République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie
Département de l'Informatique

THESE

Présentée pour l'obtention du diplôme de
Doctorat en sciences en Informatique

THEME

Techniques d'apprentissage automatique pour l'analyse et la fouille des sentiments dans les réseaux sociaux

---

Présentée par : **Nedioui Med Abdelhamid**

Devant le jury composé de :

Pr. DJEDI Noureddine Université Mohamed Khider Biskra	**Président**
Pr. MOUSSAOUI Abdelouahab Université Ferhat Abbas Sétif 1	**Rapporteur**
Pr. BABAHENINI Mohamed Chaouki Université Mohamed Khider Biskra	**Co-Rapporteur**
Pr. CHERIF Foudil Université Mohamed Khider Biskra	**Examinateur**
Pr. Mohamed Benmohammed Université de constantine 2	**Examinateur**
Dr. LEJDEL Brahim Université Hamma Lakhder EL-Oued	**Examinateur**

Année Universitaire **2020/2021**

Résumé

Aujourd’hui, avec une large diffusion des réseaux sociaux, d’énormes quantités de données sont générées sous forme de points de vue, d'émotions, d'opinions et de sentiments sur différents événements sociaux, produits, marques, politiques, etc. Les sentiments des utilisateurs exprimés sur le Web ont une grande influence sur les lecteurs, les vendeurs de produits et les politiciens. La forme non structurée de données provenant des médias sociaux doit être analysée et bien structurée et à cette fin, l'analyse des sentiments a attiré une attention considérable. L'analyse des sentiments est utilisée pour classer les sentiments exprimés de différentes manières telles que négatives, positives ou neutres. Le défi de l'analyse des sentiments est le manque d'étiquettes suffisantes de données dans les réseaux sociaux. Afin de résoudre ce problème, l'analyse des sentiments ainsi que l'analyse des réseaux sociaux ont été fusionnées pour avoir des résultats beaucoup plus pertinents. Dans cette thèse nous mettons en évidence les dernières études concernant la mise en œuvre de modèles d'analyse des sentiments tels que l'apprentissage automatique et celles basées sur un lexique à l’instar des techniques de détection de communautés de sentiments pour résoudre les différents problèmes liées à d'analyse de sentiments.

Mots clés :

Fouille de données, apprentissage automatique, analyse des sentiments, réseaux sociaux, communautés des sentiments.

Abstract

Today, with the wide spread of social media, huge amount of data is generated in the form of views, emotions, opinions and sentiments about different social events, products, brands, politicians, etc. The feelings of users expressed on the web have a great influence on readers, product sellers and politicians. The unstructured form of data coming from social media needs to be analyzed and well-structured and for this purpose sentiment analysis has attracted considerable attention. Sentiment analysis is used to classify feelings expressed in different ways such as negative, positive or neutral. The challenge of sentiment analysis is the lack of sufficient data tags in social media. In order to solve this problem, sentiment analysis as well as social media analysis have been merged to have much more relevant results. In this thesis we highlight the latest studies concerning the implementation of sentiment analysis models such as machine learning and those based on a lexicon like the techniques of detection of communities of sentiment to resolve the different problems related to sentiment analysis.

Keywords:

Data mining, machine learning, sentiment analysis, social media, sentiment communities.

ملخص

اليوم، مع الانتشار الواسع لوسائل التواصل الاجتماعي ، يتم إنشاء كميات هائلة من البيانات في شكل وجهات نظر وعواطف وآراء ومشاعر حول مختلف الأحداث الاجتماعية والمنتجات والعلامات التجارية والسياسيين ، إلخ. تؤثر مشاعر المستخدمين التي يتم التعبير عنها على الويب بشكل كبير على القراء وبائعي المنتجات والسياسيين. يحتاج الشكل غير المهيكل للبيانات القادمة من وسائل التواصل الاجتماعي إلى التحليل والتنظيم الجيد ولهذا الغرض اجتذب تحليل المشاعر اهتمامًا كبيرًا. يستخدم تحليل المشاعر لتصنيف المشاعر التي يتم التعبير عنها بطرق مختلفة مثل السلبية أو الإيجابية أو المحايدة. التحدي المتمثل في تحليل المشاعر هو عدم وجود علامات بيانات كافية في وسائل التواصل الاجتماعي. لحل هذه المشكلة ، تم دمج تحليل المشاعر بالإضافة إلى تحليل وسائل التواصل الاجتماعي للحصول على نتائج أكثر صلة. نسلط الضوء في هذه الرسالة على أحدث الدراسات المتعلقة بتنفيذ نماذج تحليل المشاعر مثل التعلم الآلي وتلك التي تعتمد على معجم مثل تقنيات اكتشاف مجتمعات المشاعر لحل المشكلات المختلفة المتعلقة بتحليل المشاعر.

الكلمات المفتاحية :

التنقيب عن البيانات ، التعلم الآلي ، تحليل المشاعر ، وسائل التواصل الاجتماعي ، مجتمعات المشاعر.

Remerciements

Tout d’abord, je voudrais remercier mon directeur de thèse, M. Moussaoui Abdelouahab, professeur à l’université de Ferhat Abbas Sétif 1 pour son soutien durant mon doctorat. Il était toujours disponible pour m’aider à avancer dans ma recherche et la rédaction de cette thèse. Son support était essentiel pour que ce travail puisse voir le jour.

J’adresse également mes remerciements à M. BABAHENINI Mohamed Chaouki, professeur à l’université de Mohamed Khider Biskra, en tant que co-directeur, il m’a guidé, encouragé et il m’a fait confiance durant toutes ces années de recherche.

Je tiens à remercier sincèrement le président de jury M. DJEDI Noureddine, professeur à l’université de Mohamed Khider Biskra, qui m’a fait le grand honneur de présider ce jury.

Mes remerciements vont aussi aux membres du jury, M. CHERIF Foudil, professeur à l’université de Mohamed Khider Biskra et à M. Mohamed Benmohammed, professeur à l’université de constantine 2 d’avoir accepté de juger mon travail et pour m’avoir fait l’honneur de participer à mon jury de soutenance en tant qu’examinateurs.

Mes remerciements vont également à M. LEJDEL Brahim, maître de conférences, MCA, à l’université Hamma Lakhdar EL-Oued, d’avoir bien voulu examiner ce travail de thèse.

Enfin, je ne saurais terminer sans remercier mes amis et toute ma famille qui ont toujours été au plus près de moi en veillant à ma réussite.

Dédicace

Je veux dédier cette thèse à l'âme de mon père, ma femme et à tous mes enfants.

Table des matières

Chapitre 3 Analyse des sentiments dans les réseaux sociaux

Chapitre 4 Les communautés de sentiments

Chapitre 5 Contributions

Table des figures

Liste des tableaux

Introduction générale

L'analyse des sentiments ou l'extraction d'opinions peut être définie comme une application particulière de la fouille de données, qui vise à agréger et à extraire des émotions et des sentiments de différents types de documents [1]. La quantité de données disponibles sur le Web augmente de façon exponentielle. Cependant, ces données sont principalement décrites dans un format non structuré et ne peuvent donc pas être traitées que par machine. Par conséquent, les techniques d'exploration de graphes et de traitement du langage naturel (PNL) peuvent contribuer à la distillation des connaissances et des opinions à partir de l'énorme quantité d'informations présentes sur le Web.

L'analyse des sentiments peut améliorer les capacités des systèmes de gestion de la relation client et de recommandation en permettant, par exemple, de découvrir par quelles fonctionnalités les clients sont particulièrement intéressés, ou à exclure des annonces les éléments qui ont reçu des commentaires défavorables. De même, elle peut être utilisée dans la communication sociale pour améliorer les systèmes antispams par exemple.

Le Business Intelligent peut également bénéficier de l'analyse des sentiments. Depuis que la prédiction de l'attitude du public envers une marque ou un produit est devenue d'une importance cruciale pour les entreprises, beaucoup d’efforts ont été déployés pour développer de nouvelles stratégies de marketing impliquant l'extraction d'opinions et de sentiments.
Les réseaux sociaux sont en effet un moyen populaire de partager des données et des idées, et ont connu une diffusion toujours croissante. La quantité de données générées en 30 secondes sur Internet est d'environ 600 Go de trafic. Ce fait confirme que l'information en ligne, avec un focus particulier sur les réseaux sociaux, est devenue une source de big data. Par exemple, compte tenu du cas spécifique de la communauté Twitter, chaque minute, plus de 320 nouveaux comptes sont créés et plus de 98 000 tweets sont publiés. Cela fait de l'analyse du microblogage Twitter un domaine de premier plan et important pour la veille économique et les stratégies marketing. Une multiplicité d'utilisateurs peuple ce réseau social, partageant différents types d'informations. L'âge moyen des utilisateurs de Twitter varie de 14 à 60 ans, répartis également entre les individus des deux sexes. Ainsi, parmi la multitude de tweets, on peut vouloir récupérer des

informations associées à des sujets pertinents spécifiques et identifier la polarité et la caractérisation affective associées.

Nous constatons, à cet effet, que la plupart des applications d'analyse des sentiments impliquent des réseaux sociaux à l'instar de Twitter, FacTwitter, Facebook ou d'autres communautés numériques en temps réel.

Dans cette thèse nous avons proposés une nouvelle approche pour la détection de communautés dans les réseaux sociaux basées sur des cliques de différentes tailles. La détection des communautés se fait par l'exploration en profondeur du graphe cible en parcourant les sommets un par un et à chaque fois qu'on tombe dans un sommet déjà visité, tous les sommets situés dans ce parcours appartenant au même circuit. La thèse est organisée de la façon suivante :
Dans le premier chapitre, nous avons présenté les définitions et la terminologie utilisées dans cette thèse. Nous avons vu dans l'ordre, le processus de l'analyse de sentiments, les applications et les tâches de l'analyse de sentiments et finalement les défis rencontrés au cours de cette analyse.

Dans le deuxième chapitre, un état de l'art résumant les approches les plus prometteuses de classification des sentiments. Ces approches se divisent en deux catégories. La première catégorie concerne les méthodes d'apprentissage automatique qui consistent à classer les documents à partir d'une base de données d'apprentissage. La deuxième catégorie concerne les méthodes de classification basée sur le lexique qui repose sur un lexique des sentiments, une collection de termes de sentiments connus et précompilés.

Dans le troisième chapitre, nous abordons d'une manière très concise les différents types de réseaux sociaux, soulignons leur historique, leur définitions et leur techniques d'analyses et en outre, discutons les différentes techniques d'analyse des sentiments dans les réseaux sociaux.

Dans le quatrième chapitre, nous décrivons un nouveau concept appelé «communauté de sentiments», pour étudier les sentiments et les relations des utilisateurs sur les réseaux sociaux. Nous présentons une analyse critique des travaux de littérature liés aux problématiques de détection communautaire des sentiments.

Dans le cinquième chapitre, nous avons proposé une approches pour la détection de communautés dans les réseaux sociaux . Cette dernière se concentre sur la découverte de communautés d'utilisateurs interconnectés qui partagent des sentiments communes sur les réseaux sociaux. Contrairement aux méthodes de détection de communauté classiques existantes qui ne prennent en compte que la connectivité dans la structure du réseau, notre approche prend en compte le nouveau concept de communauté de sentiments, qui utilise à la fois les relations des utilisateurs dans les réseaux sociaux et leurs sentiments. Il est inclue dans ce chapitre une description de la réalisation des différentes approches ainsi qu'aux différents résultats des tests qui lui ont permis d'évaluer ses méthodes.

Cette thèse se termine par une conclusion qui, en revenant sur les grandes thématiques qui nous aurons guidés tout au long de cette lecture, porte sur le bilan de notre recherche, sur l'ensemble des techniques d'analyse des sentiments apportées par cette thèse et finalement ses limites. Cette conclusion donne également l'occasion d'exprimer les perspectives de nos travaux de recherche.

Chapitre 1
L'Analyse des sentiments

Sommaire

1 Introduction

Ce que pense l’être humain a toujours été un élément essentiel de la méthodologie de choix. Ces dernières années, l'augmentation exponentielle de l'utilisation d'Internet et des échanges d'opinions et des sentiments, qui sont exprimés de différentes manières, y compris la quantité de détails donnés, le type de vocabulaire utilisé, le contexte d'écriture, les argots et les variations linguistiques rend l'analyse de sentiments manuelle fastidieuse et presque impossible [144]. Cette explosion de données a conduit à une augmentation considérable de la demande d’outils d’analyse des sentiments de la part des entreprises désireuses de suivre l’opinion des gens sur ces entreprises et sur leurs produits et services, mais aussi par les chercheurs en sciences sociales afin d'exploiter les informations qui circulent entre les différents acteurs et leurs opinons et sentiments par rapport à la qualité des services et des produits de ces entrepris. Pour répondre à cette demande croissante de tels outils, de plus en plus des chercheurs development de nouveaux outils pour effectuer une analyse des sentiments, beaucoup d'entre eux prétendant pouvoir effectuer une analyse des sentiments de tout type de document dans tous les domaines. Actuellement, il n'existe pas encore d'outils d'analyse des sentiments «prêts à l'emploi» fonctionnant dans plusieurs domaines. La principale raison pour laquelle l'analyse des sentiments est si difficile est que les mots ont souvent des significations différentes et sont associés à des émotions distinctes selon le domaine dans lequel ils sont utilisés. Il existe des situations où différentes formes d'un même mot seront associées à des sentiments différents. Par exemple, dans les commentaires des clients le mot «amélioré» était associé à des commentaires positifs, mais «améliorer» était plus souvent utilisé dans les commentaires négatifs.

2 Analyse des sentiments - définitions

Dans la littérature, l’analyse de sentiments (SA) reçoit différentes dénominations ou termes communs, on trouve entre autres l'extraction d'opinions (opinion mining OM), l'analyse de subjectivité, l'analyse des émotions et l'extraction de l'évaluation et autres. Les plus souvent utilisées dans la littérature sont l'analyse des sentiments et l'exploration d'opinions (MO). Selon [25], ce sont deux concepts similaires qui désignent le même domaine d'étude, qui lui-même peut être considéré comme un sous-domaine de l'analyse de subjectivité. La référence [25] déclare que l’analyse des sentiments est un domaine de recherche dans le domaine de l'exploration de

texte et le définit comme le traitement informatique des opinions, des sentiments et de la subjectivité du texte. Sans surprise, il y a eu une certaine confusion parmi les chercheurs sur la différence entre le sentiment et l'opinion, débattant ainsi de la question de savoir si le domaine devrait être appelé analyse des sentiments ou exploitation d’opinion.

Dans le dictionnaire collégial de Merriam-Webster [143], le sentiment est défini comme une attitude, une pensée ou un jugement suscité par le sentiment, tandis que l'opinion est définie comme une vue, un jugement ou une évaluation formé dans l'esprit sur une question particulière. La différence est assez subtile et chacun d'eux contient certains éléments de l'autre. Les définitions indiquent qu’une opinion est davantage une vue concrète d’une personne sur quelque chose, alors qu’un sentiment est davantage un sentiment. Par exemple, la phrase «Je suis préoccupé par la situation politique actuelle» exprime un sentiment, tandis que la phrase «Je pense que la politique ne va pas bien» exprime une opinion. Si quelqu'un dit la première phrase d'une conversation, nous pouvons répondre en disant "Je partage votre sentiment", mais pour la deuxième phrase, nous dirions normalement "Je suis d'accord / pas d'accord avec vous. " Cependant, les significations sous-jacentes des deux phrases sont strictement liées car le sentiment décrit dans la première phrase est susceptible d'être un sentiment causé par l'opinion dans la deuxième phrase. À l'inverse, la première phrase sentimentale implique une opinion négative sur la politique, ce que dit la deuxième phrase. Bien que dans la plupart des cas, les opinions impliquent des sentiments, certaines opinions ne le font pas, comme "Je pense qu'il gagnera à la prochaine élection présidentielle."

D’une manière générale, l’analyse de sentiment est une procédure permettant de suivre les opinions des clients/auteurs autour d'un objet ou d'un sujet spécifique. Elle comprend la création d'un cadre pour recueillir et examiner les opinions sur l'élément émises dans les entrées de blog, les remarques, les audits ou les tweets.

3 Nécessité de l’analyse des sentiments

Avec la croissance des sites des réseaux sociaux en ligne, par exemple, les forums, les sites de critiques, les blogs et les micro-blogs, l'enthousiasme pour l'extraction d'opinions s'est considérablement développé. Aujourd'hui, les sentiments en ligne se sont transformés en une sorte de profit virtuel pour les entreprises cherchant à commercialiser leurs produits, à

reconnaître les nouvelles tendances et à gérer leur position. De nombreuses organisations utilisent actuellement des systèmes d’analyse de sentiments et d'extraction d'opinions pour suivre les entrées des clients dans les sites de vente en ligne et les sites d'évaluation. L’analyse de sentiments est également utile pour les organisations pour analyser les opinions des clients sur leurs produits et fonctionnalités. Si les attributs des produits sont clairement mentionnés, la découverte de la cause principale de la faiblesse des profits nécessite de se concentrer davantage sur les sentiments individuels des clients sur ces caractéristiques. L'analyse de sentiments est une méthode étonnante pour prendre en charge de nombreuses tendances commerciales identifiées avec l'administration des transactions, la gestion du statut et la publicité. De plus, les organisations peuvent avoir la capacité d'effectuer une prédiction de modèle dans une transaction en suivant les perspectives des clients.

