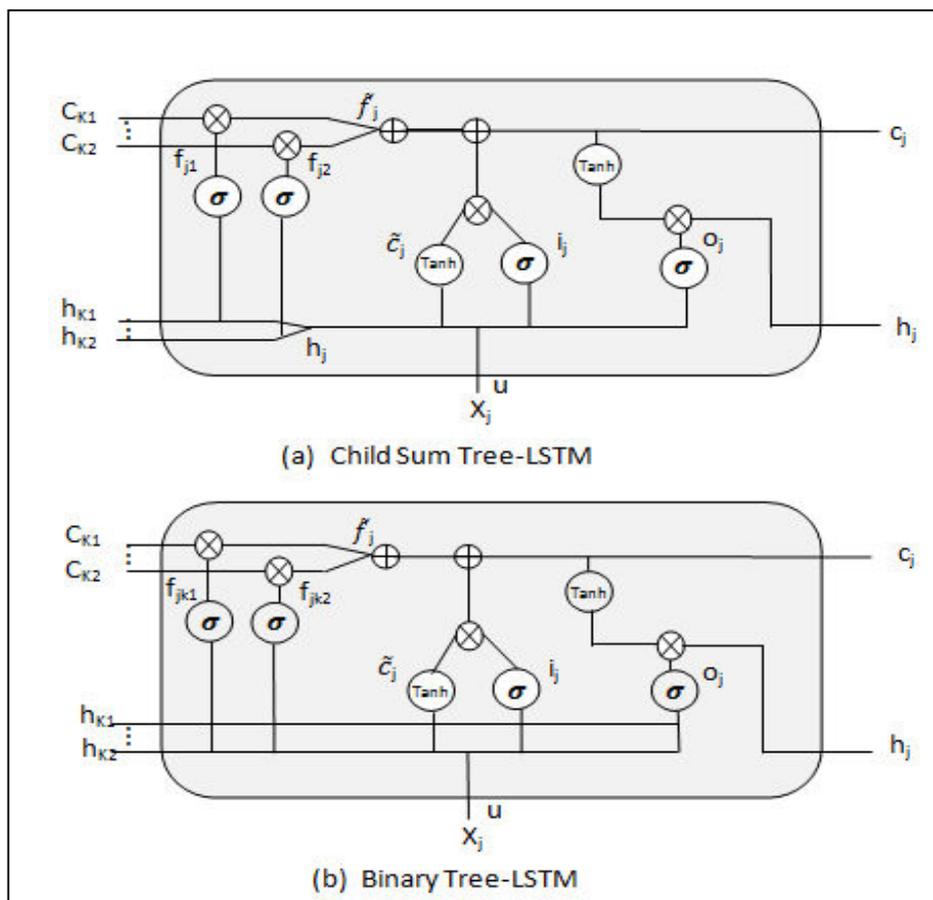


Figure 4.4: Les différents types du réseau Tree-LSTM (Tai et al., 2015).

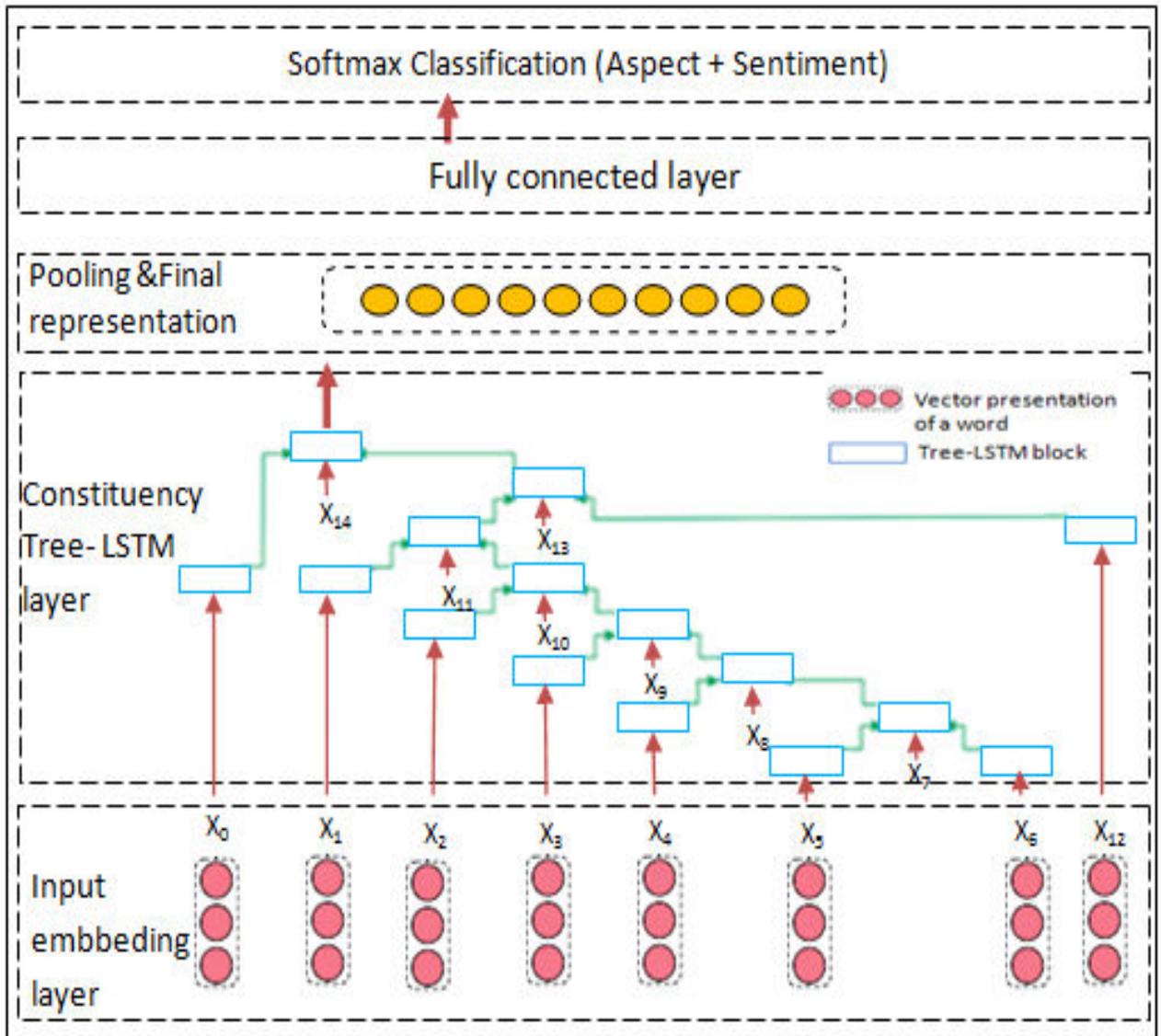


Figure 4.5: L'architecture de l'approche proposée (Bouras, Amroune, Bendjenna, & Bendib, 2020).

**a) La couche d'entrée (Input embedding layer)**

Pour décrire notre modèle, nous commençons par la présentation formelle du problème. Nous supposons que le contexte contient n-mot  $\{x_c^1, x_c^2, \dots, x_c^n\}$  et que l'aspect est dans m-mot  $\{x_a^1, x_a^2, \dots, x_a^m\}$ , x désignant un mot spécifique pour le contexte et l'aspect. Nous présentons chaque mot par son vecteur de valeur réelle, appelé présentation par incorporation de mots (vecteurs continus). Ce dernier correspond aux paramètres du réseau de neurones, pré-formés à partir d'un corpus approprié via un modèle de langage, pour être adaptés à notre arborescence du réseau LSTM. Ainsi, dans cette

étape, nous pouvons obtenir  $X_t \in \mathbb{R}^d$  de  $M^{v \times d}$ , où  $t$  est l'indice du mot,  $v$  représente la taille du vocabulaire et  $d$  présente la dimension d'incorporation.