4 Processus d’analyse des sentiments

Les étapes du processus sentimental sont illustrées ci-dessus :

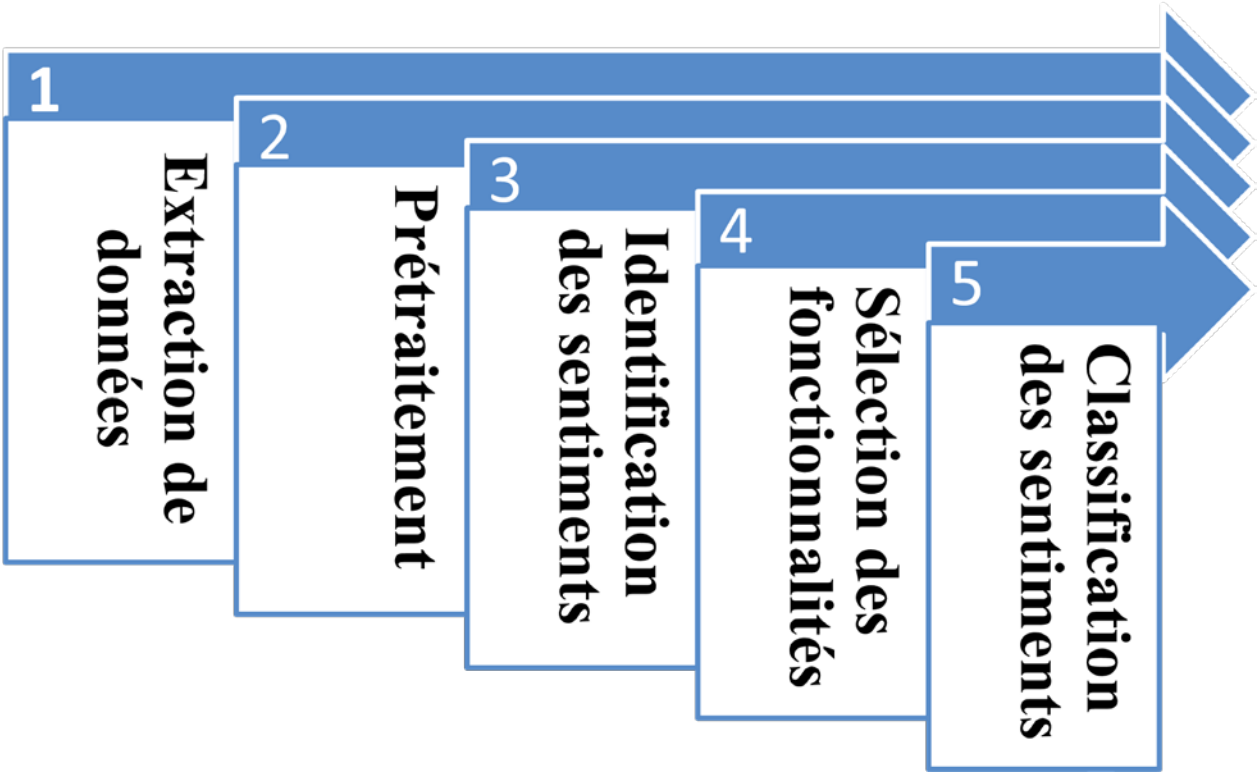


Figure 1 - Etapes de processus sentimental

Étape 1: Extraction de données

Les données sont collectées à partir des requêtes de discussions d'utilisateurs sur des forums publics tels que des blogs, des forums de discussions et des tableaux de critiques de produits, ainsi que sur des journaux privés via des sites de réseaux sociaux tels que Twitter et Facebook. Très souvent, le journal de données est volumineux, désorganisé et désintégré sur plusieurs portails stockées dans une base de données. Une fois les données extraites, elles seront ensuite préparées pour l'analyse.

Étape 2: Prétraitement

Le prétraitement est le processus de nettoyage des données préparant le texte pour la classification. Les textes en ligne contiennent généralement beaucoup de bruits et des éléments inutiles tels que des balises, des scripts. Le prétraitement des données réduit le bruit dans lequel contribue à s’améliorer les performances du classificateur. Le prétraitement accélère également le processus de classification, aidant ainsi en temps réel SA.

La préparation du texte implique le nettoyage des données extraites avant que l'analyse ne soit effectuée. Habituellement, la préparation de texte implique l'identification et l'élimination du contenu non textuel de l'ensemble de données textuelles. En outre, tout autre contenu qui n'est pas jugé pertinent pour le domaine d'étude est également supprimé de l'ensemble de données textuelles, comme par exemple des mots vides ou des mots qui ne sont pas pertinents pour le cours de l'analyse. Pour un système qui donne SA des flux de données, la stratégie de prétraitement est la suivante :

a) Suppression des mots : Certaines approches considèrent que certains mots très communs n’apportent aucune information utile pour l’analyse du texte. Dans ce cas, il est courant, d’utiliser une liste de ces mots pour filtrer le texte. La suppression de ces mots présente l’avantage de réduire la phrase à des mots pleins.

b) Structuration des phrases : Structurer une phrase permet de mieux appréhender la sémantique de chaque mot. Certaines techniques d’analyse de sentiments utilisent la structure des phrases afin d’identifier l’opinion.

c) Suppression des localisateurs de ressources uniformes (URL), hashtags, références, caractères spéciaux : Le nettoyage des données des hashtags, références, caractères spéciaux, aidera à réduire la plupart des bruits.

d) Traduction de mots d'argot - Pour cela, nous prenons l'aide du dictionnaire d'argot Internet et remplaçons les mots d'argot dans leur format significatif.

e) Suppression des lettres supplémentaires des mots : Les mots qui ont la même lettre plus de deux fois et qui ne sont pas présents dans le lexique sont réduits au mot avec la lettre répétitive n'apparaissant qu'une seule fois. Par exemple, le mot exagéré «Happyyyyyy» est réduit à «Happy».

f) ***Enracinement*** : L'enracinement donne le mot racine, est fait à l'aide de Natural Language Tool Kit (NLTK). Par exemple, des mots tels que «waiting», «waits», «waited» sont remplacés par le mot «wait».

Étape 3: Détection des sentiments

La troisième étape est la détection des sentiments. La détection des sentiments nécessite d'évaluer et d'extraire des critiques et des opinions à partir de l'ensemble de données textuelles grâce à l'utilisation de tâches de calcul. Chaque phrase est examinée pour la subjectivité. Seules les phrases avec des expressions subjectives sont conservées dans l'ensemble de données. Les phrases qui véhiculent des faits et une communication objective sont écartées de toute analyse ultérieure. La détection des sentiments se fait à différents niveaux, soit un seul terme, des phrases, des phrases complètes ou un document complet avec des techniques couramment utilisées.

Étape 4: Sélection des fonctionnalités

Comme indiqué dans [2], le but principal de la sélection de fonction est de diminuer la dimensionnalité de l'espace de fonction. Un espace de fonctionnalités réduit le coût de calcul. En tant que deuxième objectif, la sélection des fonctionnalités réduira également la sur-adaptation du schéma d'apprentissage aux données d'apprentissage. Au cours de ce processus, il est également important de trouver un bon compromis entre les richesse des fonctionnalités et des contraintes de calcul impliquées lors de la résolution de la tâche de catégorisation.

Étape 5: Classification des sentiments

La cinquième étape est la classification de la polarité qui classe chaque phrase subjective de l'ensemble de données textuelles en groupes de classification. Habituellement, ces groupes sont représentés sur deux points extrêmes d'un continuum (positif, négatif, bon, mauvais, ect).

Les techniques de classification des sentiments (SC) peuvent être divisées en deux parties, à savoir l'approche d'apprentissage automatique (Machine Learning) et l'approche basée sur le lexique. Les approches ML sont basées sur la formation d'un algorithme, principalement la classification sur un ensemble de fonctionnalités sélectionnées pour une mission spécifique, puis testées sur un autre ensemble s'il est capable de détecter les bonnes fonctionnalités et de donner les bonnes classifications. Les approches ML utilisent des algorithmes de ML en utilisant des attributs linguistiques tandis que l'approche basée sur un lexique s'inspire du «lexique des sentiments». Un état de l'art de ces approches est abordé dans le chapitre suivant.

5 Les applications de l'analyse des sentiments

La gestion de contenu axée sur l'opinion possède de nombreuses applications importantes comme la détermination des opinions des critiques concernant un certain produit via la classification des critiques de produits en ligne ou l'enregistrement de l'évolution des attitudes du public à l'égard d'un parti politique via l'extraction de sites d'actualités en ligne ou de contenu de blogs [3]. Alors que les applications basées sur l'opinion ou sur le feedback sont plus populaires, le domaine du traitement du langage naturel s'intéresse actuellement aux analyses de sentiments ainsi qu'aux systèmes d'exploration d'opinion. Les principales applications de des analyses de sentiments et de l'extraction d'opinions sont données ci-dessous:

- **Achat de produits ou de services**: lorsque vous décidez d'acheter des produits ou des services, prendre des décisions précises n'est plus un travail difficile. Grâce à cette méthode, les individus peuvent évaluer les opinions et les expériences des autres concernant tous les produits et services et comparer les marques concurrentes.
- **Amélioration de la qualité du produit ou du service**: grâce à l'exploration d'opinions et à l'analyse des sentiments, les fabricants peuvent recueillir les opinions des critiques ainsi que

les critiques positives concernant leurs produits ou services et ainsi améliorer la qualité de leurs services.

- **Recherche marketing**: les résultats des analyses de sentiment peuvent être utilisés à des fins d'étude de marché. Grâce à des méthodes d'analyse des sentiments, les tendances récentes des clients concernant des produits ou services particuliers peuvent être examinées. De même, les attitudes actuelles du public à l'égard des nouvelles politiques de l'État peuvent également être facilement examinées.
- **Systèmes de recommandation**: Grâce à la classification des opinions des individus comme positives ou négatives, le système peut déterminer laquelle est recommandée et laquelle ne l'est pas.
- **Détection de flamme**: la supervision des sites d'information, des articles de blogs ainsi que des réseaux sociaux est simplifiée grâce à des analyses de sentiment. l'exploration d'opinions et l’analyses des sentiments sont capables de détecter les mots arrogants, surchauffés, incitant à la haine, jurons utilisés dans les e-mails, les blogs ou les sites d'informations de manière automatisée.
- **Détection d’opinion Spam**: Comme Internet est également accessible à tout le monde, n'importe qui peut télécharger n'importe quoi sur Internet. Cela signifie la probabilité que le contenu soit du spam augmente de jour en jour. Les particuliers pourraient télécharger du contenu de spam dans le but d'induire les gens en erreur. L'exploration d'opinions ainsi que les analyses de sentiment sont capables de classer le contenu Internet en spam et en non spam.
- **Élaboration de politiques**: en utilisant des analyses de sentiment, les décideurs politiques sont en mesure de prendre en considération les perspectives des citoyens concernant certaines politiques et ces connaissances peuvent être utilisées pour créer de nouvelles politiques en faveur des citoyens.
- **Prise de décision**: les opinions et les expériences du public sont des facteurs très utiles lors de la prise de décisions. Une analyse OM ainsi qu'une analyse des sentiments fournissent une analyse des opinions du public qui peuvent être utilisées efficacement lors de la prise de décisions. Les sites de réseaux sociaux tels que Twitter ou Facebook ne sont pas des sources de données publiques à grande échelle à traiter légèrement. Le public les utilise pour révéler ses opinions et ses sentiments sur plusieurs sujets. L'utilisation d'analyses de sentiment sur

les avis et leur classification automatique en classes positives, négatives ou neutres peut offrir des données cruciales aux entreprises sous la forme d'études de marché.

6 Les tâches d'analyse des sentiments

Les tâches d'analyse des sentiments consistent principalement à classer la polarité d'un texte donné au niveau du document, de la phrase ou de la fonction / aspect exprimant l'opinion comme positive, négative ou neutre. L'analyse des sentiments peut être effectuée à l'un des trois niveaux: le niveau du document, le niveau de la sentence, niveau de fonctionnalité.

- ***Analyse des sentiments au niveau du document:*** le principal défi de l'analyse des sentiments au niveau du document est d'extraire du texte informatif pour déduire le sentiment de l'ensemble du document. Dans l'analyse des sentiments, un document peut être classé comme positif ou négatif ou neutre en fonction de la polarité des informations subjectives présentes dans le document. L'évaluation de la qualité de sentiment détermine si une opinion est utile ou non, et les groupes d'identification des spams divisent les sentiments en tant que spam et non spam.

- ***Analyse des sentiments au niveau de la phrase:*** Dans une classification des sentiments au niveau du phrase, la polarité de la phrase peut être donnée par trois catégories: positive, négative et neutre. Le défi auquel est confrontée la classification des sentiments au niveau des phrases réside dans les caractéristiques d'identification indiquant si les phrases sont sur le sujet, ce qui est une sorte de problème de coréférence [4]. Enfin, l’analyse de sentiments dans des phrases relatives incorpore la reconnaissance de phrases similaires et la concentration des données à partir d'elles.

- ***Analyse des sentiments au niveau de la fonctionnalité:*** les fonctionnalités du produit sont définies comme des attributs ou des composants de produit. L'analyse de ces fonctionnalités pour identifier le sentiment du document est appelée analyse de sentiment basée sur la fonctionnalité. Dans cette approche, une opinion positive ou négative est identifiée à partir des caractéristiques déjà extraites. Dans certaines, l’analyse de sentiments au niveau des fonctionnalités aide beaucoup à extraire les informations de polarité pour une fonctionnalité ou un attribut particulier d'un produit.

7 Les défis de l'analyse des sentiments

L'analyse des sentiments concerne principalement le traitement des avis, les commentaires sur différentes personnes et leur traitement pour en obtenir des informations significatives [5]. Différents facteurs affectent le processus de SA et doivent être traités correctement pour obtenir le rapport final de classification ou de regroupement. Quelques-uns de ces défis sont abordés ci-dessous:

- ***Résolution de coréférence:*** Ce problème se réfère principalement à savoir ce qu'indique un pronom ou un proverbe? » Par exemple, dans la phrase Après avoir regardé le film, nous sommes partis manger; c'était bien. » À quoi se réfère le mot «c'était »; que ce soit le film ou la nourriture? Ainsi, lorsque l'analyse du film est en cours, si la phrase concerne le film ou la nourriture? C'est une préoccupation pour l'analyste. Ce type de problème se produit principalement dans le cas d'une SA orientée aspect.

- ***Association avec une période:*** le moment de la collecte d'avis est une question importante dans le cas de l'AS. Le même utilisateur ou groupe d'utilisateurs peut donner une réponse positive pour un produit à un moment donné, et il peut y avoir un cas où il peut donner une réponse négative. C'est donc un défi pour l'analyseur de sentiment à un autre moment. Ce type de problème survient principalement dans la SA comparative.

- ***Gestion du sarcasme:*** l'utilisation de mots qui signifient le contraire de ce qu'ils informent sont surtout connus comme des mots de sarcasme. Par exemple, la phrase «Quel bon batteur il est, il marque zéro dans toutes les autres manches.» Dans ce cas, le mot positif «bon» a un sens négatif. Ces phrases sont difficiles à trouver et par conséquent, elles affectent l'analyse du sentiment.
- ***Indépendance de domaine :*** Dans le SA, les mots sont principalement utilisés comme fonction d'analyse. Mais, le sens des mots n'est pas fixé de bout en bout. Il y a peu de mots dont la signification change d'un domaine à l'autre. En dehors de cela, il existe des mots qui ont une signification opposée dans différentes situations connues sous le nom de contronyme. Ainsi, il est difficile de connaître le contexte pour lequel le mot est utilisé, car il affecte l'analyse du texte et finalement le résultat.

- ***Négations:*** Les mots négatifs présents dans un texte peuvent totalement changer le sens de la phrase dans laquelle il est présent. Ainsi, lors de l'analyse des critiques, ces mots doivent être pris en compte. Par exemple, les phrases « Ceci est un bon livre. » Et « Ce n'est pas un bon livre. » Ont une signification opposée, mais lorsque l'analyse est effectuée en utilisant un seul mot à la fois, le résultat peut être différent. Pour gérer ce type de situations, une analyse en n-gramme est préférable.

- ***Détection de spam:*** le Web contient à la fois des contenus authentiques et du spam. Pour une classification efficace des sentiments, ce contenu de spam doit être éliminé avant le traitement. Cela peut être fait en identifiant les doublons, en détectant les valeurs aberrantes et compte tenu de la réputation du critique.

- ***Mots orthographiques***: les gens utilisent des mots orthographiques pour exprimer leur excitation, leur bonheur, par exemple: le mot Sooo, Sweeettt, Haappy ou s'ils sont pressés, ils insistent sur les mots par exemple : comeeeee, fasssssst , waitttngggg…ect
- ***Manière d'exprimer son sentiment***: les gens n'expriment pas toujours leurs sentiments de la même manière. Le sentiment de chaque individu est différent la façon de penser, la manière d'exprimer varie d'une personne à l'autre.
- ***Asymétrie dans la disponibilité des logiciels d'extraction d'opinion:*** le logiciel d'extraction d'opinion est très coûteux et actuellement abordable uniquement pour les grandes organisations et le gouvernement. C'est au-delà des attentes des citoyens ordinaires. Cela doit être accessible à tous, afin que chacun en profite [6].

8 Conclusion

Dans ce premier chapitre, nous avons présenté les définitions et la terminologie utilisées dans cette thèse. Nous avons vu dans l'ordre, le processus de l'analyse de sentiments, les applications et les tâches de l'analyse de sentiments et finalement les défis rencontrés au cours de cette analyse.

Pour achever notre étude bibliographique, nous allons présenter, dans le chapitre qui suit, un état de l'art sur les différentes approches d'analyse des sentiments en mettant l'accent sur les plus importants entre eux.

Chapitre 2
Etat de l'art

Sommaire

1 Introduction

Nous nous intéressons dans ce chapitre aux travaux relatifs à l'analyse des sentiments. La classification des sentiments est l'étape essentielle de l'analyse des sentiments. Un problème de classification de texte. La classification des sentiments est un problème de classification de texte traditionnelle, qui classe principalement les documents de différents sujets, par exemple la politique, les sciences et les sports. Dans ces classifications, les mots liés au sujet sont les principales caractéristiques. Cependant, dans la classification des sentiments, les mots de sentiment ou d'opinion qui indiquent des opinions positives ou négatives sont plus importants.

La détection d'opinions est une tâche qui permet d'extraire les opinions d'un ensemble de documents pertinents pour un sujet donné. Le sentiment (ou l'opinion) peut être exprimé de manière très variée et subtile et donc il est difficile de le déterminer. La classification du sentiment (polarité) est une sous-tâche de la détection d'opinions. Elle consiste de façon générale à déterminer si l'opinion du document sur le sujet est positive ou négative. La détection d'opinions se fait au niveau du document, du paragraphe ou de la phrase. Dans ce chapitre, nous présentons les travaux relatifs à la classification des sentiments.

2 Approches de classification des sentiments

Les approches de classification des sentiments peuvent être généralement divisées en approche d'apprentissage automatique (Machine Learning), approche basée sur le lexique et approche hybride. L'approche Machine Learning (ML) applique les célèbres algorithmes ML et utilise des fonctionnalités linguistiques qui peuvent être généralement divisées en plusieurs parties, à savoir les méthodes d'apprentissage supervisé et non supervisé. Les méthodes d'apprentissage supervisé font appel à un grand nombre de documents de formation labellisés. Dans le cas où il est difficile de trouver les documents de formation étiquetés, les méthodes non supervisées sont utilisées.

L'approche basée sur le lexique repose sur un lexique des sentiments, une collection 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 basée sur un dictionnaire qui dépend de la recherche de mots de semences d'opinion, puis recherche le dictionnaire de leurs synonymes et antonymes. L'approche basée sur le corpus commence par une liste de départ 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.

L'approche hybride combine les deux techniques et est très courante, les lexiques de sentiment jouant un rôle clé dans la majorité des méthodes. Les différentes techniques et les algorithmes les plus populaires de classification des sentiments sont illustrées sur la figure suivante comme mentionné précédemment.

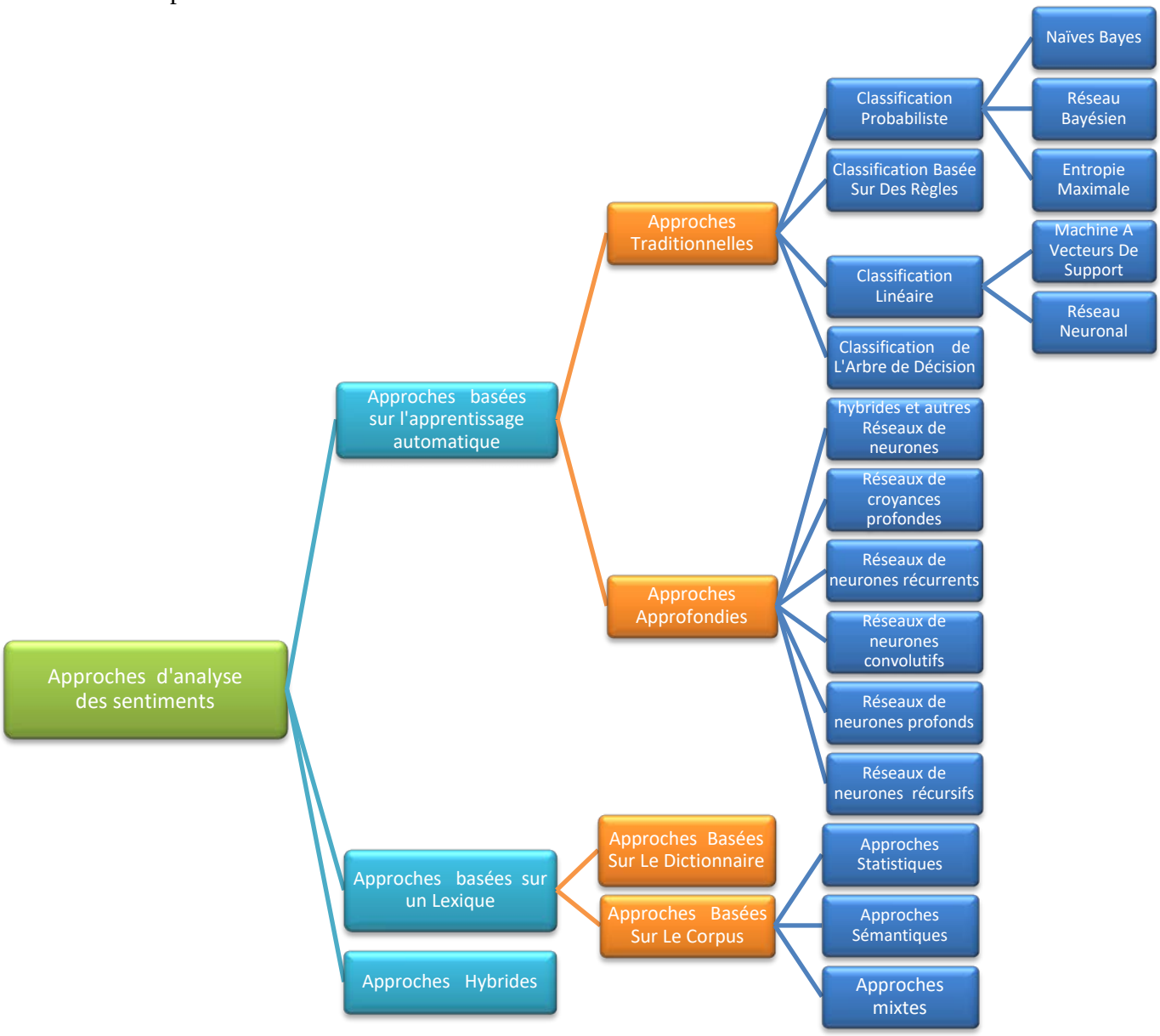


Figure 2- Approches d'analyse des sentiments

2.1 Approches d'apprentissage automatique

Les techniques d'apprentissage automatique reposent sur le fameux algorithme de machine learning (ML) pour résoudre l'analyse des sentiments en tant que problème de classification de texte régulier qui utilise des fonctionnalités syntaxiques et/ou linguistiques. Les algorithmes de machine leaning sont très souvent utiles pour classer et prédire si un document représente un sentiment positif ou négatif. Les techniques d'apprentissage automatique sont classées en deux types appelés techniques traditionnelles et approche approfondis.

Les techniques traditionnelles sont à leur tours divisées en deux sous classes, les algorithmes d'apprentissage automatique supervisés et non supervisés. L'algorithme supervisé utilise un ensemble de données étiqueté où chaque document de l'ensemble d'apprentissage est étiqueté avec le sentiment approprié. Alors que l'apprentissage non supervisé inclut un ensemble de données non étiqueté où le texte n'est pas étiqueté avec les sentiments appropriés

2.1.1 Approches traditionnelles

Ces techniques utilisent des classificateurs. Des données qui représentent des phrases subjectives (ou des documents avec opinion) sont fournies au classificateur pour l'apprentissage. Le classificateur génère un modèle, qui sera utilisé dans la partie test. Il existe de nombreux types de classificateurs dans la littérature, à savoir supervisés et non supervisés. Dans les sous-sections suivantes, nous présentons en bref quelques-uns des classificateurs les plus fréquemment utilisés dans l'analyse des sentiments.

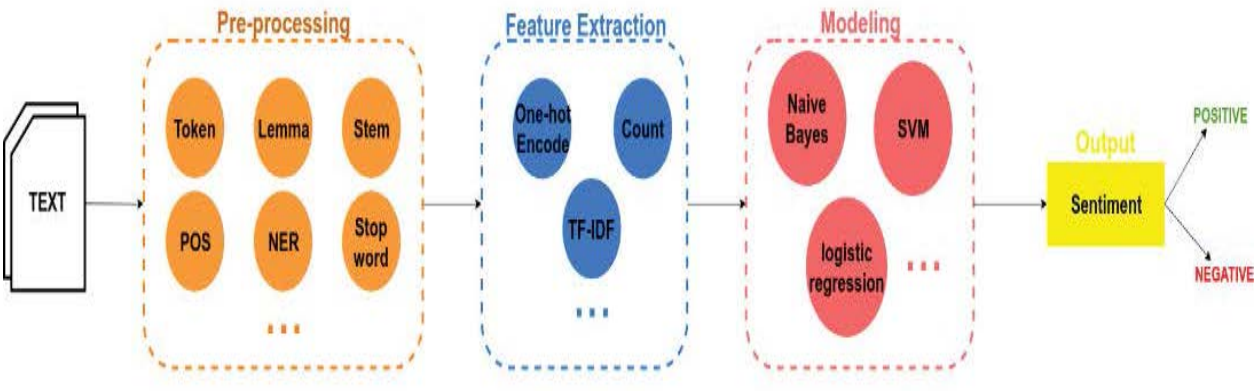


Figure 3- Modèle machine learning de classification des sentiments [7]

2.1.1.1 Classificateurs probabilistes

Les classificateurs probabilistes utilisent des modèles de mélange pour la classification. Le modèle de mélange suppose que chaque classe est un composant du mélange. Chaque composant du mélange est un modèle génératif qui fournit la probabilité d'échantillonnage d'un terme particulier pour ce composant. Ces types de classificateurs sont également appelés classificateurs génératifs. Trois des classificateurs probabilistes les plus célèbres sont discutés dans les suivantes sous-sections.

A. Classificateur Bayes naïfs (NB)

Les classificateurs bayésiens sont les classificateurs les plus simples en apprentissage supervisé basés sur le théorème de Bayes. Ils peuvent prédire la classe probabilités d'appartenance, telles que la probabilité qu'un échantillon donné appartient à une classe particulière. Les classificateurs supposent que l'effet d'une valeur d'attribut sur une classe donnée est indépendant des valeurs des autres attributs. Cette hypothèse est appelée indépendance conditionnelle de classe. Il est fait pour simplifier le calcul impliqué et, en ce sens, est considéré comme « naïf ».

Soit $X = \{x_1, x_2, \ldots, x_n\}$ être un échantillon, dont les composants représentent les valeurs faites sur un ensemble de n attributs. En termes bayésiens, X est considéré l'échantillon de données observé. Soit H une hypothèse telle que les données X appartiennent à une classe spécifique C. Pour les problèmes de classification, notre objectif est de déterminer P(H|X), la probabilité que l'hypothèse H soit vérifiée étant donné les l'échantillon de données observé X. En d'autre mots, nous recherchons la probabilité d'appartenance de l'échantillon X à la classe C, étant donné que nous connaissons la description des attributs de X.

Selon le théorème de Bayes, la probabilité que nous voulons calculer P(H|X) peut être exprimé en termes de probabilités P(H), P(X|H) et P(X) comme suit :

$$\mathbf{P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}}$$

P (H|X) représente la probabilité postérieure de H conditionnée à X, c'est-à-dire la probabilité qu'une hypothèse se vérifie étant donné la valeur de X, P(H) représente la probabilité antérieure de H, c'est-à-dire la probabilité que H est vrai indépendamment des valeurs de l'échantillon, P(X|H) représente la probabilité postérieure de X conditionnée à H, c'est-à-dire la probabilité que X aura certaines valeurs pour une hypothèse donnée , P(X) représente la probabilité antérieure de X, c'est-à-dire la probabilité que X aura certaines valeurs.

Beineke et al. [8], ont utilisé le modèle Naïve Bayes pour la classification des sentiments. Ils ont extrait une paire de caractéristiques dérivées qui sont linéairement combinables pour prédire le sentiment. Pour améliorer le niveau de précision, ils ont ajouté des caractéristiques dérivées supplémentaires au modèle et utilisé des données étiquetées pour estimer l'influence relative. Parallèlement à cela, ils ont également utilisé le concept de mots d'ancrage, c'est-à-dire les mots à sens multiple pour l'analyse. Ils ont considéré cinq mots d'ancrage positifs et cinq mots d'ancrage négatifs qui, après combinaison, produisent 25 paires possibles pour l'analyse. Ils ont suivi l'approche de Turney [9], qui génère effectivement un nouveau corpus de document d'étiquette à partir du document existant.

L'article [10] présente une méthode d'analyse des sentiments, sur l'examen effectué par les utilisateurs de films. La classification des avis dans les classes positives et négatives est basée sur un algorithme naïf de Bayes. Comme données d'entrainements, ils ont utilisés une collection (pré-classée en positif et négatif) de phrases extraites des critiques de films. Pour améliorer la classification, ils ont supprimé les mots insignifiants et introduit dans les groupes de classification des mots (n-grammes). Une amélioration substantielle de la classification pour n = 2 groupes.

Le système proposé par [11] extrait des aspects de commentaires des clients sur les produits. Les noms et les phrases nominales sont extraits de chaque phrase de révision. Le seuil de support minimum est utilisé pour trouver tous les aspects fréquents des phrases de révision données. Un algorithme bayésien naïf utilisant une approche basée sur le comptage supervisé des termes est utilisé pour identifier si la phrase est une opinion positive ou négative et en identifier également le nombre.

B. Classificateur réseau Bayésien (RB)

Les réseaux Bayésiens constituent un ensemble de méthodes statistiques utilisées pour modéliser des problèmes, extraire de l'information et prendre des décisions. Ils sont un formalisme de raisonnement probabiliste utilisé dans plusieurs domaines tels que l'industrie, la santé, finance et le traitement d'images. L'hypothèse principale du classificateur réseau Bayésien est l'indépendance des caractéristiques. L'autre hypothèse extrême est de supposer que toutes les fonctionnalités sont entièrement dépendantes. Cela conduit au modèle de réseau bayésien qui est un graphe acyclique dirigé dont les nœuds représentent des variables aléatoires et les arêtes

représentent des dépendances conditionnelles. Les réseaux bayésiens ont été largement utilisés dans de nombreuses applications de fouille de texte, comme le filtrage du spam et la récupération d'informations, il est considéré comme un modèle complet pour les variables et leurs relations.

Un problème des classificateurs RB est qu'ils ne conviennent pas aux ensembles de données avec de nombreuses caractéristiques [12]. La raison en est qu'essayer de construire un très grand réseau n'est tout simplement pas faisable en termes de temps et d'espace. Un dernier problème est qu'avant l'induction, les caractéristiques numériques doivent être discrétisées dans la plupart des cas. Dans le texte mining, cette complexité de calcul de RB est très coûteuse; c'est pourquoi, il n'est pas fréquemment utilisé.

Dans [13]**,** une approche bayésienne est proposée pour identifier les spams en utilisant un classificateur Bayes naïf. L'intuition est que certains mots ont des probabilités particulières de se produire dans les courriers indésirables et dans les instances de courriels légitimes, les mots «gratuit» et «crédit» apparaîtront fréquemment dans les courriers indésirables, mais se produiront rarement dans d'autres courriels. Pour former le filtre, l'utilisateur doit indiquer manuellement si un e-mail est du spam ou non pour un ensemble de formation. Avec un tel ensemble de données de formation, les filtres anti-spam bayésiens apprendront une probabilité de spam pour chaque mot, par exemple, une forte probabilité de spam pour les mots "gratuit" et "crédit", et une probabilité de spam relativement faible pour des mots tels que les noms d'amis. Ensuite, la probabilité de spam de l'e-mail est calculée sur tous les mots de l'e-mail, et si le total dépasse un certain seuil, le filtre marquera l'e-mail comme spam.

Les auteurs de [14] ont utilisé les réseaux Bayésiens pour considérer un problème du monde réel dans lequel l'attitude de l'auteur est caractérisée par trois variables cibles différentes (mais liées). Ils ont proposé l'utilisation de classificateurs de réseaux bayésiens multidimensionnels. Il a rejoint les différentes variables cibles dans la même tâche de classification afin d'exploiter les relations potentielles entre eux. Ils ont étendu le cadre de classification multidimensionnelle au domaine semi-supervisé afin de tirer parti de l'énorme quantité d'informations non étiquetées disponibles dans ce contexte. Ils ont montré que leur approche multidimensionnelle semi-supervisée dépasse les techniques d'analyse de sentiments les plus courantes et que leur classificateur est la meilleure solution dans un cadre semi-supervisé car il correspond à la structure de domaine sous-jacente réelle.