**b) La couche LSTM avec arbre de circonscription (The constituency Tree-LSTM layer)**

Les mots d'une phrase sont fortement dépendants les uns des autres et que Ttree-LSTM apprend très bien les dépendances à long terme et peut lutter contre la disparition progressive des données structurées. Nous avons utilisé les réseaux Tree-LSTM pour apprendre la sémantique des mots cachés. Plus formellement, notre modèle Tree-LSTM traite l'arborescence de la circonscription associée à une phrase en entrée de manière ascendante (c'est-à-dire des feuilles au nœud racine). Ceci est effectué en traitant de manière récursive les nœuds de l'arborescence, en utilisant leurs représentations enfants en tant qu'entrée avec  $n = 2$  pour chaque nœud. La fonction de transition pour un nœud «  $t$  » et un ensemble d'enfants  $C(t)$  (gauche et droite) est donnée par l'ensemble des équations suivantes:

$$h_t = O_t \odot \mathbf{tanh}(C_t) \dots \dots \dots (32)$$

$$O_t = \sigma(W_o X_t + U_o^L h_{t-1}^L + U_o^R h_{t-1}^R + b_o) \dots \dots \dots (33)$$

$$C_t = f_t^L \odot C_{t-1}^L + F_t^R \odot C_{t-1}^R + i_t \odot u_t + b_c \dots (34)$$

$$f_t^L = \sigma(W_f X_t + u_f^{LL} h_{t-1}^L + u_f^{LR} h_{t-1}^R + b_f) \dots \dots \dots (35)$$

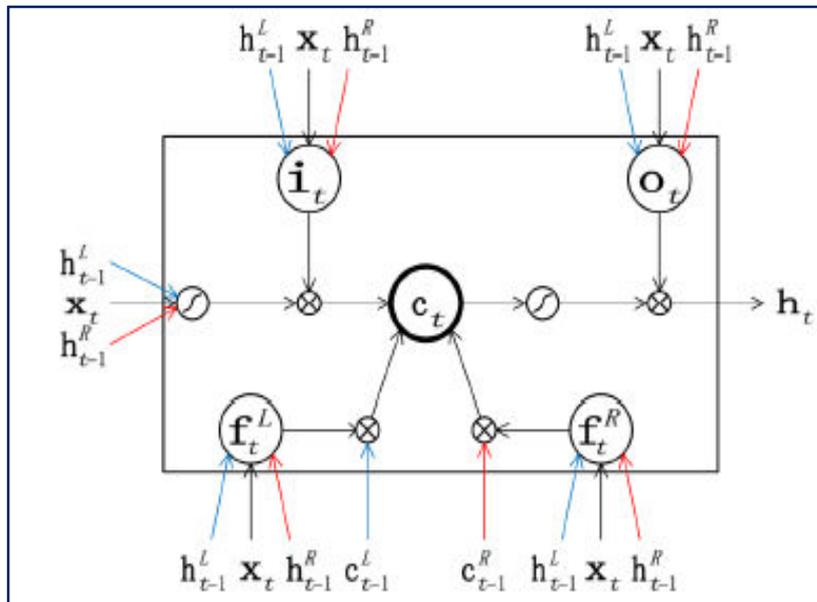
$$f_t^R = \sigma(W_f X_t + u_f^{RL} h_{t-1}^L + u_f^{RR} h_{t-1}^R + b_f) \dots \dots \dots (36)$$

$$i_t = \sigma(W_i X_t + u_i^L h_{t-1}^L + u_i^R h_{t-1}^R + b_i) \dots \dots \dots (37)$$

$$u_t = \mathbf{tanh}(W_c X_t + u_c^L h_{t-1}^L + u_c^R h_{t-1}^R) \dots \dots \dots (38)$$

Où  $\odot$  est la multiplication élément par élément,  $\sigma$  signifie une fonction sigmoïde (fonction logistique),  $X_t$  est l'entrée du noeud  $t$ ,  $h_t$  est l'état caché du  $j^{ième}$  fils. Chaque unité Tree-LSTM de circonscription de notre modèle est une collection de vecteurs: une porte d'entrée  $i_t$ , la porte d'oublie  $f_t^{LR}$  avec  $f_t^L$  pour le fils gauche et  $f_t^R$  pour fils droit, une porte de sortie  $o_t$ , une cellule mémoire  $C_t$  et un état caché  $h_t$ .  $b$  est le vecteur de biais,  $W$  et  $U$  sont les matrices de pondération pour l'étape de transformation. En général, à chaque noeud, un vecteur d'entrée  $X_t$  et les vecteurs cachés de deux fils (le fils gauche  $h_{t-1}^L$  et le droit  $h_{t-1}^R$ ) sont pris en tant qu'input pour calculer le vecteur caché  $h_{t-1}$  du nœud parent. Ces informations sont également utilisées pour configurer les quatre portes, c'est-à-dire la porte d'entrée  $i_t$ , la porte de sortie  $o_t$ , ainsi que les deux portes d'oublies  $f_t^L$  et  $f_t^R$ . La cellule de mémoire

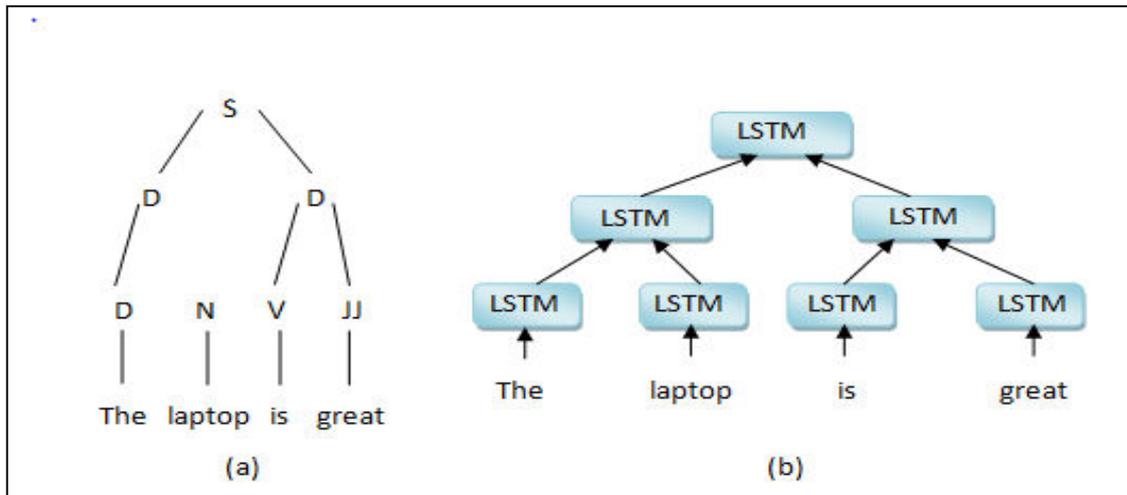
$c_t$  considère les vecteurs cellulaires de chaque enfant,  $c_{t-1}^L$  et  $c_{t-1}^R$ , qui sont déclenchés respectivement par les portillons oubliés gauche et droit, comme le montre la figure (4.6).