C. Classificateur d'Entropie Maximale (EM)

Le classificateur d'entropie maximale (appelé classificateur exponentiel conditionnel) convertit les ensembles d'entités étiquetés en vecteurs à l'aide du codage. Ce vecteur codé est ensuite utilisé pour calculer les poids de chaque entité qui peuvent ensuite être combinés pour déterminer l'étiquette la plus probable pour un ensemble d'entités. L'entropie maximale maximise l'entropie définie dans la distribution de probabilité conditionnelle. Il traite de la même manière décrite dans l'algorithme naïf de Bayes.

Le classificateur d'Entropie Maximale a été utilisé par Kaufmann [15] pour détecter des phrases parallèles entre n'importe quelle paire de langues avec de petites quantités de données d'apprentissage. Les autres outils développés pour extraire automatiquement des données parallèles de corpus non parallèles utilisent des techniques spécifiques au langage ou nécessitent de grandes quantités de données de formation. Leurs résultats ont montré que les classificateurs ME peuvent produire des résultats utiles pour presque toutes les paires de langues. Cela peut permettre la création de corpus parallèles pour de nombreuses nouvelles langues.

Dans l'article [16], une nouvelle méthode est utilisée pour collecter divers messages Twitter des apprenants. Sur ce jeu de données, un prétraitement pour l'analyse des sentiments est effectué. Il implique diverses opérations intermédiaires pour lever l'ambiguïté. L'ensemble de données pré-traité est utilisé pour construire la classification de l'état émotionnel de l'utilisateur et les classificateurs SVM, ME et bayésiens naïfs sont appliqués et les résultats sont très efficaces.

2.1.1.2 Classificateurs linéaires

Les classificateurs linéaires font partie des méthodes de classification les plus pratiques. Dans l'analyse des sentiments, le classificateur linéaire associe un coefficient au décompte de chaque mot de la phrase, et tente de déterminer de bons séparateurs linéaires (hyperplan) entre les différentes classes. Les classificateurs linéaires sont ceux pour lesquels la sortie du prédicteur linéaire est défini comme étant $p = A \cdot X + b$, où $X = (x_1, x_2, \ldots, x_n)$ est la valeur normalisée document vecteur de fréquence de mot, $A = (a_1, a_2, \ldots, a_n)$ est un vecteur de coefficients linéaires avec la même dimensionnalité que l'espace d'entité, et b est un scalaire. Une interprétation naturelle du prédicteur $p = A \cdot X + b$ dans le scénario discret (étiquettes de classe catégorielles) serait une séparation hyperplan entre les différentes classes. Deux des classificateurs linéaires les plus célèbres sont décrits dans la suite:

A. Classificateur de machines à vecteurs de support (SVM)

Les SVMs sont les classificateurs d'apprentissage automatique largement utilisé pour la catégorisation de texte, qui ont été proposés pour la première fois par Vapnik en 1990. Les SVMs sont des techniques d'apprentissage supervisé qui reposent sur deux notions principales : la notion de marge maximale et la notion de fonction noyau [20]. L'objectif principal des SVM est de déterminer dans l'espace de recherche des séparateurs appelés hyperplans, qui peuvent séparer au mieux les différentes classes.

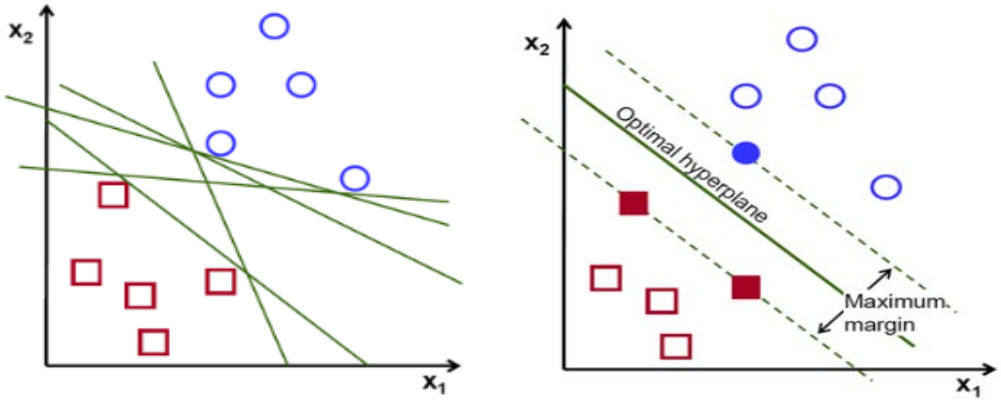


Figure 4- Exemple d'un SVM

Dans le cas où le problème est linéairement séparable, le choix de l'hyperplan séparateur n'est pas évident. Il existe en effet une infinité d'hyperplans capables de séparer parfaitement les deux classes d'exemples. Le principe des SVMs est de choisir celui qui va maximiser la distance minimale entre l'hyperplan et les exemples d'apprentissage. Vapnik a montré dans [17] qu'il existe un unique hyperplan optimal, défini comme l'hyperplan qui maximise la marge (la marge est la distance entre la frontière de séparation et les échantillons les plus proches) entre les échantillons et l'hyperplan séparateur.

Afin de pouvoir traiter des cas où les données ne sont pas linéairement séparables, l'idée clé des SVMs est de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension, dans lequel il est probable qu'il existe une séparation linéaire grâce à des fonctions noyaux qui permettent de transformer un produit scalaire dans un espace de grande dimension.

Les SVMs sont utilisés dans de nombreuses applications, parmi lesquelles celle de [18], qui a effectué une analyse des sentiments sur les Tweets à l'aide d'un SVM et a atteint une précision d'environ 60% en fonction des fonctionnalités utilisées. De plus, ils ont prétraité les Tweets en

remplaçant les acronymes par leur pleine signification et en remplaçant les émoticônes par leur état émotionnel. Il convient de noter que la même méthode a atteint des précisions d'environ 75% lors de la conduite de la classification binaire, ignorant la classe neutre et n'ayant que des classes positives ou négatives.

Les auteurs de [19] ont appliqué l'algorithme SVM pour l'analyse des sentiments où des valeurs sont attribuées à quelques mots sélectionnés, puis les ont combinées pour former un modèle de classification. Parallèlement à cela, différentes classes d'entités ayant une proximité avec le sujet sont affectées aux valeurs favorables, qui aident à la classification. Les auteurs ont présenté une comparaison de leur approche proposée avec les données, ayant annotation de sujet et annotation manuelle. Leur méthode proposée a montré un meilleur résultat par rapport à celui de l'annotation de sujet alors que les résultats ont besoin d'être améliorés lors de la comparaison avec des données annotées à la main.

Dans [21]**,** les auteurs ont utilisé deux techniques basées sur SVM multiclasse: SVM un contre tous (SVM One-versus-All) et SVM multiclasse mono-machine (Single-Machine Multiclass SVM) pour catégoriser les revues. Ils ont proposé une méthode pour évaluer la qualité des informations dans les revues de produits en le considérant comme un problème de classification. Ils ont également adopté un cadre de qualité de l'information (QI) pour trouver un ensemble de fonctionnalités orientées information. Ils ont travaillé sur des appareils photo numériques et Revues MP3. Leurs résultats ont montré que leur méthode peut classer les avis en fonction de leur qualité.

Li et Li [22] utilisent les SVMs comme classificateurs de polarité des sentiments. Contrairement au problème de classification binaire, ils ont argumenté que la subjectivité de l'opinion et la crédibilité de l'exprimeur devraient également être prises en considération. Ils ont proposé un cadre qui fournit un résumé numérique compact des opinions sur les plateformes de micro-blogs. Ils ont identifié et extrait les sujets mentionnés dans les opinions associées aux requêtes des utilisateurs, puis classé les opinions à l'aide de SVMs. Ils ont travaillé sur des publications Twitter pour leur expérience. Ils ont découvert que la prise en compte de la crédibilité des utilisateurs et de la subjectivité des opinions est essentielle pour agréger les opinions des micro-blogs. Ils ont prouvé que leur mécanisme peut effectivement découvrir l'intelligence de marché (MI) pour aider les décideurs en établissant un système de suivi pour suivre les opinions externes sur différents aspects d'une entreprise en temps réel.

Balahur dans [23] a atteint une précision de 85% en utilisant un SVM qui avait été formé avec les données des réseaux sociaux. Il a utilisé les méthodes suivantes pour prétraiter les Tweets:

- Supprimer la ponctuation répétée- Les textes informels écrits sur les réseaux sociaux contiennent souvent plusieurs symboles de ponctuation, par ex. '!!!' ou '??'. Ceux-ci ont été normalisés en une seule occurrence du symbole de ponctuation, par ex. "!" et "?".
- Remplacement d'émoticônes - En comparant l'émoticône trouvé dans le Tweet avec le dictionnaire d'émotions, Balahur a remplacé les émoticônes véhiculant un sentiment positif par le mot "positif" et les émoticônes véhiculant un sentiment négatif avec "négatif". Les émoticônes considérés comme neutres ont été supprimées
- Remplacement de l'argot - Le remplacement de l'argot est effectué dans le but de normaliser la langue utilisée dans le texte. Cela a été fait en utilisant un site spécialisé avec des remplacements pour les mots d'argot.
- Normalisation des mots - Les textes dans les médias sociaux contiennent souvent des mots qui ont été accentués en répétant certaines des lettres du mot. Par exemple, interroger Twitter pour le mot «haaate» génère une grande quantité de résultats. Afin de faire face à cela, Balahur a vérifié l'existence du mot dans un dictionnaire, et si aucun n'a été trouvé, les lettres accentuées ont été supprimées jusqu'à ce qu'un mot soit trouvé. Par exemple, «haaate» deviendrait «haate» et enfin «hate».

Dans [140], les SVMs combinés avec des lexiques spécifiques à un domaine sont implémentés pour la classification des aspects et l'identification de la polarité de revue du produit. SVM est formé pour modéliser la classification d'aspect et ce SVM formé est utilisé pour la classification de polarité par aspect. Les résultats expérimentaux indiquent que les techniques proposées ont atteint une précision d'environ 78%. Les données Web sont appliquées au sous-système d'extraction des causes émotionnelles et la méthode de sélection des fonctionnalités complémentaires, basée sur la sortie de ces fonctionnalités, est fusionnée.

B. Classificateurs de réseau neuronal (RN)

Les réseaux de neurones fabriquées de structures cellulaires artificielles, constituent une approche permettant d'aborder sous des angles nouveaux les problèmes de perception, de mémoire, d'apprentissage et de raisonnement (en d'autres termes l'intelligence artificielle I.A.).

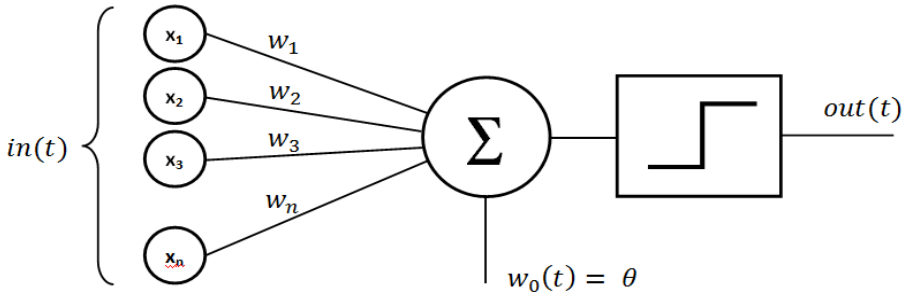


Figure 5- Représentation graphique d'un perceptron avec une seule sortie [24]

Ils s'avèrent aussi des alternatives très prometteuses pour contourner certaines des limitations des méthodes numériques classiques. Grâce à leur traitement parallèle de l'information et à leurs mécanismes inspirés des cellules nerveuses (neurones), ils infèrent des propriétés émergentes permettant de solutionner des problèmes jadis qualifiés de complexes. Les réseaux de neurones artificiels fondés sur des modèles qui tentent de simuler les cellules du cerveau humain et leurs interconnexions. Le but, est d'exécuter des calculs complexes et de trouver, par apprentissage, une relation non linéaire entre des données numériques et des paramètres.

Les réseaux de neurones multicouche sont utilisés pour les problèmes non linéaires. Ces couches multiples sont utilisées pour induire plusieurs problèmes linéaires par morceaux, et sont utilisées pour approximer les régions fermées appartenant à une classe particulière. Les sorties des neurones des couches précédentes alimentent les neurones des couches ultérieures. Le processus d'apprentissage est plus complexe car les erreurs doivent être retransmises sur différentes couches.

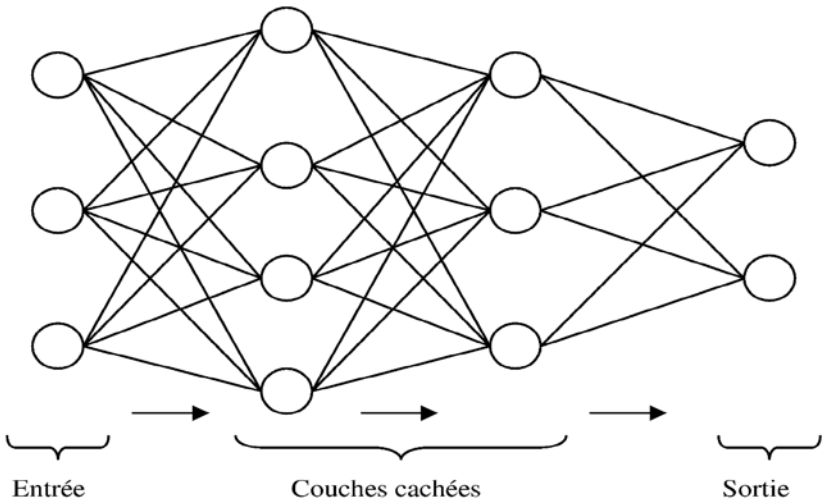


Figure 6- Représentation graphique d'un réseau de neurones multicouche [26]

Les premières méthodes de travail neuronal pour l'apprentissage en ligne ont été proposées dans [27]. Dans ces méthodes, le classificateur commence par définir tous les poids du réseau neuronal à la même valeur. L'exemple d'entraînement entrant est classé avec le réseau neuronal. Dans le cas où le résultat du processus de classification est correct, les poids ne sont pas modifiés. D'un autre côté, si la classification est incorrecte, les pondérations des termes sont augmentées ou diminuées selon la classe à laquelle appartient l'exemple de formation. Plus précisément, si la classe à laquelle appartient l'exemple de formation est une instance positive, les poids des termes correspondants (dans le document de formation) sont augmentés de α. Sinon, les poids de ces termes sont réduits de α. La valeur de α est également connue sous le nom de taux d'apprentissage.

De nombreuses autres variantes sont possibles en termes de modification des poids. Par exemple, la méthode de [28] utilise une règle de mise à jour multiplicative, dans laquelle deux constantes multiplicatives $\alpha_1 > 1$ et $\alpha_2 < 1$ sont utilisées pour le processus de classification. Les poids sont multipliés par α1, lorsque l'exemple appartient à la classe positive, et sont multipliés par α_2 autrement.

Freund et Schapire [29] ont créé un algorithme appelé, vote-perceptron, qui stocke plus d'informations pendant la formation et utilise ensuite ces informations élaborées pour générer de meilleures prédictions sur les données de test. Les informations qu'il conserve pendant l'entraînement sont la liste de tous les vecteurs de prédiction qui ont été générés après chaque erreur. Pour chacun de ces vecteurs, il compte le nombre d'itérations qu'il «survit» jusqu'à ce que l'erreur suivante soit commise; les auteures appellent ce compte le «poids» du vecteur de prédiction. Pour calculer une prédiction, l'algorithme calcule la prédiction binaire de chacun des vecteurs de prédiction et combine toutes ces prédictions au moyen d'un vote majoritaire pondéré.

Dans [30], la classification des sentiments est effectuée à l'aide du réseau neuronal à propagation arrière (BPN). Cette approche utilise le gain d'information, le BPN et la connaissance de la subjectivité présents dans le lexique des sentiments. Le résultat a montré une dimensionnalité et une précision réduites dans la classification effectuée sur les films et les critiques d'hôtels. Zhu Jian a proposé un modèle individuel basé sur des réseaux de neurones artificiels pour diviser le corpus de critiques de films en tons positifs, négatifs et flous, basé sur l'algorithme avancé de formation à la propagation des moindres carrés récursifs.

Duncan et Zhang [31] ont utilisé un réseau de neurones à action directe pour effectuer une analyse des sentiments sur les tweets collectés à l'aide de l'API Twitter. La mémoire devient un problème lors de la formation du réseau de modèles à action directe si le vocabulaire devient trop volumineux.

2.1.1.4 Classificateurs basés sur des règles

Dans les classificateurs basés sur des règles, l'espace de données est modélisé avec un ensemble de règles. Le côté gauche représente une condition sur l'ensemble des fonctionnalités exprimée sous forme normale disjonctive tandis que le côté droit est l'étiquette de classe. Les conditions sont sur le terme présence. L'absence de terme est rarement utilisée car elle n'est pas informative dans les données rares. Il existe un certain nombre de critères pour générer des règles, la phase de formation construit toutes les règles en fonction de ces critères. Les deux critères les plus courants sont le soutien et la confiance [32]. La prise en charge est le nombre absolu d'instances dans l'ensemble de données de formation qui sont pertinentes pour la règle. La confiance fait référence à la probabilité conditionnelle que le côté droit de la règle soit satisfait si le côté gauche est satisfait.

Zhang et al. ont proposé une approche fondée sur des règles pour le classement des commentaires [33]. Leur approche comprend deux phases, à savoir l'analyse du sentiment de la phrase et l'agrégation du sentiment du document. Ils décomposent le document en ses phrases constitutives et découvrent la polarité de chaque phrase. Ensuite, le score de polarité de toutes les phrases est combiné pour calculer la polarité globale du document. Ils ont considéré l'ensemble de données Euthanasia composé de 851 articles chinois et l'ensemble de données AmazonCN composé de 458522 critiques de six catégories différentes, à savoir livres, musique, film, appareil électrique, produit numérique et appareil photo. Ils ont utilisé les techniques SVM, NB et Arbre de décision pour classer les évaluations.

Kai et al. [34] proposent une approche basée sur des règles pour la détection des composants de cause d'émotion pour les micro-blogs chinois. Il présente le modèle émotionnel et extrait les composants de cause correspondants dans les émotions à grain fin. Le lexique émotionnel peut être construit manuellement et automatiquement à partir du corpus. Pendant ce temps, les proportions des composantes de cause peuvent être calculées dans l'influence des caractéristiques

multilingues basées sur la probabilité bayésienne. Les résultats de l'expérience montrent la faisabilité de l'approche.

Dans [35], une approche basée sur des règles est utilisée en définissant diverses règles pour obtenir l'opinion ou elle utilise des règles lexicales, créées en jetant chaque phrase dans chaque document, puis en testant chaque jeton ou mot pour sa présence. Si le mot est là et a un sentiment positif, une note +1 lui a été appliquée. Chaque message commence par un score neutre de zéro et était considéré comme positif si le score de polarité final était supérieur à zéro, ou négatif si le score global était inférieur à zéro. En cela, certaines règles doivent se former et ensuite les sentiments doivent être analysés en fonction de cela. Le résultat de l'approche basée sur les règles crée les règles en prenant:

a) après la sortie de l'approche basée sur des règles, il vérifiera ou demandera si la sortie est correcte ou non. Si la phrase d'entrée contient un mot qui n'est pas présent dans la base de données, ce qui peut aider à l'analyse de la critique de film, alors ces mots doivent être ajoutés à la base de données.

b) il s'agit d'un apprentissage supervisé dans lequel le système est formé pour apprendre si de nouvelles entrées sont fournies.

c) cette approche augmentera toujours l'efficacité du système.

2.1.1.4 Classificateur d'arbre de décision

Parmi les algorithmes de classification, l'un des plus simples d'utilisation et d'interprétation, tout en gardant des performances très respectables, est l'arbre de décision. Existant sous plusieurs formes, l'arbre de décision est reconnu par le résultat de l'algorithme qui produit un modèle constitué d'un ensemble de règles de classification qu'il est possible de représenter sous forme d'arbre. Le classificateur 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. La condition ou le prédicat est la présence ou l'absence d'un ou plusieurs mots. La division de l'espace de données se fait récursivement jusqu'à ce que les nœuds terminaux contiennent un nombre minimal d'enregistrements qui sont utilisés à des fins de classification.

Les auteurs de [36] proposent d'explorer les variations émotionnelles de l'adolescence et les raisons de ces changements à l'aide de techniques d'exploration de données. En classant les

émotions et en utilisant l'arbre de décision, différentes variations émotionnelles sont analysées. Les règles **if-then** sont également générées à partir de l'arbre de décision. L'analyse des valeurs aberrantes est utilisée pour identifier la variation émotionnelle chez l'enfant ayant tout type de handicap.

Dans [37], les fonctionnalités de révision de film obtenues à partir d'IMDB ont été extraites en utilisant la fréquence inverse du document et l'importance du mot trouvé. L'analyse des composants principaux et CART a été utilisée pour la sélection des fonctionnalités en fonction de l'importance du travail par rapport à l'ensemble du document. La précision de classification obtenue par LVQ était de 75%.

Les arbres de décision et les règles de décision ont tendance à coder les règles sur l'espace des fonctionnalités, mais l'arbre de décision à tendance à atteindre cet objectif avec une approche hiérarchique. Quinlan [38] a étudié les problèmes d'arbre de décision et de règle de décision dans un cadre unique; car un certain chemin dans l'arbre de décision peut être considéré comme une règle pour la classification de l'instance de texte. La principale différence entre les arbres de décision et les règles de décision est que les arbres de décision est un partitionnement hiérarchique strict de l'espace de données, tandis que les classificateurs basés sur des règles permettent des chevauchements dans l'espace de décision.

2.1.2 Approches d'apprentissage en profondeur

L'apprentissage en profondeur ou Deep learning (DL) est une branche émergente des algorithmes d'apprentissage automatique, qui s'inspire des réseaux de neurones artificiels. Il offre des moyens d'apprendre les représentations de données de manière supervisée et non supervisée à l'aide de la hiérarchie des couches, qui permettent un traitement multiple et fourni les meilleures solutions à de nombreux problèmes dans les domaines de la reconnaissance d'image et de la parole, ainsi que dans le traitement du langage naturel.

L'apprentissage en profondeur est une méthodologie d'apprentissage et de représentation à plusieurs niveaux, obtenue en composant des modules plus simples mais non linéaires (NL) où chacun transmute la représentation en un seul niveau (à partir de l'entrée brute) en une représentation dans un niveau abstrait supérieur. Avec la compilation de ces transmutations adéquates, des fonctions exceptionnellement complexes sont apprises.

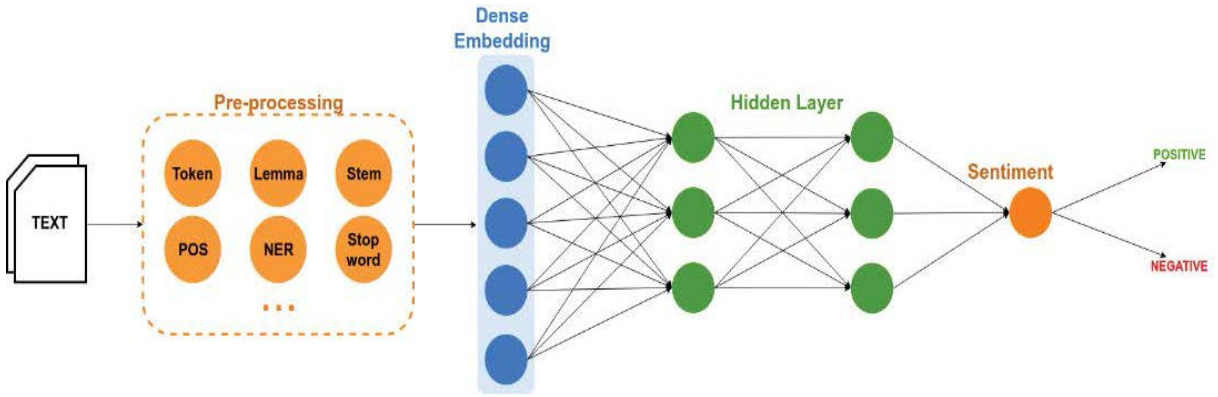


Figure 7- Modèle deep learning de classification de sentiments [7]

Récemment, les algorithmes d'apprentissage en profondeur ont fourni des performances impressionnantes dans les applications de traitement du langage naturel (NLP) englobant l'analyse des sentiments (SA) dans de nombreux ensembles de données. De tels modèles n'ont pas besoin de fonctionnalités prédéfinies qui sont sélectionnées par un ingénieur, mais ils pourraient apprendre eux-mêmes des fonctionnalités sophistiquées à partir de l'ensemble de données. Bien que chaque unité de ces réseaux neuronaux (NN) soit assez simple, grâce à l'empilement de couches d'unités à l'arrière les unes des autres, ces modèles sont compétents pour apprendre des limites de décision hautement sophistiquées. Les mots sont signifiés dans un espace vectoriel de grande dimension, et l'extorsion d'entité est laissée au NN. En conséquence, ces modèles pourraient mapper des mots ayant des propriétés syntaxiques et sémantiques identiques à des emplacements adjacents dans leur système de coordonnées, d'une manière qui évoque la compréhension de la signification des mots.

Les algorithmes basés sur les réseaux de neurones profonds ont atteint la performance de pointe dans la classification des sentiments de nos jours. Plusieurs types de techniques d'apprentissage en profondeur sont abordés dans cette section.

2.1.2.1 Réseaux de neurones profonds (DNN)

Un réseau neuronal profond est un réseau neuronal avec plus de deux couches, dont certaines sont des couches cachées (Figure 8). Les réseaux de neurones profonds utilisent une modélisation mathématique sophistiquée pour traiter les données de différentes manières. Un réseau neuronal est un modèle ajustable de sorties en fonction des entrées, qui se compose de plusieurs couches: une couche d'entrée, y compris les données d'entrée; couches cachées, y

compris les nœuds de traitement appelés neurones; et une couche de sortie, comprenant un ou plusieurs neurones, dont les sorties sont les sorties du réseau. Cette architecture était acceptable pour trouver un certain nombre de problèmes, mais le taux d'erreur était encore assez élevé. Ainsi, une architecture profonde des réseaux de neurones a été développée. Elle a une couche d'entrée, plusieurs couches cachées et une couche de sortie. Cette architecture a été développée pour améliorer la précision, mais au prix de la puissance et de son application elle n'a pas été possible jusqu'à ce que les GPUs modernes soient venus pour améliorer l'efficacité. La précision des réseaux de neurones s'est avéré augmentée à mesure que le nombre de couches cachées augmentait. En d'autres termes, à mesure que le réseau neuronal devenait «plus profond» en termes d'architecture, il fonctionnait de mieux en mieux. Cependant, ce n'est qu'un seul facteur qui améliore la précision. D'autres facteurs incluent peu de pools de moyenne et de progrès, etc.

Les méthodes basées sur DNN ont considérablement amélioré les performances de l'algorithme de classification des sentiments. De plus, les méthodes basées sur DNN simplifient également le processus d'exécution de la tâche de classification des sentiments, qui intègre les étapes d'extraction et de classification des fonctionnalités dans un traitement de bout en bout.

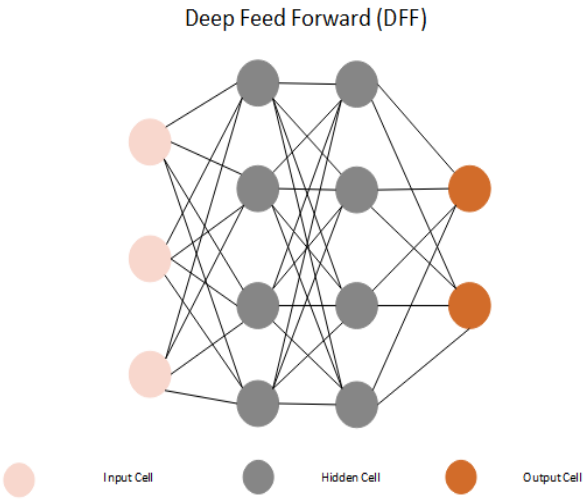


Figure 8- Réseau de neurone profond [140]

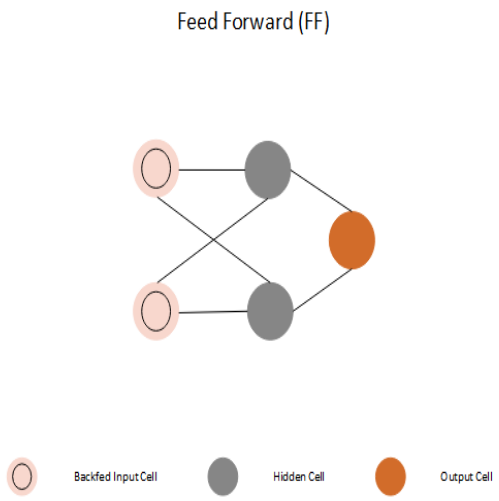


Figure 9- Réseau de neurone simple [140]

Une version non supervisée du DNN est introduite par [39] est l'auto-encodeur (AE). Lors de l'apprentissage de ce réseau, la supervision est réalisée avec les informations d'entrée. Il apprend à reconstruire le vecteur d'entrée en passant par une couche cachée de taille inférieure. Ils sont principalement utilisés dans deux cas : le premier à des fins de débruitage et de réduction de dimensions et le second comme préapprentissage des poids des couches cachées de DNN.

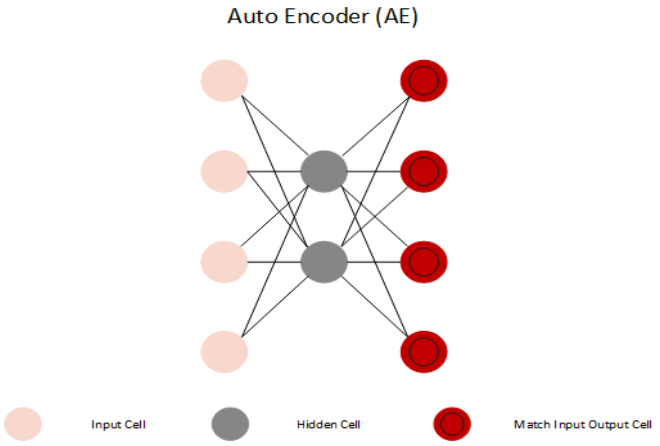


Figure 10- Auto-encodeur (AE) [140]

Le pré-entrainement d'un réseau profond à l'aide d'auto-encodeur réduit les possibilités de converger vers un optimum local lors de l'étape d'apprentissage globale.

Un type spécial de l'auto-encodeur est l'auto-encodeur variationnel (VAE) [40] qui est introduit comme un lien entre l'inférence variationnelle et les DNN.

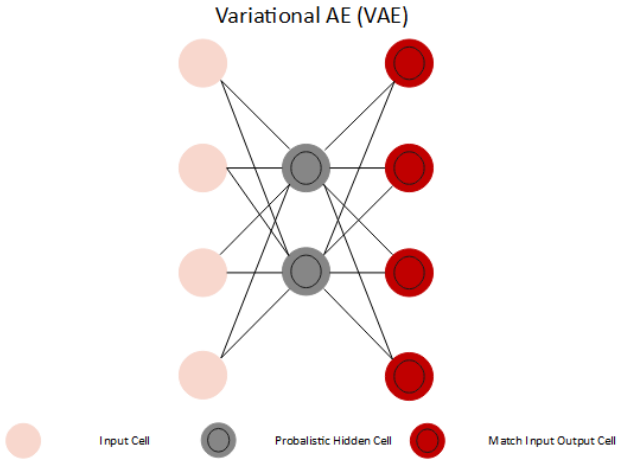


Figure 11- Auto-encodeur variationnel (VAE) [140]

Dans cette étude [41], une architecture de réseau neuronal profond a été proposée pour évaluer la similitude des documents. L'architecture a été formée en utilisant plusieurs nouvelles du marché pour produire des articles de vecteurs ennemis. Les actualités T&C ont été utilisées comme ensemble de données. La similarité cosinus a été calculée parmi les articles étiquetés et la polarité des documents a été prise en compte, mais le contenu n'a pas été pris en compte. La méthode proposée a permis d'obtenir des performances supérieures en termes d'estimation de similitude des articles en fonction de la polarité.

Dans cette étude [42], l'auteur a proposé un modèle d'analyse des sentiments prenant en compte les contenus visuels et textuels des réseaux sociaux. Ce nouveau schéma a utilisé un modèle de réseau de neurones profond tel que les encodeurs automatiques de débruitage (DAE) . La base du schéma était le modèle CBOW (Continuous Bag-Of-Words). Le modèle proposé comprenait deux parties CBOW-LR (régression logistique) pour le contenu textuel et a été développé sous la forme CBOW-DA-LR. La classification a été effectuée en fonction de la polarité des informations visuelles et textuelles. Quatre jeux de données ont été évalués, à savoir, Sanders Corpus, Sentiment140, SemEval-2013 et SentiBank Twitter. Le modèle proposé a surpassé le CBOWS + SVM et le FSLM (modèle de langage probabiliste entièrement supervisé). Peut-être que l'ESLAM (modèle de langage probabiliste entièrement supervisé étendu) en termes de petites données de formation avait surpassé le modèle actuel. L'apprentissage des fonctionnalités et les sauts de grammes nécessitaient de grands ensembles de données pour une meilleures performances.

Abburi et al. [43] ont présenté un schéma pour repérer les critiques sentimentales des produits en ligne en Hindi centrées sur ses multiples natures de modalité (texte avec audio). Pour chaque entrée audio, les caractéristiques des «coefficients cepstraux de fréquence de Mel» (MFCC) ont été extorquées. Ces fonctionnalités ont été utilisées pour construire une conception de sentiment en utilisant les classificateurs DNN et GMM (Gaussian Mixture Models). D'après les résultats, il a été perçu que le classificateur DNN offrait de meilleurs résultats contrairement au GMM. D'autres caractéristiques du texte ont été extorquées de la transcription de l'entrée audio en utilisant des vecteurs Doc2vec. D'après les résultats expérientiels, il a été perçu que l'intégration du texte et des fonctionnalités audio a amélioré la performance pour repérer le sentiment des critiques des produits en ligne.