**Figure 4.6:** Le bloc de mémoire à chaque nœud du réseau de type ‘circonscription arborescente-LSTM (Bouras et al., 2020).

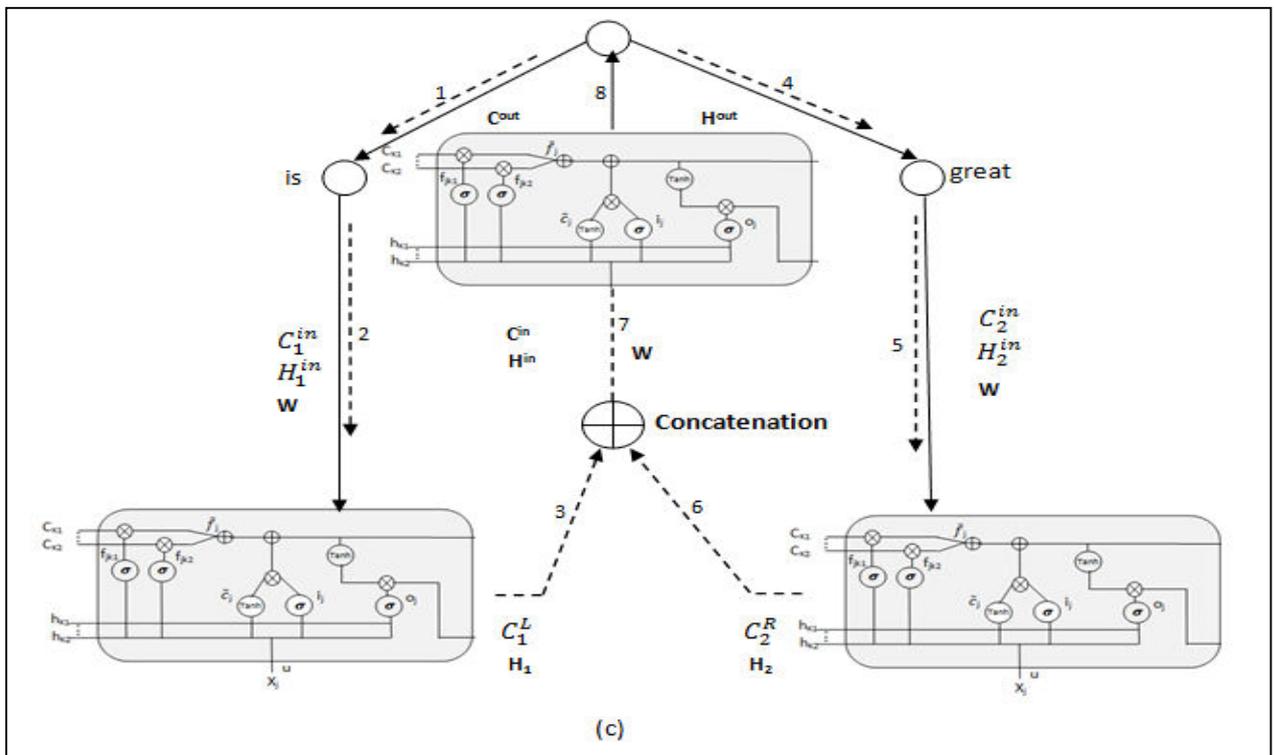
Tout d'abord, la structure arborescente binaire de chaque phrase de notre proposition est fournie à l'aide de l'analyseur de PNL de Stanford. Où les annotations d'aspect-polarité sont codées avec une chaîne de 18 caractères attachée au nœud racine de chaque phrase. La position dans la chaîne de 18 caractères est liée à la présence ou à l'absence de l'aspect et le caractère «1, 2, 3» présentes respectivement un sentiment positif, neutre ou négatif. Cependant, le codage «0» signifie que l'aspect est manquant (pas d'aspect dans la phrase). Ensuite, le modèle Tree-LSTM traite l'arborescence de la circonscription ( $n = 2$ ) associée à une phrase d'entrée des feuilles au nœud racine que tous les autres nœuds enfants, comme expliqué dans les figures (4.7) et (4.8).

Enfin, nous obtenons les états cachés en tant que représentations de mot finales de la phrase avec son codage en chaîne de 18 caractères pour les alimenter vers le calque entièrement connecté avec une fonction d'activation « softmax » pour la classification d'opinion au niveau aspect. L'étiquette avec la probabilité la plus élevée est définie comme résultat final.



**Figure 4.7:** (a) un exemple d'arbre de circonscription, (b) le réseau tree-LSTM correspondant (Bouras et al., 2020).

Dans la figure (a), nous utilisons les abréviations suivantes: S pour phrase, N pour nom, V pour verbe, D pour déterminant, JJ pour adjectif.



**Figure 4.8:** Le comportement de notre modèle Tree-LSTM sur une partie de l'arbre de circonscription de la phrase «The laptop is great» (Bouras et al., 2020).

### c) La couche de sortie (The fully-connected layer (output))

Comme nous l'avons mentionné précédemment, la Tree-LSTM traite l'arbre de circonscription de manière récurrente de ses feuilles à ses nœuds parents afin de fournir une représentation pour le nœud racine, c'est-à-dire le nœud qui synthétise tous les autres nœuds. Cette représentation prend en compte la phrase entière. Le résultat de la couche « circonscription tree-LSTM » est entré dans une couche entièrement connectée avec une fonction d'activation softmax et génère l'aspect sur une position de 18 caractères et une distribution de polarité sur trois étiquettes de polarité: «positif», «négatif» et «neutre». L'indice avec la probabilité estimée la plus élevée devient l'étiquette de polarité prédite pour l'aspect donné.

## 4.4 Expérimentation

Nous menons des expériences sur deux sous-tâches de la fouille d'opinion: la sous-tâche d'extraction des d'aspect et la sous-tâche de détection de polarité. Dans notre travail, ces deux types de tâches sont pris en compte et ils sont collectivement nommés en tant que la fouille d'opinion basée sur les aspects (ABOM).

### 4.4.1 Jeux de données ( datasets)

Pour vérifier la validité du modèle proposé, nous avons effectué des expériences sur les jeux de données SemEval 2014 Task 4 (Pontiki et al., 2014), qui constituent des références largement utilisées dans de nombreux travaux de la littérature, l'un appartenant au domaine des ordinateurs portables et l'autre au domaine des restaurants. Dans notre proposition, nous faisons référence à l'ensemble de données d'un ordinateur portable. Le tableau (4.1) indique le nombre d'instances d'apprentissage et de test dans cet ensemble de données.

Jeux de données	Positive.(#)	Neutre.(#)	Negative.(#)
Laptop- apprentissage	994	464	870
Laptop- Test	341	169	128

**Tableau 4.1:** Statistique SemEval 2014 du jeu de données des ordinateurs portable.

```

</sentence>
<sentence id="1316">
  <text>The tech guy then said the service center does not do 1-to-
  1 exchange and I have to direct my concern to the "sales" team,
  which is the retail shop which I bought my netbook from.</text>
  <aspectTerms>
    <aspectTerm term="service center" polarity="negative" from="27"
    to="41" />
    <aspectTerm term=""sales" team" polarity="negative" from="109"
    to="121" />
    <aspectTerm term="tech guy" polarity="neutral" from="4" to="12" />
  </aspectTerms>
</sentence>

```

**Figure 4.9** Exemple de jeu de données SemEval 2014 Task 4 sur les ordinateurs portables (Pontiki et al., 2014).

#### 4.4.2 Environnement et paramètres d'exécution

Nous implémentons les modèles dans Tensorflow. Nous avons utilisé 300 nouveaux mots imbriqués dans Word2vecs pré-formés et des mots hors vocabulaire en échantillonnant dans la distribution uniforme  $U(-0,5, 0,5)$ . Les valeurs initiales de toutes les matrices de poids sont échantillonnées à partir de la distribution uniforme  $U(-0,1, 0,1)$ . Toutes les dimensions d'intégration sont définies sur 300 et la taille du lot sur 32. Nous minimisons la fonction de perte pour entraîner nos modèles à l'aide d'AdaGradOptimizer (Duchi, Hazan, & Singer, 2011) avec un taux d'apprentissage défini à 0,05.

Pour éviter les sur-apprentissages (overfitting) de notre réseau, nous adoptons la stratégie dropout (Semeniuta, Severyn, & Barth, 2016; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) après la couche d'intégration avec  $p = 0,6$  et le coefficient de normalisation L2 dans la fonction de perte est défini sur 0,01.

#### 4.4.3 Mesures d'évaluation

En général, L'évaluation comprend principalement la sélection et le traitement des données, l'ajustement du modèle de prévision et la comparaison du modèle avec d'autres modèles implémentés sur les mêmes jeux de données.

Premièrement le jeu de donnée a été divisé en deux parties indissociables, un ensemble de d'apprentissage étant utilisé pour la formation de modèle, et un ensemble de test utilisé pour tester l'effet d'apprentissage du modèle.

Les performances de notre modèle proposé sont évaluées par les métriques : la *Précision* (P), le *Recall* (R), le *F-mesure* (F) pour la sous-tâche d'extraction d'aspect et la métrique d' *Accuracy* (Acc) pour la sous-tâche de détection de polarité; ces mesures sont définies comme suit:

$$\text{Precision} = \frac{TP}{TP + FP} \dots \dots \dots (39)$$

$$\text{Recall} = \frac{TP}{TP + FN} \dots \dots \dots (40)$$

$$F(\text{ score}) = \frac{2 * P * R}{P + R} \dots \dots \dots (41)$$

$$ACC = \frac{TP}{TP + TN + FP + FN} \dots \dots \dots (42)$$

Où TP (Vrai Positif), TN (Vrai Négatif), FP (Faux Positif) et FN (Faux Négatif) sont les cas correctement classés ou mal classés, et F-score est la moyenne des métriques de précision et de rappel.

La précision mesure le pourcentage d'échantillons correctement prédits dans tous les échantillons. Généralement, un système bien exécuté a une plus grande précision.

**4.5 Discussion**

**4.5.1 Comparaison des modèles**

Dans cette section, nous allons comparer notre modèle avec d'autres modèles sur le même jeu de donnée pour tester la performance de notre modèle de manière exhaustive. Nous utilisons les jeux de données de la tâche 4 de SemEval-2014 qui contient deux jeux de données spécifiques au domaine des ordinateurs portables.

**4.5.1.1 La sous tâche d'extraction d'aspect**

Pour évaluer notre modèle Tree-LSTM proposé pour la sous-tâche1: extraction d'aspect, nous avons effectué une étude de comparaison avec les deux groupes de modèles suivants:

**a- Modèles basés sur le CRF**



- **CRF**: champs aléatoires conditionnels avec présentation de base des fonctionnalités et des incorporations de mots;
- **Semi-CRF**: champs aléatoires conditionnels de semi-Markov et modèle d'entité;
- **WDEmb**: CRF avec incorporations de mots, incorporations de chemins de dépendance et incorporations de contexte linéaires;
- **DLIREC**: l'un des systèmes les mieux classés sur la base de CRF dans la sous-tâche ATE dans les défis SemEval ABSA ;
- **SemEval1, SemEval2**: les deux systèmes gagnants pour le défi 2014 de SemEval 2014 (sous-tâche 1);
- **WDEmb + B + CRF**: modèle proposé onutilisant les intégrations de mot et de chemin de dépendance combinées avec des fonctionnalités d'intégration de contexte linéaire, des fonctionnalités d'intégration de contexte de dépendance et des fonctionnalités manuelles en tant qu'entrée CRF.

**b- Modèles basés sur les réseaux de neurones**

- **LSTM**: est un modèle vanille basé sur BiLSTM ;
- **RNCRF**: réseau neuronal récurrent basé sur un arbre de dépendance avec extracteur de CRF;
- **MIN**: est un cadre d'apprentissage multitâche comportant (1) deux LSTM pour l'extraction conjointe d'aspects et d'opinions, et (2) un troisième LSTM pour la discrimination de peines sentimentales et non sentimentales;
- **BiLSTM-CNN-CRF**: État de la technique issu de la communauté de reconnaissance d'entités nommées (NER).

<b>Modèle</b>	<b>F1-mesure sous tâche1 (%)</b>
CRF	74.01
Semi-CRF	68.75
WDEmb	75.16
DLIREC	73.78
SemEval1	74.55
SemEval2	73.78
WDEmb+B+CRF	75.16
LSTM	75.25
RNCRF	77.26
MIN	77.58

BiLSTM-CNN-CRF	77.80
<b>Our Constituency Tree-LSTM</b>	<b>79.65</b>

**Tableau 4.2:** Table de comparaison de note approche avec les approches existantes en termes de F1- mesure pour le jeu de données de SemEval 2014, tâche 4 (Bouras et al., 2020).

#### 4.5.1.2 La sous tâche de detection de la polarité

Pour évaluer notre modèle Tree-LSTM proposé pour la sous-tâche2: détection de polarité, nous réalisons une étude comparative avec les deux groupes de modèles suivants:

##### a- Modèles basés sur le paradigme traditionnels

- **Majority:** est une méthode de base de base, qui attribue l'étiquette de sentiment majoritaire dans l'ensemble d'apprentissage à chaque instance de l'ensemble de tests.
- **SVM simple:** est un classificateur SVM avec des fonctionnalités simples unigrammes ou bigrams;
- **SVM avec caractéristiques améliorées:** est un classificateur SVM avec des fonctions n-gram, des fonctions d'analyse et des fonctions de lexique.

##### b- Modèles basés sur un réseau de neurones LSTM

- **LSTM:** présente un LSTM standard pour la tâche de classification par exploration d'opinion basée sur l'aspect;
- **TD-LSTM:** étend le LSTM en tenant compte de l'aspect et utilise deux réseaux LSTM, l'un en amont et l'autre en arrière;
- **AE-LSTM:** représente les cibles avec des incorporations d'aspect. Tout d'abord, cette méthode modélise les mots de contexte via les réseaux LSTM, puis associe le mot états cachés à des imbrications d'aspects pour superviser la génération de vecteurs d'attention ;
- **ATAE-LSTM:** est développé sur la base de AE-LSTM. ATAE-LSTM renforce encore l'effet d'incorporation d'aspects et ajoute les incorporations d'aspects à chaque vecteur d'incorporation de mots pour représenter le contexte. L'autre conception de ATAE-LSTM est identique à celle de AE-LSTM ;
- **IAN:** implémenté via deux réseaux de neurones LSTM avec les mécanismes d'attention interactifs ;

- **RAM**: utilise le vecteur de localisation relative pour attribuer des pondérations aux vecteurs d'état cachés du mot de contexte d'origine, puis apprend le vecteur d'attention dans un mécanisme d'attention récurrent constitué d'une cellule GRU;
- **Dépendances entre aspects LSTM**: modélisez les représentations potentielles basées sur les aspects en tant que séquence pour capturer les dépendances entre aspects.