Dans [44], les auteurs ont suggéré une structure d'extension de contenu (c'est-à-dire), intégrant des articles et des commentaires connectés à une conversation de microblog destinée à l'extorsion de fonctionnalités. Un auto-encodeur à convolution a été utilisé qui pouvait extorquer des informations contextuelles lors de conversations de microblog qui étaient utilisées comme fonctionnalités destinées aux articles. Un DNN personnalisé, qui a été intégré à de nombreuses couches de RBM (Machine Boltzmann restreinte), a été exécuté pour initialiser la structure NN. Les couches RBM pourraient prendre des échantillons de distribution de probabilité des données entrées pour apprendre des structures cachées pour une représentation fine des caractéristiques de niveau supérieur. Une couche de classe RBM (Classification RBM) a été utilisée pour atteindre l'étiquette de classification sentimentale finale destinée aux messages. Les résultats expérientiels ont montré qu'avec des paramètres et des structures appropriés, la performance du DNN suggéré sur la classification sentimentale était meilleure si l'on considérait les modèles d'apprentissage comme NB ou SVM, ce qui confirmait que le modèle DNN suggéré était pertinent pour une classification de document plus courte avec la fonctionnalité suggérée.

2.1.2.2 Réseaux de neurones convolutifs (CNN)

Un réseau neuronal convolutif est un type spécial de réseau neuronal de propagation vers l'avant (feed-forward) introduits en 1989 et utilisé à l'origine dans des domaines tels que la vision par ordinateur, les systèmes de recommandation et le traitement du langage naturel. En principe, CNN a trois types de couches, à savoir la couche d'entrée, les couches d'extraction d'entités et la

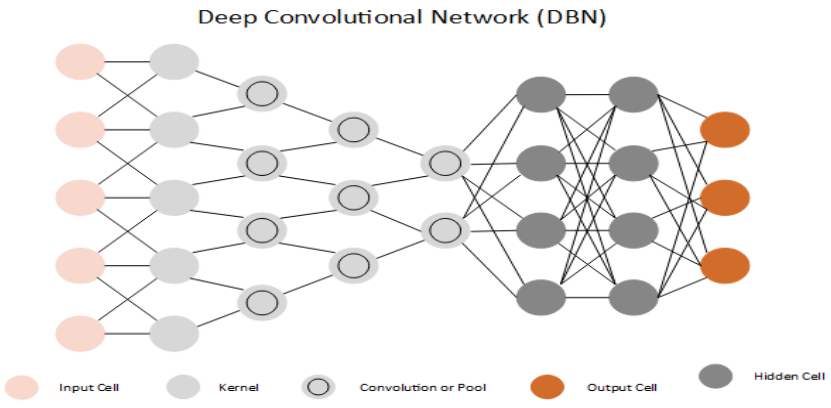


Figure 12- Réseaux de neurones convolutifs (CNN) [140]

couche de classification. La couche d'entrée prend les entrées brutes et produit les plongements. Ensuite, les couches d'extraction d'entités, qui incluent les couches de convolution et de mise en commun, apprennent les caractéristiques pertinentes. La couche de convolution applique des filtres appelés détecteurs d'entités pour apprendre les entités et produire la carte des entités. La couche de mise en commun, également connue sous le nom de méthode de réduction dimensionnelle, est utilisée pour extraire les caractéristiques pertinentes, en laissant celles qui ne sont pas nécessaires. Enfin, les entités produites par les couches d'entités sont transmises à la couche de classification, qui est constituée d'un réseau entièrement connecté avec un classificateur.

Dans l'étude de [45], les chercheurs ont représenté un cadre à sept couches pour analyser les sentiments des phrases. Ce travail de trame dépend de réseau neuronal convolutif et de Word2vec pour l'analyse des sentiments et pour calculer la représentation vectorielle, respectivement. Word2vec a été proposé par Google. La technologie Dropout, la normalisation et l'unité linéaire paramétrique rectifiée (PReLU), ont été utilisées pour améliorer l'exactitude et la généralisabilité du modèle proposé. Le cadre a été vérifié sur l'ensemble des données de rottentomatoes.com qui contient le corpus d'extraits de critiques de films, l'ensemble des données se compose de cinq étiquettes positives, plutôt positives, neurales, négatives et quelque peu négatives. En comparant le modèle proposé avec d'autres modèles tels que le réseau neuronal récursif Matrix-Vector (MV-RNN) et le réseau neuronal récursif (RNN), le modèle proposé a surpassé ces modèles avec une précision de 45,5%.

Les auteurs [46] ont proposé le système d'apprentissage en profondeur pour l'analyse des sentiments de Twitter. L'objectif principal de ce travail était d'initialiser le poids des paramètres du réseau de neurones convolutionnels et il est essentiel de former le modèle avec précision tout en évitant la nécessité d'ajouter de nouvelles fonctionnalités. Un langage neuronal est utilisé pour initialiser le mot d'intégration et est formé par un grand groupe de tweets non supervisés. Pour affiner davantage l'incorporation sur un corpus supervisé volumineux, un réseau neuronal conventionnel est utilisé. Pour initialiser le réseau, des mots et des paramètres précédemment intégrés ont été utilisés, ayant la même architecture et la même formation sur le corpus supervisé qu'au Semeval-2015. Les composants utilisés dans les travaux proposés sont les activations, le regroupement de matrices de phrases, les couches softmax et convolutionnelles. Pour former le

réseau, des algorithmes de descente de gradient stochastique (SGD) et d'optimisation de fonction non convexe ont été utilisés et pour calculer l'algorithme de propagation inverse des gradients. La technique de décrochage a été utilisée pour améliorer la régularisation des réseaux de neurones. Le modèle d'apprentissage en profondeur est appliqué à deux tâches: la tâche au niveau du message et la tâche au niveau de la phrase de Semeval-2015 pour prédire la polarité et obtenir des résultats élevés. En appliquant six ensembles de tests, le modèle proposé se situe au premier rang en termes de précision.

Une recherche détaillée de [47] a présenté un aperçu de l'analyse des sentiments liés au micro-blog. Le but de cet effort était d'obtenir les opinions et les attitudes des utilisateurs sur les événements chauds en utilisant le réseau de neurones convolutifs. L'utilisation de CNN surmonte le problème de l'extraction de caractéristiques explicites et apprend implicitement grâce aux données d'apprentissage. Pour collecter les données de la cible, l'URL d'entrée et le crawler ciblé ont été utilisés, 1000 commentaires de micro-blogs ont été collectés sous forme de corpus et divisés en trois étiquettes, à savoir 274 émotions neutres, 300 émotions négatives et 426 émotions positives. Le modèle proposé a été comparé aux précédents des études comme celles-ci qui avaient utilisé CRF, SVM et d'autres algorithmes traditionnels pour effectuer une analyse des sentiments à un prix élevé. Cependant, la performance prouve que le modèle proposé est raisonnable et suffisant pour améliorer la précision en termes d'analyse des émotions.

Cette étude [48] a proposé un nouveau cadre de réseau neuronal convolutif pour l'analyse des sentiments visuels afin de prédire les sentiments du contenu visuel. CNN a été implémenté en utilisant Caffe et Python sur une machine Linux. L'approche d'apprentissage par transfert et l'hyper-paramètre ont été utilisés dans les biais et les pondérations sont utilisées à partir de GoogLeNet préformé. Alors que CNN améliore ses performances en augmentant sa taille et sa profondeur, un modèle CNN très profond, inspiré de GoogLeNet, est proposé avec 22 couches pour l'analyse des sentiments. Il est optimisé en utilisant l'algorithme SGD (Stochastic Gradient Descent). La stratégie avec 60 époques a été réalisée pour la formation du réseau comme GoogLeNet a réalisé 250 époques. Pour le travail expérimental, un ensemble de données de Twitter contenant 1269 images est sélectionné et une rétro-propagation est appliquée. Amazon Mechanical Turk (MTurk) et l'intelligence populaire sont utilisés pour étiqueter les images. Cinq travailleurs ont été impliqués pour générer une étiquette de sentiment en faveur de chaque image.

Le modèle proposé a été évalué sur cet ensemble de données et a acquis de meilleures performances que les systèmes existants. Les résultats montrent que le système proposé atteint des performances élevées sans ajustement précis sur l'ensemble de données Flickr. Cependant AlexNet a été utilisé dans des travaux précédents et GoogleNet a fourni près de 9% de progression des performances par rapport à AlexNet. En convertissant GoogLeNet en un cadre d'analyse des sentiments visuels, une meilleure extraction des fonctionnalités a été obtenue. Un état stable et fiable a été obtenu en utilisant des hyper paramètres.

Tao et al [49], ont suggéré une méthodologie de division et de conquête qui a initialement catégorisé les phrases en types disparates, puis a exécuté l'analyse des sentiments séparément sur des phrases comme pour chaque type. Surtout, il a été constaté que les phrases avaient tendance à être extrêmement complexes si elles comprenaient des mots plus sentimentaux. Ainsi, il a été suggéré d'utiliser un modèle de séquence centré sur NN pour classer les phrases avec opinion en trois types selon le nombre de cibles transpirées dans une phrase. Chaque groupe de phrases a ensuite été fourni à un CNN à une dimension séparément pour la classification sentimentale. Cette approche a été évaluée sur quatre ensembles de données de classification sentimentale et contrastée avec des références étendues. Les résultats expérientiels ont montré que la catégorisation du type de phrase pouvait augmenter les performances de l'analyse des sentiments au niveau de la phrase.

Gichang et al. [50] ont recommandé une méthodologie pour reconnaître les mots clés différenciant les phrases négatives et positives en utilisant une méthodologie d'apprentissage faiblement supervisée centrée sur un CNN. Dans ce modèle, tous les mots étaient signifiés comme un vecteur à valeur continue tandis que toutes les phrases étaient signifiées comme une matrice dont les lignes correspondaient au vecteur de mots utilisé dans la phrase. Par la suite, le CNN a été formé en utilisant ces matrices de phrases comme entrées, en plus, les étiquettes de sentiment en tant que sortie. Après la formation, le schéma d'attention aux mots a été mis en œuvre pour reconnaître les mots à contribution plus élevée afin de classer les résultats avec la carte d'activation de classe en utilisant les poids. Pour valider la méthodologie recommandée, l'exactitude de la classification et le taux de mots de polarité parmi les mots ayant le score le plus élevé ont été évalués à l'aide de deux jeux de données d'examen de films. Le résultat expérientiel

a confirmé que le modèle recommandé pouvait catégoriser correctement la polarité de la phrase et reconnaître avec succès les mots correspondants avec les scores de polarité les plus élevés.

2.1.2.3 Réseaux de neurones récurrents (RNN)

Un autre type de réseau neuronal de propagation vers l'avant (feed-forward) proposé par Elman en 1990, largement utilisé et populaire dans l'apprentissage en profondeur, est le réseau de neurones récurrent. Les RNNs sont également compétents pour comprendre efficacement la structure des phrases. Ces modèles font partie des modèles DL qui sont le meilleur choix pour des tâches telles que l'analyse de sentiments. Les RNNs sont une classe de réseaux de neurones dont les connexions entre les neurones forment un cycle dirigé, qui jouent un rôle important dans la propagation de la fonction d'activation vers la séquence d'entrée entrante. La fonction principale de RNN est le traitement des informations séquentielles sur la base de la mémoire interne capturée par les cycles dirigés. Contrairement aux réseaux de neurones traditionnels, RNN peut se souvenir du calcul précédent des informations et peut les réutiliser en les appliquant à l'élément suivant dans la séquence d'entrées.

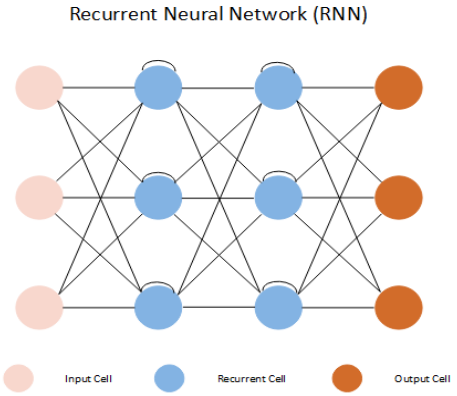


Figure 13- Réseaux de neurones récurrents (RNN) [140]

Un type spécial de RNN est la mémoire à court long-terme (LSTM) [51], qui est capable d'utiliser une mémoire longue qui permet à la cellule de mémoire de conserver les informations pendant une longue période ou de jeter les résultats de calcul précédents. Cependant, le LSTM est accusé d'avoir une structure complexe.

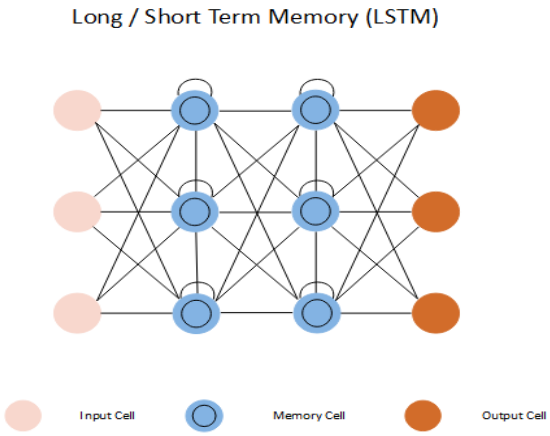


Figure 14- Mémoire à court long-terme (LSTM) [140]

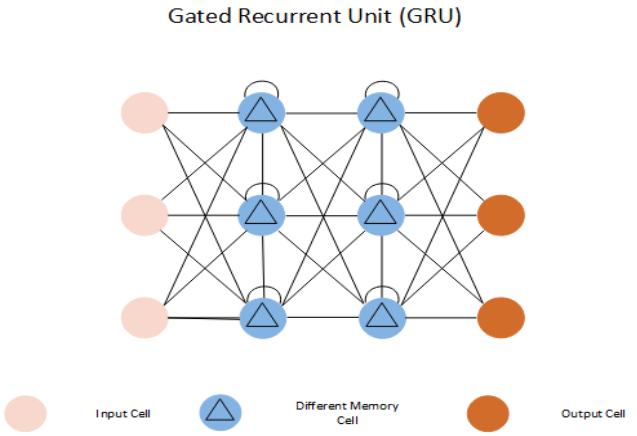


Figure 15- Unité récurrente fermée (GRU) [140]

Ainsi, l'unité récurrente fermée (GRU) [52] a été conçue comme une variante simplifiée de LSTM et plus efficace en termes de puissance de calcul par rapport au LTSM et le RNN classique.

Wenge et al. [53] ont recommandé un modèle d'analyse des sentiments centré sur le RNN, qui a pris une partie d'un document en entrée, puis les parties suivantes ont été utilisées pour prévoir la distribution de l'étiquette sentimentale. La méthodologie recommandée a appris la représentation des mots et aussi la distribution sentimentale. Des études expérientielles ont été exécutées sur des ensembles de données couramment utilisés et les résultats ont prouvé son potentiel propice.

Ces auteurs [54] ont proposé un modèle de séquence pour se concentrer sur l'intégration des revues de nature temporelle aux produits, car ces revues étaient moins focalisées dans les études existantes. La combinaison d'un GRU avec un RNN est utilisée pour apprendre des représentations dispersées de produits et d'utilisateurs. Pour la classification des sentiments, ces

représentations ont alimenté le classificateur d'apprentissage automatique. L'approche a été évaluée sur trois ensembles de données collectés auprès de Yelp et IMDB. Chaque avis étiqueté en fonction de la note. Pour former le réseau, l'algorithme de rétro-propagation avec la méthode d'optimisation stochastique d'Adam a été utilisé. Les résultats montrent que la modélisation séquentielle des produits dispersés et l'apprentissage de la représentation des utilisateurs améliorent la classification du sentiment de performance au niveau du document et l'approche proposée permet d'obtenir des résultats de haute technologie sur les ensembles de données de référence. Le résultat du modèle proposé par rapport à de nombreuses lignes de base, y compris les réseaux de neurones récursifs, le réseau de neurones du produit utilisateur, word2vec, le vecteur de paragraphe et l'algorithme JMARS.

Fei et al. [55] ont suggéré une conception centrée sur le LSTM qui était sensible aux mots qui existaient dans le vocabulaire; par conséquent, les mots clés influencent la sémantique du document complet. Le modèle suggéré a été évalué dans une tâche d'analyse des sentiments de texte court sur deux ensembles de données comme IMDB et SemEval-2016. Les résultats expérientiels ont indiqué que la conception a surpassé la LSTM de base de 1% à 2% en termes de précision et a été efficace avec une amélioration notable des performances par rapport à de nombreuses conceptions sémantiques latentes non RNN (en particulier dans la gestion de textes courts). Il a également intégré l'idée de modèle d'unité récurrente fermée (GRU) et a atteint de bonnes performances, ce qui a confirmé que cette méthodologie était adéquate pour augmenter les modèles DL disparates.