Modèle	Accuracy sous tâche2(%)
Majority	65.00
Simple SVM	66.9
Feature Enhanced SVM	72.1
LSTM	66.4
TD-LSTM	68.13
AE-LSTM	68.9
ATAE- LSTM	68.7
IAN	72.1
RAM	71.32
Inter-aspect Dependencies LSTM	72.5
<b>notre Constituency Tree-LSTM</b>	<b>92.26</b>

**Tableau 4.3:** Comparaison de notre approche avec les approches existantes en terme d'Accuracy sur une prédiction à 3 classes concernant le jeu de données SemEval 2014 Task 4 (Bouras et al., 2020).

Notre modèle	Sous-tâche1			Sous-tâche2
	P (%)	R (%)	F1 (%)	Acc (%)
Constituency Tree-LSTM	<b>83.61</b>	<b>76.04</b>	<b>79.65</b>	<b>92.26</b>

**Tableau 4.4:** Résultats de notre approche en termes de Précision (P), Recall (R), et F-mesure (F1) pour la fouille d'opinion basée aspects (Bouras et al., 2020).

## 4.6 Analyse des résultats

### 4.6.1 La sous tâche d'extraction d'aspect

Le tableau (4.2) décrit les résultats des expériences et montre les performances de notre modèle, par rapport aux méthodes présentées dans l'état de l'art basées sur la CRF ou sur les méthodes basées sur un réseau neuronal sur le jeu de données de l'ordinateur portable SemEval 2014. Nous fournissons le même jeu de données annoté (les étiquettes d'aspect et les étiquettes d'opinion sont incluses pour la formation), notre proposition basée sur le réseau Tree-LSTM réalise une performance considérable par rapport à tous les autres modèles. Notre modèle surperforme CRF, WDEmb et WDEmb avec

respectivement 5,64%, 4,4% et 4,4%. D'une part, le modèle Semi-CRF présente les performances les plus faibles par rapport aux autres modèles. D'autre part, notre modèle de Tree-LSTM présente une performance supérieure à comparer avec le BiLSTM-CNN-CRF avec 1,85%. Le modèle LSTM présente les performances les plus faibles par rapport à notre proposition, principalement parce que le modèle basé sur LSTM mélange l'aspect et le contexte, de sorte qu'il ne peut pas donner les informations complètes sur l'aspect pour prédire la polarité correspondante.

#### 4.6.2 La sous tâche de la détection de la polarité

Nos résultats expérimentaux fournis par notre modèle pour la détection de polarité sous-tâche2 sont résumés dans le tableau (4.3). nous avons comparé notre arborescence de circonscription-LSTM à deux groupes de modèles: le modèle traditionnel basé sur un paradigme et les modèles basés sur un réseau neuronal LSTM avec un paradigme d'apprentissage en profondeur. Il est remarquable que Tree-LSTM avec l'arbre de circonscription surpasse les systèmes existants en matière de sous-tâche de classification des sentiments et offre une précision comparable à l'ensemble de données de pointe sur l'ordinateur portable SemEval. En particulier, nous constatons qu'il surpasse le modèle de base avec une marge remarquable et qu'il est similaire à la fois pour le SVM simple et pour celui doté de fonctionnalités améliorées. D'une part, le modèle proposé dépasse de manière significative les performances de LSTM et de TD-LSTM. D'autre part, comparé à AE-LSTM et ATAE-LSTM, où les deux modèles sont assez similaires et que le dernier est une extension du précédent, notre modèle améliore beaucoup plus la précision des jeux de données d'ordinateurs portables par rapport aux modèles indiqués.

Le modèle Tree-LSTM avec un arbre de circonscription améliore la précision des jeux de données pour ordinateurs portables par rapport à trois modèles récents: IAN, RAM et les dépendances inter-aspects LSTM sur le même jeu de données, ceci pour la simple raison que notre proposition effectue l'extraction d'opinion basée sur les aspects sur la sous-tâche1 et la sous-tâche2 sans ingénierie de fonctionnalité supplémentaire. La principale contribution est que les modèles Tree-LSTM de circonscription reçoivent une supervision à la fois à la racine et aux nœuds intermédiaires de l'arborescence.

#### 4.7 Conclusion

Dans ce chapitre nous avons présenté deux contributions dans le domaine de la fouille d'opinion basée sur les aspects. La première contribution consiste à présenter une étude comparative méticuleuse, dans laquelle plus de 30 approches de recherche ont été résumées, classées, et comparées selon leurs catégories relatives (supervisées, semi-supervisées et non supervisées). D'après notre comparaison nous avons constaté que la majorité des travaux de recherche récente dans le domaine de la fouille d'opinion, particulièrement au niveau aspect sont orientés vers les approches basées sur l'apprentissage en profondeur.

A cet effet, selon notre première contribution et suivant la nouvelle tendance technologique, cette contribution a dressé un état récapitulatif des principales techniques et méthodes existantes au domaine de l'analyse d'opinion. Cette contribution a montré clairement les points forts et faibles des travaux existants. En se basant sur les résultats fournis par la première contribution, la deuxième contribution a consisté à développer une approche Deep Learning basée sur les réseaux de neurone de type LSTM arborescente (Tree-LSTM) pour l'extraction et la classification des textes porteurs d'opinion au niveau aspect. Afin de présenter notre contribution, nous avons défini les différentes étapes pour la démarche suivie pour élaborer notre réseau Tree-LSTM. Nous avons décrit de manière détaillée l'architecture de notre réseau Deep avec ses différentes couches utilisées pour l'extraction des aspects et la classification des opinions. Pour valider notre modèle et tester leur efficacité. Nous avons effectué une expérimentation avec le jeu de données de référence « SemEval 2014 ». En comparant les résultats expérimentaux de notre réseau Tree-LSTM avec d'autres approches existantes. Les résultats illustrés ont confirmé la performance de l'approche proposée.

## Conclusion générale

Le travail réalisé dans le cadre de cette thèse s'inscrit dans un domaine émergent appelé « la fouille d'opinion », qui est à son tour intégré dans le domaine de l'intelligence artificielle qui fait appel à plusieurs domaines de recherche tel que le traitement de la langue naturelle, l'apprentissage automatique, et l'apprentissage en profondeur. Nous nous sommes particulièrement concentrés sur les problématiques de l'extraction des aspects et la détection de polarité des opinions exprimées dans des textes porteurs d'opinion, écrits en langage naturel.

Nous avons développé une nouvelle approche basée sur l'apprentissage en profondeur, en utilisant du réseau de neurones de type Tree-LSTM, celui-ci manipule des données arborescentes pour l'extraction des aspects et la classification des opinions. Pour ce faire, nous avons présenté les contributions suivantes :

- En premier lieu, nous avons présenté une étude comparative entre les différentes approches existantes dans la littérature. Dans laquelle plus de 30 approches de recherche ont été résumées, classées, et comparées selon leurs catégories relatives (supervisées, semi-supervisées et non supervisées). D'après notre comparaison nous avons constaté que la majorité des travaux de recherche récente dans le domaine de la fouille d'opinion, particulièrement au niveau aspect sont orientés vers les approches basées sur l'apprentissage en profondeur. Ces modèles d'apprentissage en profondeur qui constituent une branche de l'apprentissage automatique basé sur des réseaux de neurones, récemment utilisés dans de nombreuses tâches de traitement de la langue naturelle ont présenté les meilleurs résultats, sans ingénierie manuelle des caractéristiques et avec le moins de restrictions possibles.
- En deuxième lieu, et selon nos remarques retirés de la première contribution. Nous avons présenté une approche basée sur les méthodes d'apprentissage en profondeur. Dans notre approche, l'extraction des d'aspect et la classification des opinions sont simultanément gérées dans un même réseau de neurones en profondeur de type Tree-LSTM. Les modèles traditionnels reposent souvent sur une ingénierie manuelle laborieuse des caractéristiques. Les réseaux de neurones peuvent apprendre automatiquement des représentations de caractéristiques efficaces

pour des informations sémantiques et syntaxiques dans des phrases. Nous appliquons des réseaux de neurones Tree-LSTM pour traiter les deux sous-tâches d'une manière successive.

- La comparaison de nos résultats avec d'autres modèles existants dans la littérature en terme des mesures d'évaluation, à savoir : la Précision, le Recale et le Score montrent l'efficacité de notre approche proposée et évaluée sur dataset de SemEval 2014 sur des critiques de micro-portable (laptop dataset).
- Les résultats nous ont incités à continuer à travailler sur cette thématique pour améliorer l'architecture du processus de la fouille d'opinion basé aspect.

En guise de conclusion, les résultats que nous avons obtenus sont encourageants par rapport aux résultats des autres approches proposées dans la littérature.

## Perspectives

Naturellement, l'une des caractéristiques attrayantes de la recherche est la question "comment ?". Tout en répondant à cette question de recherche, d'autres questions se posent au cours du processus. Cela s'est également produit au cours de notre thèse. Bien que la majorité de ces questions restent sans réponse, nous voudrions en abordant quelques-unes qui sont à notre avis, semblent les plus intéressantes, et pourraient constituer des avancées potentielles pour le domaine de la fouille d'opinion basée sur des aspects. Les réponses sur ses questions seront les perspectives de notre thèse comme suit :

- La première perspective porte sur le développement des approches dédiées à l'extraction des aspects implicites.
- La deuxième piste, qui nous semble importante, est l'implémentation d'un autre modèle d'apprentissage approfondi avec d'autre type de réseau de neurone.
- Il serait également important de développer des approches dédiées à l'extraction des aspects sur des datasets en langue Arabe et dialecte algérien.
- Malgré que, les trois itérations réussies de la tâche partagée de la fouille d'opinion basée sur les aspects lors des ateliers SemEval 2014, 2015 et 2016, le développement des datasets de taille importante dédiées à la fouille basée sur les aspects reste toujours une perspective à considérer.

## References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Isard, M. (2016). *Tensorflow: A system for large-scale machine learning*. Paper presented at the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16).
- Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.
- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). *Sentiment analysis of twitter data*. Paper presented at the Proceedings of the Workshop on Language in Social Media (LSM 2011).
- Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25, 821-837.
- ALMasri, M., Berrut, C., & Chevallet, J.-P. (2016). *A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information*. Paper presented at the European conference on information retrieval.
- Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media*. Paper presented at the Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01.
- Bafna, K., & Toshniwal, D. (2013). Feature based summarization of customers' reviews of online products. *Procedia Computer Science*, 22, 142-151.
- Bagheri, A., Saraee, M., & de Jong, F. (2013). *An unsupervised aspect detection model for sentiment analysis of reviews*. Paper presented at the International Conference on Application of Natural Language to Information Systems.
- Balikas, G., & Amini, M.-R. (2016). *Multi-label, multi-class classification using polylingual embeddings*. Paper presented at the European Conference on Information Retrieval.
- Bancken, W., Alfarone, D., & Davis, J. (2014). *Automatically detecting and rating product aspects from textual customer reviews*. Paper presented at the Proceedings of the 1st international workshop on interactions between data mining and natural language processing at ECML/PKDD.
- Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine Learning research*, 3(Feb), 1137-1155.
- Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). *Greedy layer-wise training of deep networks*. Paper presented at the Advances in neural information processing systems.
- Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2), 157-166.
- Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., . . . Bengio, Y. (2010). *Theano: A CPU and GPU math compiler in Python*. Paper presented at the Proc. 9th Python in Science Conf.
- Birdsall, W. F. (2007). Web 2.0 as a social movement. *Webology*, 4(2), 5-11.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.
- Bouras, D., Amroune, M., & Ben Djanna, H. (2019). Techniques and Trends for Fine-Grained Opinion Mining and Sentiment Analysis: Recent Survey. *Recent Patents on Computer Science*, 12, 1-13.
- Bouras, D., Amroune, M., Bendjenna, H., & Bendib, I. (2020). Improving Fine-Grained Opinion Mining Approach with a Deep Constituency Tree-Long Short Term Memory Network and Word Embedding. *Recent Advances in Computer Science and Communications (formerly recent patents on computer science and communications)*, 13(1). doi: <https://doi.org/10.2174/2666255813999200922142212>



- Cardie, C., Wiebe, J., Wilson, T., & Litman, D. J. (2003). *Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering*. Paper presented at the New directions in question answering.
- Chen, P., Sun, Z., Bing, L., & Yang, W. (2017). *Recurrent attention network on memory for aspect sentiment analysis*. Paper presented at the Proceedings of the conference on empirical methods in natural language processing.
- Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). *Torch7: A matlab-like environment for machine learning*. Retrieved from
- Collobert, R., & Weston, J. (2008). *A unified architecture for natural language processing: Deep neural networks with multitask learning*. Paper presented at the Proceedings of the 25th international conference on Machine learning.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine Learning research*, 12(Aug), 2493-2537.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., & Vallejo, C. G. (2013). ‘Long autonomy or long delay?’ The importance of domain in opinion mining. *Expert Systems with Applications*, 40(8), 3174-3184.
- Das, S., & Chen, M. (2001). *Yahoo! for Amazon: Extracting market sentiment from stock message boards*. Paper presented at the Proceedings of the Asia Pacific finance association annual conference (APFA).
- Dave, K., Lawrence, S., & Pennock, D. M. (2003). *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. Paper presented at the Proceedings of the 12th international conference on World Wide Web.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *Imagenet: A large-scale hierarchical image database*. Paper presented at the 2009 IEEE conference on computer vision and pattern recognition.
- Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4), 197-387.
- Dini, L., & Mazzini, G. (2002). Opinion classification through information extraction. *WIT Transactions on Information and Communication Technologies*, 28.
- Du, J., Chan, W., & Zhou, X. (2014). *A product aspects identification method by using translation-based language model*. Paper presented at the 2014 22nd International Conference on Pattern Recognition.
- Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine Learning research*, 12, 2121-2159.
- Eirinaki, M., Pissal, S., & Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4), 1175-1184.
- Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2), 179-211.
- Fahrni, A., & Klenner, M. (2008). *Old wine or warm beer: Target-specific sentiment analysis of adjectives*. Paper presented at the Proc. of the Symposium on Affective Language in Human and Machine, AISB.
- Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4), 82-89.
- Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.
- Ganu, G., Elhadad, N., & Marian, A. (2009). *Beyond the stars: improving rating predictions using review text content*. Paper presented at the WebDB.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*: MIT press.