2.1.2.4 Réseaux de neurones récursifs (RvNN)

Le réseau neuronal récursif (RvNN) introduit en 1996 par Goller et al [56], réside dans l'apprentissage supervisé est un type de réseau neuronal qui peut être considéré comme une généralisation de RNN. Les RvNN permettent aux réseaux de neurones de gérer des entrées structurées de n'importe quelle forme, comme des arbres et des graphes. Les réseaux de neurones récursifs sont généralement utilisés pour apprendre une structure de graphe acyclique dirigée à partir de données. Ils sont très puissants simplement parce qu'ils représentent chaque mot comme un vecteur et un opérateur très intuitif, de sorte qu'un mot peut agit sur le mot suivant et change sa polarité en faisant tourner le vecteur de fin pour signifier autre chose. Cependant, ces réseaux ont besoin de beaucoup de connaissances en formation (des arbres d'analyse sont nécessaires pour former ces réseaux). Les RvNN ont montré des résultats impressionnants dans l'analyse des

sentiments en raison de leur capacité à représenter l'entrée de séquence sous la forme d'arbre et de capacité à représenter le contexte dans lequel un mot apparaît, en plus, leur capacité à apprendre des informations sémantiques et syntaxiques à partir des entrées les conduit à réussir dans l'analyse des sentiments.

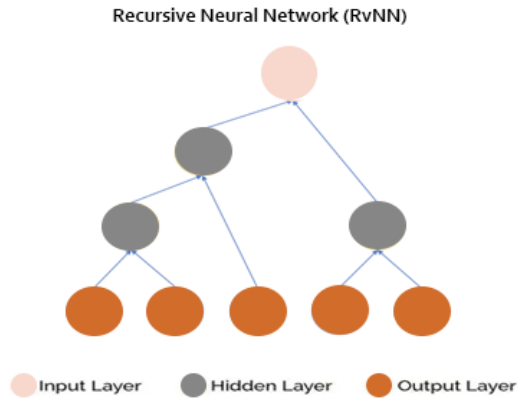


Figure 16- Réseaux de neurones récursifs (RvNN)

Dans l'analyse des sentiments, la variante couramment utilisée du RvNN est le réseau tenseur neural récursif (RNTN) qui a été introduit en 2013 [57]. Ce réseau a été appliqué avec succès à l'analyse des sentiments. Son intention principale est de capturer le sentiment d'une phrase de n'importe quelle longueur en ne se basant pas seulement sur ses composants mais aussi explorer l'ordre dans lequel les mots sont regroupés syntaxiquement.

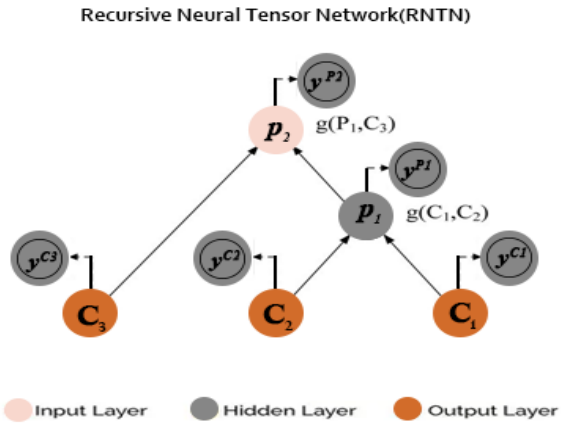


Figure 17- Réseau tenseur neural récursif (RNTN)

Ainsi, RNTN réalise cette représentation basée sur une fonction de composition basée sur le tenseur, qui est appliquée à tous les nœuds de l'arbre.

En 2011, Socher et al. [58] ont proposé une architecture RvNN capable de gérer les entrées de différentes modalités. Les auteurs montrent un exemple d'utilisation de RvNN pour classer des phrases en langage naturel. Alors, qu'une phrase est divisée en mots. RvNN calcule le score d'une paire possible pour les fusionner et construire un arbre syntaxique. Pour chaque paire d'unités, RvNN calcule un score pour la plausibilité de la fusion. La paire avec le score le plus élevé est ensuite combinée en un vecteur de composition. Après chaque fusion, RvNN générera (1) une plus grande région de plusieurs unités, (2) un vecteur de composition représentant la région, et (3) l'étiquette de classe (par exemple, si les deux unités sont deux mots nominaux, l'étiquette de classe pour la nouvelle région serait une expression nominale). La racine de l'arborescence RvNN est la représentation vectorielle compositionnelle de toute la région. La figure 1 montre un exemple d'arbre RvNN.

Dans cette étude [58], un modèle comprenant RNTN et Sentiment Treebank a été proposé pour clarifier correctement les effets de composition à différents niveaux de phrases, c'est-à-dire des phrases positives et négatives. Le modèle proposé a été comparé à tous les modèles existants. Dans les modèles existants, le sens des phrases longues ne peut pas être exprimé efficacement par des espaces de mots sémantiques. Par conséquent, pour la détection des sentiments, des ressources d'évaluation et de formation plus riches et supervisées sont nécessaires car elles nécessitent des modèles de composition plus influents. Le RNTN a atteint une précision de 80,7% dans la prédiction des sentiments en effectuant un étiquetage à grain fin sur toutes les phrases et a dépassé les modèles précédents.

Les travaux proposés en [59] construisent un Treebank pour les sentiments des Chinois sur les données sociales afin de surmonter la carence de corpus étiquetés et volumineux dans les modèles existants. Pour prédire les étiquettes au niveau de la phrase, c'est-à-dire positif ou négatif, le modèle neuronal profond récursif (RNDM) a été proposé et a obtenu de plus hautes performances que SVM, Naïve Bayes et Entropie maximale. 2270 critiques de films ont été collectées sur le site Web et l'outil de segmentation de mots chinois ICTCLAS a été utilisé pour segmenter ces commentaires. Cinq classes ont été réglées pour chaque phrase et l'analyseur syntaxique Standford a demandé l'analyse syntaxique de la phrase. Le modèle proposé a amélioré la prédiction des étiquettes de sentiment des phrases en concluant 13550 phrases chinoises et 14964 mots.

Cette étude [60] a fourni un cadre généralisé et évolutif pour reconnaître les meilleurs vendeurs de cartes/logiciels malveillants. Le modèle est basé sur un apprentissage approfondi pour l'analyse des sentiments et utilisé dans la classification des fils et l'échantillonnage des boules de neige pour évaluer la qualité du service/produit du vendeur en analysant les commentaires des clients. L'évaluation du modèle proposé a été menée sur le forum de cardage russe et un robot d'exploration Web a été utilisé pour rassembler les éléments de conversation du forum. Une banque d'arbres de sentiments a été utilisée et elle a été formée en utilisant un réseau de tenseur neuronal récursif sur un corpus de revue en ligne. Pour évaluer la validité et l'efficacité, deux expériences ont été menées dans lesquelles le modèle proposé a été comparé aux modèles basés sur Naïve Bayes, KNN et SVM. Cette étude a recherché les vendeurs qui sont très bien notés pour les services/produits malveillants et l'efficacité de l'apprentissage en profondeur pour reconnaître ces vendeurs. Les résultats ont indiqué que les techniques d'apprentissage en profondeur atteignent des résultats supérieurs aux classificateurs peu profonds et il a été établi que les vendeurs de cartes ont moins de notes que les vendeurs de logiciels malveillants.

2.1.2.5 Réseaux de croyances profondes (DBN)

Les réseaux de croyances profondes sont des méthodes génératives probabilistes de type réseaux de neurones feed-forward qui sont initié en 2006. Les DBNs sont composées d'une couche de neurones visibles et de plusieurs couches cachées de variables latentes stochastiques qui aident à apprendre les représentations de niveau supérieur à partir des variables d'entrée.

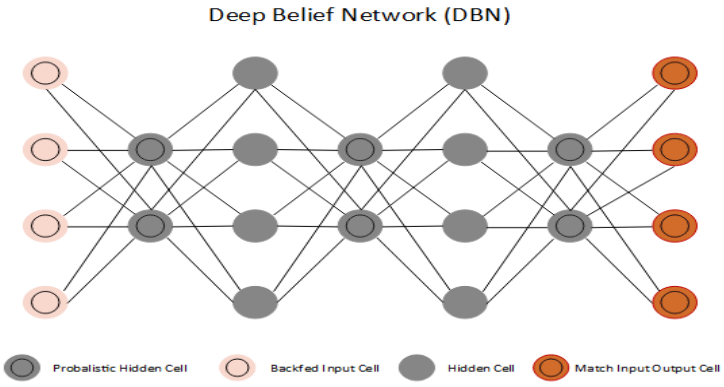


Figure 18- Réseaux de croyances profondes (DBN) [140]

Les DBNs comprennent des machines Boltzmann restreintes (RBM) qui traitent des caractéristiques de niveau supérieur de manière non supervisée. Néanmoins, les unités entre les couches cachées dans les DBNs sont bidirectionnelles, ce qui rend les DBNs différents des autres réseaux de neurones feed-forward. De plus, il n'y a pas de connexions entre les unités au sein des mêmes couches. Cette différence est réalisée par les couches RBMs dans la phase pré-entraînée. Le principal avantage des DBNs est qu'ils peuvent exploiter une grande quantité de données non étiquetées car ils fonctionnent sur le principe de la préformation par couche. Cependant, le principal inconvénient des DBNs réside dans le fait qu'ils sont difficiles à former. Le seul inconvénient du DBN est qu'il est coûteux et prend du temps.

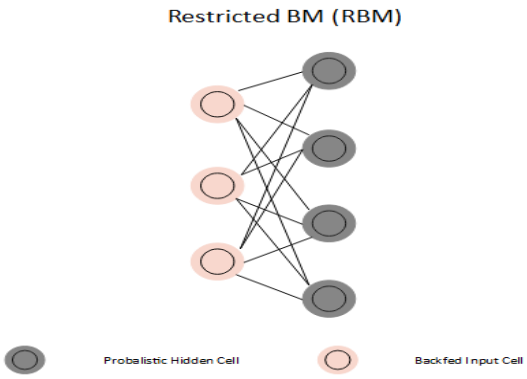


Figure 19- Machines Boltzmann restreintes (RBM) [140]

Une autre étude de [61] a utilisé un réseau de croyances profondes avec un vecteur de mots pour la détection politique dans des articles coréens. Le modèle proposé a utilisé SVM pour le calcul des biais, un pipeline en cinq étapes pour la détection des biais politiques, un robot d'indexation python pour rassembler des articles de presse. L'ensemble de données contenait 50 000 articles politiques du 01 janvier 2014 au 28 février 2015. Les résultats ont montré une précision de 81,8% en prédisant correctement les étiquettes et les résultats contenaient une erreur quadratique moyenne de 0,120.

Shusen et al. [62] ont présenté une méthodologie d'apprentissage semi-supervisé (SSL) en deux étapes appelée DBN floue (FDBN) pour la classification sentimentale. Principalement, le DBN commun a été formé par le SSL en utilisant l'ensemble de données d'apprentissage. Ensuite, une fonction d'appartenance floue (FMF) a été conçue pour toutes les classes de revues centrées sur l'architecture DL. Comme la formation DBN mappe chaque revue sur l'espace de sortie DBN, la diffusion de la totalité des échantillons de formation sur l'espace a été évaluée comme une

connaissance antérieure, en outre, a été codée par des séquences de FMF. Deuxièmement, fondée sur les fonctions d'appartenance floues, le DBN atteint à l'étape principale, une architecture FDBN a été construite et l'étape d'apprentissage supervisé a été utilisée pour augmenter les performances de classification du FDBN. FDBN a hérité de la puissante compétence d'abstraction de DBN et a délimité la compétence de classification floue attrayante pour le traitement des données sentimentales.

Cette recherche [63] a proposé un réseau de croyances profondes avec sélection de caractéristiques (DBNFS) pour surmonter les problèmes de vocabulaire, le réseau a utilisé un corpus d'entrée ainsi que de nombreuses couches cachées. La technique Chi-Squared de sélection des caractéristiques a été utilisée pour améliorer le DBN dans le but de réduire la complexité de la saisie de vocabulaire et d'éliminer les caractéristiques non pertinentes. En appliquant la technique du Chi-Squared, la phase d'apprentissage de DBN a été améliorée en DBNFS. Dans ce travail, deux nouvelles fonctionnalités de tâches, la sélection et la réduction ont été utilisées ainsi que de nombreuses autres tâches des techniques de classification existantes, telles que le portionnement des données, l'extraction des fonctionnalités, la formation et les tests de modèle. La performance du DBNFS a été démontrée et le temps de formation et la précision du DBNFS proposé ont également été comparés à d'autres algorithmes. Cinq ensembles de données de classification des sentiments ont été utilisés pour l'estimation, les jeux de données sont des livres (BOO), de l'électronique (ELE), des DVD (DVD), des appareils de cuisine (KIT) et des critiques de films (MOV). Pour une comparaison équitable, les paramètres d'apprentissage étaient les mêmes que pour les travaux existants. La précision a été évaluée en comparant la quantité de fonctionnalités avant et après la sélection et la réduction des fonctionnalités. Les résultats de précision ont été comparés aux travaux précédents et se sont avérés meilleurs DBNFS que DBN. Le temps de formation était également plus faible en DBNFS qu'en DBN. Le temps de formation a été amélioré en raison de la structure profonde simple et de la méthode de sélection des caractéristiques proposées.

Yong et al. [64] ont suggéré une forme positionnelle de mot ainsi qu'une représentation matricielle mot à segment pour intégrer les informations de position aux DBNs pour la classification sentimentale. Par la suite, la performance a été évaluée par la précision totale. Par conséquent, ces résultats expérimentaux ont montré qu'en incluant des informations de position sur dix petits ensembles de données de texte, la représentation matricielle était la plus efficace.

En examinant le formulaire de contribution de position linéaire, il a en outre suggéré que les informations de position soient prises en compte pour les tâches SA ou NLP.

2.1.2.6 Réseaux de neurones hybrides

Les algorithmes DL individuels ont été largement utilisés et se sont révélés produire des résultats impressionnants en NLP, en particulier pour l'analyse des sentiments. Par conséquent, différents chercheurs ont tenté de combiner ces techniques pour améliorer les performances de leurs modèles en obtenant les avantages offerts par chaque type. Par exemple, les CNN sont des modèles bien reconnus pour extraire des entités locales. D'un autre côté, les RNN sont bien connus pour gérer les dépendances à longue distance. Par conséquent, l'idée intéressante est de les intégrer afin que le modèle puisse extraire les deux types de fonctionnalités. Récemment, des chercheurs ont proposé divers modèles hybrides pour accomplir diverses tâches dans l'analyse des sentiments.

Dans cette étude de recherche [65], un modèle hybride a été proposé qui consiste en un réseau neuronal probabiliste (PNN) et un Boltzmann restreint à deux couches (RBM). Le but de proposer ce modèle hybride d'apprentissage en profondeur est d'atteindre une meilleure précision de la classification des sentiments. La polarité, c'est-à-dire que les avis négatifs et positifs varient en fonction du contexte afin de résoudre ce type de problème, ce modèle fonctionne bien, les avis neutres ne sont pas pris en compte. La précision a été améliorée pour cinq ensembles de données en les comparants à l'existant état de l'art de Arnold et al. [66]. Il n'y a pas de ressources externes dans l'approche proposée, telles que le tagueur de point de vente et le dictionnaire des sentiments etc, donc c'est plus rapide que le concurrent.

Shusen et al. [67] ont suggéré un algorithme appelé réseau profond actif (ADN). ADN a été construit par RBM (Restricted Boltzmann Machine) avec un apprentissage non supervisé centré sur des évaluations non étiquetées et maximales. Après cela, la structure construite a été modifiée au moyen d'un apprentissage supervisé centré sur la descente de gradient ayant une fonction de perte exponentielle. Deuxièmement, un apprentissage actif a été utilisé pour reconnaître les avis qui étaient marqués comme données de formation, après quoi, ont utilisé les avis sélectionnés et tous les avis non étiquetés pour la formation de l'architecture ADN. De plus, pour intégrer la densité de l'information avec la méthodologie suggérée de l'IADN (Information ADN), qui pourrait utiliser la densité de l'information de l'ensemble des avis non étiquetés pour sélectionner les avis étiquetés manuellement. Des expériences sur cinq ensembles de données de

classification sentimentale ont confirmé que l'IADN et l'ADN ont surpassé les algorithmes classiques et les techniques DL utilisées pour la classification sentimentale.

Cette étude [68] a proposé deux techniques d'apprentissage approfondi pour la classification des sentiments des données Twitter thaïlandaises, à savoir le réseau CNN et LSTM. Le traitement des données a été effectué correctement. Les données ont été collectées auprès des utilisateurs et de leurs abonnés de Twitter thaïlandaises. Après filtrage des données, seuls les utilisateurs avec des tweets thaïlandais et des tweets avec des caractères thaïlandais ont été sélectionnés. Cinq expériences ont été menées pour obtenir les meilleurs paramètres d'apprentissage en profondeur, pour comparer l'apprentissage en profondeur avec des techniques classiques et pour obtenir l'importance de la séquence de mots. Une validation croisée triple a été utilisée pour vérifier le processus. Les résultats ont conclu que la précision est élevée en DNN que LSTM et les deux techniques d'apprentissage en profondeur sont plus précises que SVM et Naïve Bayes mais moins que l'entropie maximale. Une précision élevée a été trouvée plus dans les phrases originales que dans les phrases mélangées, la séquence des mots est donc importante.