- Graham, P. (2005). Web 2.0. . Retrieved from <http://www.paulgraham.com/web20interview.html>
- Graves, A., Jaitly, N., & Mohamed, A.-r. (2013). *Hybrid speech recognition with deep bidirectional LSTM*. Paper presented at the IEEE workshop on automatic speech recognition and understanding.
- Graves, A., Mohamed, A.-r., & Hinton, G. (2013). *Speech recognition with deep recurrent neural networks*. Paper presented at the 2013 IEEE international conference on acoustics, speech and signal processing.
- Hai, Z., Chang, K., & Cong, G. (2012). *One seed to find them all: mining opinion features via association*. Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management.
- Hai, Z., Chang, K., Kim, J.-J., & Yang, C. C. (2013). Identifying features in opinion mining via intrinsic and extrinsic domain relevance. *IEEE Transactions on knowledge and data engineering*, 26(3), 623-634.
- Hatzivassiloglou, V., & McKeown, K. R. (1997). *Predicting the semantic orientation of adjectives*. Paper presented at the Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1984). *Distributed representations: Carnegie-Mellon University Pittsburgh, PA*.
- Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527-1554.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- Hoffman, T. (1999). *Probabilistic latent semantic indexing*. Paper presented at the Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, 1999.
- Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.
- Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems*, 52(3), 674-684.
- Huang, B., Ou, Y., & Carley, K. M. (2018). *Aspect level sentiment classification with attention-over-attention neural networks*. Paper presented at the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation.
- Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). *Improving word representations via global context and multiple word prototypes*. Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.
- Huang, S., Liu, X., Peng, X., & Niu, Z. (2012). *Fine-grained product features extraction and categorization in reviews opinion mining*. Paper presented at the 12th International Conference on Data Mining Workshops.
- Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3), 574-591.
- Jabreel, M., Hassan, F., & Moreno, A. (2018). Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks *Advances in Hybridization of Intelligent Methods* (pp. 39-55): Springer.

- Jebbara, S., & Cimiano, P. (2016). *Aspect-based sentiment analysis using a two-step neural network architecture*. Paper presented at the Semantic Web Evaluation Challenge.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., . . . Darrell, T. (2014). *Caffe: Convolutional architecture for fast feature embedding*. Paper presented at the Proceedings of the 22nd ACM international conference on Multimedia.
- Jindal, N., & Liu, B. (2006a). *Identifying comparative sentences in text documents*. Paper presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.
- Jindal, N., & Liu, B. (2006b). *Mining comparative sentences and relations*. Paper presented at the AAAI.
- Joachims, T. (1996). *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*.
- Jordan, M. (1986). *Serial order: a parallel distributed processing approach*. Technical report, June 1985-March 1986.
- Joshi, M., Das, D., Gimpel, K., & Smith, N. A. (2010). *Movie reviews and revenues: An experiment in text regression*. Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.
- Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., . . . Borchers, A. (2017). *In-datacenter performance analysis of a tensor processing unit*. Paper presented at the 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA).
- Kim, S.-M., & Hovy, E. (2004). *Determining the sentiment of opinions*. Paper presented at the Proceedings of the 20th international conference on Computational Linguistics.
- Kim, S.-M., & Hovy, E. (2006). *Extracting opinions, opinion holders, and topics expressed in online news media text*. Paper presented at the Proceedings of the Workshop on Sentiment and Subjectivity in Text.
- Kim, Y. (2014a). *Convolutional neural networks for sentence classification*. Paper presented at the In Proceedings of the 2014 Conference on EMNLP.
- Kim, Y. (2014b). *Convolutional neural networks for sentence classification*.
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*: MIT press.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. Paper presented at the Advances in neural information processing systems.
- Kuzi, S., Shtok, A., & Kurland, O. (2016). *Query expansion using word embeddings*. Paper presented at the Proceedings of the 25th ACM international on conference on information and knowledge management.
- Lakkaraju, H., Socher, R., & Manning, C. (2014). *Aspect specific sentiment analysis using hierarchical deep learning*. Paper presented at the NIPS Workshop on deep learning and representation learning.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). *Neural architectures for named entity recognition*. *arXiv preprint arXiv:1603.01360*.
- Le, P., & Zuidema, W. (2015). *Compositional distributional semantics with long short term memory*. Paper presented at the in the Joint Conference on Lexical and Computational Semantics (\*SEM), 2015.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *nature*, 521(7553), 436.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *nature* 521.
- LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). *Handwritten digit recognition with a back-propagation network*. Paper presented at the Advances in neural information processing systems.

- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
- Li, S., Wang, R., & Zhou, G. (2012). *Opinion target extraction using a shallow semantic parsing framework*. Paper presented at the Twenty-sixth AAAI conference on artificial intelligence.
- Li, Y., Qin, Z., Xu, W., & Guo, J. (2015). A holistic model of mining product aspects and associated sentiments from online reviews. *Multimedia Tools and Applications*, 74(23), 10177-10194.
- Li, Y., Wang, H., Qin, Z., Xu, W., & Guo, J. (2014). *Confidence estimation and reputation analysis in aspect extraction*. Paper presented at the 2014 22nd International Conference on Pattern Recognition.
- Ling, W., Luís, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., . . . Trancoso, I. (2015). Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*.
- Lipsman, A. (2007). Online consumer-generated reviews have significant impact on offline purchase behavior. *comScore. Inc. Industry Analysis*, 2-28.
- Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*: Springer Science & Business Media.
- Liu, B. (2012). Sentiment analysis and opinion mining, volume 5 of Synthesis lectures on human language technologies. *Morgan & Claypool Publishers*, May.
- Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., & Zhou, M. (2007). *Low-quality product review detection in opinion summarization*. Paper presented at the Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).
- Liu, K., Xu, H. L., Liu, Y., & Zhao, J. (2013). *Opinion target extraction using partially-supervised word alignment model*. Paper presented at the Twenty-Third International Joint Conference on Artificial Intelligence.
- Liu, K., Xu, L., & Zhao, J. (2012). *Opinion target extraction using word-based translation model*. Paper presented at the Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning.
- Liu, K., Xu, L., & Zhao, J. (2015). Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. *IEEE Transactions on knowledge and data engineering*, 27(3), 636-650.
- Liu, P., Joty, S., & Meng, H. (2015). *Fine-grained opinion mining with recurrent neural networks and word embeddings*. Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing.
- Liu, Q., Liu, B., Zhang, Y., Kim, D. S., & Gao, Z. (2016). *Improving opinion aspect extraction using semantic similarity and aspect associations*. Paper presented at the Thirtieth AAAI Conference on Artificial Intelligence, Barcelona, Spain.
- Ma, B., Zhang, D., Yan, Z., & Kim, T. (2013). An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews. *Journal of Electronic Commerce Research*, 14(4), 304.
- Ma, B. L. W. H. Y., & Liu, B. (1998). *Integrating classification and association rule mining*. Paper presented at the Proceedings of the fourth international conference on knowledge discovery and data mining.
- Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision support systems*, 53(4), 680-688.
- Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2013). *Opinion zoom: a modular tool to explore tourism opinions on the web*. Paper presented at the Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03.

- Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*, 41(17), 7764-7775.
- Marrese-Taylor, E., Velásquez, J. D., Bravo-Marquez, F., & Matsuo, Y. (2013). Identifying customer preferences about tourism products using an aspect-based opinion mining approach. *Procedia Computer Science*, 22, 182-191.
- Maynard, D., & Funk, A. (2011). *Automatic detection of political opinions in tweets*. Paper presented at the Extended Semantic Web Conference.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4), 1093-1113.
- Mejova, Y., & Srinivasan, P. (2011). *Exploring feature definition and selection for sentiment classifiers*. Paper presented at the Fifth international AAAI conference on weblogs and social media.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). *Recurrent neural network based language model*. Paper presented at the Eleventh annual conference of the international speech communication association.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. Paper presented at the Advances in neural information processing systems.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4), 235-244.
- Moghaddam, S., & Ester, M. (2010). *Opinion digger: an unsupervised opinion miner from unstructured product reviews*. Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management.
- Mohammad, S., Dunne, C., & Dorr, B. (2009). *Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus*. Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2.
- Mohammad, S. M., & Yang, T. W. (2011). *Tracking sentiment in mail: How genders differ on emotional axes*. Paper presented at the Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis.
- Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002). *Mining product reputations on the web*. Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.
- Nasukawa, T., & Yi, J. (2003). *Sentiment analysis: Capturing favorability using natural language processing*. Paper presented at the Proceedings of the 2nd international conference on Knowledge capture.
- Nickolls, J., Buck, I., & Garland, M. (2008). *Scalable parallel programming with CUDA*. Paper presented at the 2008 IEEE Hot Chips 20 Symposium (HCS).
- O'Brien, C. (2011). The emergence of the social media empowered consumer. *Irish marketing review*, 21(1/2), 32-40.
- O'Reilly, T. (2005). What is web 2.0? Retrieved from <https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html>
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1-2), 1-135.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques*. Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.

- Pascanu, R., Mikolov, T., & Bengio, Y. (2013). *On the difficulty of training recurrent neural networks*. Paper presented at the International conference on machine learning.
- Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation*. Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).
- Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.-S., . . . De Clercq, O. (2016). *Semeval-2016 task 5: Aspect based sentiment analysis*. Paper presented at the Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016).
- Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2014). *Semeval-2014 task 4: Aspect based sentiment analysis*. Paper presented at the in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)
- Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). *Semeval-2015 task 12: Aspect based sentiment analysis*. Paper presented at the Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).
- Popescu, A., & Etzioni, O. (2005). Product features and Opinions from Reviews in Proceedings of HLT-EMNLP 2005. *Vancouver, Canada: ACL*, 339-346.
- Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108, 42-49.
- Poria, S., Cambria, E., Ku, L.-W., Gui, C., & Gelbukh, A. (2014). *A rule-based approach to aspect extraction from product reviews*. Paper presented at the Proceedings of the second workshop on natural language processing for social media (SocialNLP).
- Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). Challenges of sentiment analysis in social networks: An overview *Sentiment analysis in social networks* (pp. 1-11): Elsevier.
- Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., & Chen, C. (2010). DASA: dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications*, 37(9), 6182-6191.
- Quan, C., & Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences*, 272, 16-28.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81-106.
- Raina, R., Madhavan, A., & Ng, A. Y. (2009). *Large-scale deep unsupervised learning using graphics processors*. Paper presented at the Proceedings of the 26th annual international conference on machine learning.
- Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89, 14-46.
- Read, J., & Carroll, J. (2009). *Weakly supervised techniques for domain-independent sentiment classification*. Paper presented at the Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.
- Ruder, S., Ghaffari, P., & Breslin, J. G. (2016). A hierarchical model of reviews for aspect-based sentiment analysis. *arXiv preprint arXiv:1609.02745*.
- Sadikov, E., Parameswaran, A., & Venetis, P. (2009). *Blogs as predictors of movie success*. Paper presented at the Third International AAAI Conference on Weblogs and Social Media.
- Samha, A. K., Li, Y., & Zhang, J. (2014). Aspect-based opinion extraction from customer reviews. *arXiv preprint arXiv:1404.1982*.
- Semeniuta, S., Severyn, A., & Barth, E. (2016). Recurrent dropout without memory loss. *arXiv preprint arXiv:1603.05118*.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Singhal, P., & Bhattacharyya, P. (2016). Sentiment analysis and deep learning: a survey. *Center for Indian Language Technology, Indian Institute of Technology, Bombay*.
- Smith, A., & Anderson, M. (2016). Online shopping and e-commerce. *Pew Research Center*, 19.

- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank*. Paper presented at the Proceedings of the conference on empirical methods in natural language processing.
- Song, Y., Wang, J., Jiang, T., Liu, Z., & Rao, Y. (2019). Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.
- Stevens, L. (2016). Survey shows rapid growth in online shopping. *The Wall Street Journal*.
- Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36, 10-25.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks*. Paper presented at the Advances in neural information processing systems.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). *Going deeper with convolutions*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. doi:arXiv preprint arXiv:1503.00075
- Tan, P.-N., Steinbach, M., & Kumar, V. others. 2006. Introduction to data mining. Vol. 1: Pearson Addison Wesley Boston.
- Tang, D., Qin, B., & Liu, T. (2016). Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.
- Thanaki, J. (2017). *Python Natural Language Processing*: Packt Publishing Ltd.
- Thet, T. T., Na, J.-C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6), 823-848.
- Tong, R. M. (2001). *An operational system for detecting and tracking opinions in on-line discussion*. Paper presented at the Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification.
- Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). *Predicting elections with twitter: What 140 characters reveal about political sentiment*. Paper presented at the Fourth international AAAI conference on weblogs and social media.
- Turney, P. D. (2002). *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*. Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics.
- Vapnik, V. (1995). *The Nature of Statistical Learning Theory*: USA Springer Google Scholar New York, NY.
- Walaa, M., Ahmed, H., & Hoda, K. (2014). Combined Algorithm for Data Mining using Association rules. *Ain Shams of Electrical Engineering J*, 1(1).
- Wang, W., Pan, S. J., Dahlmeier, D., & Xiao, X. (2016). Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679*.
- Wei, C.-P., Chen, Y.-M., Yang, C.-S., & Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and E-Business Management*, 8(2), 149-167.
- Wiebe, J., & Mihalcea, R. (2006). *Word sense and subjectivity*. Paper presented at the Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics.
- Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis*. Paper presented at the Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.

- Xu, H., Liu, B., Shu, L., & Yu, P. S. (2018). Double embeddings and cnn-based sequence labeling for aspect extraction. *arXiv preprint arXiv:1805.04601*.
- Xue, W., & Li, T. (2018). Aspect based sentiment analysis with gated convolutional networks. doi:arXiv preprint arXiv:1805.07043
- Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management*, 52(7), 850-858.
- Yang, B., & Cardie, C. (2013). *Joint inference for fine-grained opinion extraction*. Paper presented at the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
- Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M., & Zhou, M. (2016). Unsupervised word and dependency path embeddings for aspect term extraction. doi:arXiv preprint arXiv:1605.07843
- Yu, J., Zha, Z.-J., Wang, M., & Chua, T.-S. (2011). *Aspect ranking: identifying important product aspects from online consumer reviews*. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.
- Zamani, H., & Croft, W. B. (2016). *Embedding-based query language models*. Paper presented at the Proceedings of the 2016 ACM international conference on the theory of information retrieval.
- Zaremba, W., & Sutskever, I. (2014). Learning to execute. doi:arXiv preprint arXiv:1410.4615
- Zhang, L., & Liu, B. (2011). *Identifying noun product features that imply opinions*. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2.
- Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253.
- Zhao, Q., Wang, H., Lv, P., & Zhang, C. (2014). *A bootstrapping based refinement framework for mining opinion words and targets*. Paper presented at the Proceedings of the 23rd ACM international conference on conference on information and knowledge management.