2.2 Approches basées sur un lexique

L'application d'un lexique est l'une des deux principales techniques de l'analyse des sentiments et consiste à calculer le sentiment à partir de l'orientation sémantique du mot ou des phrases qui apparaissent dans un texte. Avec cette approche, un dictionnaire de mots positifs et négatifs est nécessaire, avec une valeur de sentiment positive ou négative attribuée à chacun des mots. Différentes techniques de création de dictionnaires ont été proposées, notamment des techniques manuelles et automatiques. D'une manière générale, dans les techniques basées sur le lexique, un morceau de phrase texte est représenté comme un sac de mots. Suite à cette représentation de la phrase, les valeurs de sentiment du dictionnaire sont attribuées à tous les mots ou expressions positifs et négatifs de la phrase. Une fonction de combinaison, telle que somme ou moyenne, est appliquée afin de faire la prédiction finale concernant le sentiment général pour la phrase. Outre une valeur sentimentale, l'aspect du contexte local d'un mot est généralement pris en considération, comme la négation ou l'intensification. Les deux techniques automatisées sont présentées dans les sous-sections suivantes.

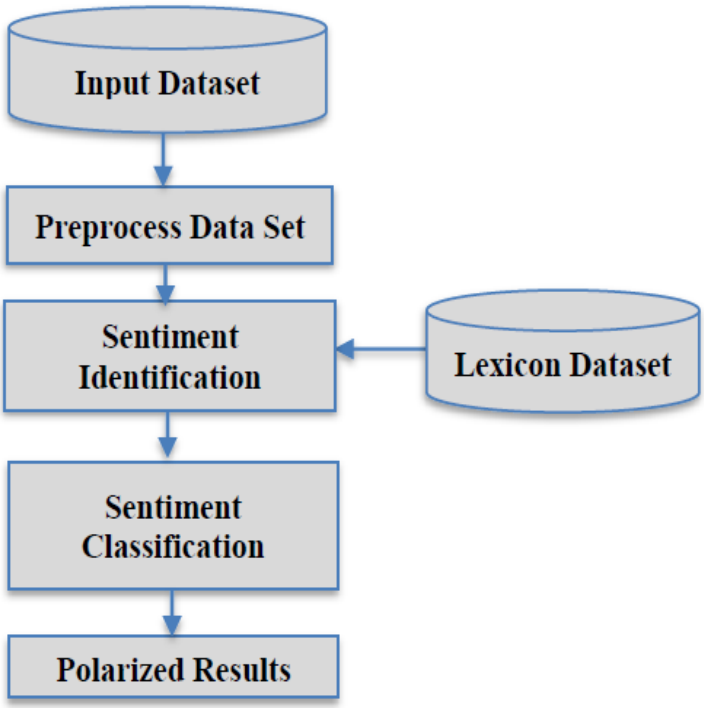


Figure 20- Modèle basée sur un lexique de classification de sentiments

2.2.1 Approches basées sur le Dictionnaire

L'utilisation d'un dictionnaire pour compiler les mots de sentiment est une approche évidente parce que la plupart des dictionnaires (par exemple, WordNet [69]) énumèrent des synonymes et des antonymes pour chaque mot. Ainsi, une technique simple dans cette approche consiste à utiliser quelques mots de sentiment de départ pour démarrer en fonction de la structure des synonymes et des antonymes d'un dictionnaire. Plus précisément, cette méthode fonctionne comme suit: Un petit ensemble de mots de sentiment (graines) avec des orientations positives ou négatives connues est d'abord collecté manuellement, ce qui est très facile. L'algorithme développe ensuite cet ensemble en recherchant dans le WordNet ou un autre dictionnaire en ligne leurs synonymes et antonymes. Les mots nouvellement trouvés sont ajoutés à la liste des graines. La prochaine itération commence. Fin du processus itérative quand plus aucun nouveau mot ne peut être trouvé. Une fois le processus terminé, une inspection manuelle peut être effectuée pour supprimer ou corriger les erreurs.

Turney et Littman [70], utilisent la méthode basée sur le PMI (Pointwise Mutual Information) comme dans [9]. Cette mesure a été utilisée pour calculer l'orientation des sentiments d'un mot donné en mesurant le degré de dépendance statistique entre deux termes. Plus précisément, elle calcule l'orientation du mot à partir de la force de son association avec un ensemble de mots positifs (bon, gentil, excellent, positif, chanceux, correct et supérieur), moins la force de son association avec un ensemble de mots négatifs (mauvais, méchant, pauvre, négatif, malheureux, mauvais et inférieur).

Les auteurs de [71] proposent une approche, qui a utilisé une méthode basée sur la distance WordNet pour déterminer l'orientation des sentiments d'un adjectif donné. La distance d (t_1, t_2) entre les termes t_1 et t_2 est la longueur du chemin le plus court qui relie t_1 et t_2 dans WordNet. L'orientation d'un terme adjectif t est déterminée par sa distance relative entre deux termes de référence bons et mauvais, SO(t) = (d (t, mauvais) - d (t, bon)) / d (bon, mauvais). t est positif si SO (t) > 0, et est négatif sinon. La valeur absolue de SO(t) donne la force du sentiment.

Esuli et Sebastiani [72] ont utilisé l'apprentissage supervisé pour classer les mots en classes positives et négatives. Étant donné un ensemble P de mots de départ positifs et un ensemble N de mots de départ négatifs, les deux ensembles de semences sont d'abord développés à l'aide de relations de synonymes et d'antonymes dans un dictionnaire en ligne (par exemple, WordNet) pour générer les ensembles étendus P' et N', et former l'ensemble de formation. L'algorithme utilise ensuite toutes les gloses du dictionnaire pour chaque terme de P' U N' pour générer un vecteur caractéristique. Un classificateur binaire est ensuite construit en utilisant différents algorithmes d'apprentissage. Le processus peut également être exécuté de manière itérative. C'est-à-dire que les termes positifs et négatifs nouvellement identifiés et leurs synonymes et antonymes sont ajoutés à l'ensemble d'apprentissage, un classificateur mis à jour peut être construit et ainsi de suite. Dans [73], les auteurs ont également inclus l'objectif de catégorie. Pour élargir l'ensemble de semences objectif, les hyponymes ont été utilisés en plus des synonymes et des antonymes. Ils ont ensuite essayé des différentes stratégies pour effectuer la classification en trois classes.

Andreevskaia et Bergler [74] ont proposé une méthode d'amorçage (bootstrapping) avec plusieurs techniques pour étendre les ensembles initiaux positifs et négatifs et pour nettoyer les ensembles développés (en supprimant les non adjectifs et les mots dans les ensembles positifs et négatifs). De plus, leur algorithme effectue également plusieurs exécutions du processus

d'amorçage à l'aide de sous-ensembles de semences non chevauchants. Chaque exécution trouve généralement un ensemble de mots de sentiments légèrement différent. Un score de chevauchement net pour chaque mot est ensuite calculé sur la base du nombre de fois où le mot est découvert dans les passages en tant que mot positif et en tant que mot négatif. Le score est ensuite normalisé à [0, 1] sur la base de la théorie des ensembles flous.

En [75] et [76], de nombreux heuristique ont été utilisées pour construire un lexique de sentiments à partir de documents HTML basés sur des structures de mise en page Web. Par exemple, un tableau dans une page Web peut avoir une colonne indiquant clairement les orientations positives ou négatives (par exemple, les avantages et les inconvénients) du texte entourant. Ces indices peuvent être exploités pour extraire un grand nombre de phrases d'opinion positives et négatives à partir d'un large ensemble de pages Web. Des phrases adjectives sont ensuite extraites de ces phrases et des orientations de sentiment attribuées en fonction de différentes statistiques de leurs occurrences dans les ensembles de phrases positifs et négatifs respectivement.

Dans [77], une méthode d'amorçage différente a été proposée, qui utilisait un ensemble de semences positives, négatives et également neutres. L'approche fonctionne sur la base d'un graphe sémantique dirigé et pondéré où les nœuds voisins sont des synonymes ou des antonymes de mots dans WordNet et ne font pas partie de l'ensemble neutre d'origine. L'ensemble neutre est utilisé pour arrêter la propagation des sentiments à travers des mots neutres. Les poids des bords sont pré-attribués en fonction d'un paramètre de mise à l'échelle pour différents types de bords, c'est-à-dire des bords synonymes ou antonymes. Chaque mot est ensuite noté (donnant une valeur de sentiment) à l'aide d'une version modifiée de l'algorithme de propagation d'étiquette [78]. Au début, chaque mot de graine positive reçoit le score de +1, chaque graine négative reçoit le score de -1 et tous les autres mots reçoivent le score de 0. Les scores sont révisés au cours du processus de propagation. Lorsque la propagation s'arrête après un certain nombre d'itérations, les scores finaux après une mise à l'échelle logarithmique sont attribués aux mots en tant que degrés de positif ou de négatif.

Dans [79], trois méthodes d'apprentissage semi-supervisées basées sur des graphes ont été essayées pour séparer les mots positifs et négatifs en fonction d'un ensemble positif de graines, d'un ensemble négatif de graines et d'un graphe de synonyme extrait de WordNet. Il a été démontré que le Mincut et le Mincut aléatoire ont produit de meilleurs scores F, mais la

propagation des étiquettes a donné des précisions significativement plus élevées avec de faibles rappels.

Qiu et He [80] ont utilisé une approche basée sur un dictionnaire pour identifier les phrases de sentiment dans la publicité contextuelle. Ils ont proposé une stratégie publicitaire pour améliorer la pertinence des annonces et l'expérience utilisateur. Ils ont utilisé l'analyse syntaxique et le dictionnaire des sentiments et ont proposé une règle basée sur approche pour aborder l'extraction des mots du sujet et l'identification de l'attitude des consommateurs dans l'extraction des mots clés publicitaires. Ils ont travaillé sur des forums Web de automotvieforums.com. Leurs résultats ont démontré l'efficacité de l'approche proposée pour l'extraction de mots clés publicitaires et la sélection d'annonces.

Hassan et Radev [81] ont présenté un modèle de marche aléatoire de Markov sur un graphe de parenté de mots pour produire une estimation de sentiment pour un mot donné. Il utilise d'abord des synonymes et des hypernymes WordNet pour créer un graphe de parenté de mots. Une mesure, appelée le temps moyen de frappe h (i | S), et ensuite définie et utilisée pour évaluer la distance d'un nœud i à un ensemble de nœuds (mots) S, qui est le nombre moyen de pas d'un marcheur aléatoire, à partir de l'état $i \epsilon S$, pour entrer la première fois dans un état $k \epsilon S$. Étant donné un ensemble de mots de départ positifs S^+ et un ensemble de mots de départ négatifs S^-, pour estimer l'orientation sentimentale d'un mot donné w, il calcule les temps de frappe $h(w \mid S^+)$ et $h(w \mid S^-)$. Si $h(w \mid S^+)$ est supérieur à $h(w \mid S^-)$, le mot est classé comme négatif, sinon positif.

Velikovich et al. [82] ont également proposé une méthode pour construire un lexique sensible à l'aide de pages Web. Il était basé sur un algorithme de propagation de graphe sur un graphe de similarité de phrase. Il a de nouveau supposé comme entrée un ensemble de phrases de départ positives et un ensemble de phrases de départ négatives. Les nœuds du graphe de phrases étaient les phrases candidates sélectionnées parmi tous les n-grammes jusqu'à la longueur 10 extraits de 4 milliards de pages Web. Seulement 20 millions de phrases candidates ont été sélectionnées en utilisant plusieurs heuristiques, par exemple, la fréquence et les informations mutuelles de limites des mots. Un vecteur de contexte pour chaque expression candidate a ensuite été construit sur la base d'une fenêtre de mots de taille six agrégée sur toutes les mentions de l'expression dans les 4 milliards de documents. L'ensemble des bords a été construit par calcul de similitude en cosinus des vecteurs de contexte des phrases candidates. Tous les bords (v_i,v_j) ont été rejetés s'ils n'étaient pas l'un des 25 bords les plus pondérés adjacents au nœud v_i ou v_j. Le poids du bord a

été réglé sur la valeur de similitude cosinus correspondante. Une méthode de propagation graphique a été utilisée pour calculer le sentiment de chaque phrase comme l'ensemble de tous les meilleurs chemins vers les mots de départ.

Dans [83], une autre méthode d'amorçage (Boostrap), mais très différente, a été proposée à l'aide de WordNet. Étant donné un ensemble de mots de départ, au lieu de simplement suivre le dictionnaire, les auteurs ont proposé un ensemble de règles d'inférence sophistiquées pour déterminer les orientations des sentiments des autres mots à travers un processus déductif. Autrement dit, l'algorithme prend en entrée des mots avec des orientations de sentiment connues (les germes) et produit des synsets (ensembles de synonymes) avec des orientations. Les synsets avec les orientations déduites peuvent ensuite être utilisés pour déduire davantage les polarités d'autres mots.

Dans le document [84], un système d'exploration des opinions basé sur les aspects nommé «Système d'orientation des sentiments basé sur les aspects» est proposé, qui extrait la caractéristique et les opinions des phrases et détermine si les phrases données sont positives, négatives ou neutres pour chaque caractéristique. La négation est également gérée par le système. Pour déterminer l'orientation sémantique des phrases, une technique basée sur un dictionnaire de l'approche non supervisée est adoptée. Pour déterminer les mots d'opinion et leurs synonymes et antonymes, WordNet est utilisé comme dictionnaire. Toutes les caractéristiques du produit sur lesquelles les avis sont donnés seraient identifiées et l'orientation de la phrase pour chaque caractéristique serait déterminée. La polarité de la phrase donnée est déterminée sur la base de la majorité des mots d'opinion. En fin de compte, le système générera le résumé des fonctionnalités des phrases positives, négatives et neutres qui seront plus faciles à lire, à analyser et à aider les utilisateurs à décider si le produit doit être acheté ou non.

En résumé, nous notons que l'avantage d'utiliser une approche basée sur un dictionnaire est que l'on peut facilement et rapidement trouver un grand nombre de mots de sentiment avec leurs orientations. Bien que la liste résultante puisse contenir de nombreuses erreurs, une vérification manuelle peut être effectuée pour la nettoyer, ce qui prend du temps mais ce n'est qu'un effort ponctuel. L'approche par dictionnaire présente un inconvénient majeur qui est l'incapacité à trouver des mots d'opinion avec des orientations spécifiques au domaine et au contexte. L'approche basée sur le corpus ci-dessous peut aider à résoudre ce problème.

2.2.2 Approches basées sur le corpus

Le principal inconvénient de l'approche par dictionnaire est que les orientations sentimentales des mots ainsi collectés sont générales ou indépendantes du domaine et du contexte. En d'autres termes, il est difficile d'utiliser l'approche basée sur un dictionnaire pour trouver des orientations dépendantes du domaine ou du contexte des mots de sentiment. Comme indiqué précédemment, de nombreux mots de sentiment ont des orientations dépendantes du contexte. Cependant, l'approche basée sur le corpus aide à résoudre ce problème. Ses méthodes commencent par une liste de départ de mots de sentiment connus (souvent à usage général), et puis découvrir d'autres mots de sentiment et leurs orientations à partir d'un corpus de domaine. Mais, le problème est plus compliqué que de simplement construire un lexique de sentiments spécifiques à un domaine car dans le même domaine, le même mot peut être positif dans un contexte mais négatif dans un autre. Ci-dessous, nous discutons certains des travaux existants qui ont tenté de résoudre ces problèmes. Notez bien que l'approche basée sur un corpus puisse également être utilisée pour créer un lexique de sentiment général si un corpus très grand et très divers est disponible, l'approche basée sur un dictionnaire est généralement plus efficace pour cela car un dictionnaire a tous les mots.

L'une des idées clés et aussi des premières a été proposée par Hazivassiloglou et McKeown [85]. Les auteurs ont utilisé un corpus et quelques mots adjectifs sentimentaux pour trouver des adjectifs sentimentaux supplémentaires dans le corpus. Leur technique a exploité un ensemble de règles ou conventions linguistiques sur les connecteurs pour identifier des mots de sentiment plus adjectifs et leurs orientations à partir du corpus. L'une des règles concerne la conjonction et ET, qui dit que les adjectifs conjoints ont généralement la même orientation. Par exemple, dans la phrase «Cette voiture est belle ET spacieuse», si «belle» est connue pour être positive, on peut en déduire que «spacieuse» est également positive. La conjonction ET dit par exemple que les adjectifs conjoints ont généralement la même orientation. Cette idée est appelée cohérence des sentiments, qui n'est pas toujours cohérente dans la pratique.

Il existe également des expressions négatives telles que MAIS, Cependant qui sont indiquées comme des changements d'opinion. Afin de déterminer si deux adjectifs conjoints ont des orientations identiques ou différentes, l'apprentissage est appliqué à un grand corpus. Ensuite, les liens entre les adjectifs forment un graphe et le regroupement est effectué sur le graphe pour produire deux ensembles de mots: positif et négatif.

Kanayama et Nasukawa [86] ont étendu l'approche en introduisant les concepts de cohérence sentimentale intra-sententielle (dans une phrase) et inter-sententielle (entre phrases voisines), qu'ils appellent cohérence. La cohérence inter-sententielle applique simplement l'idée aux phrases voisines. Autrement dit, la même orientation sentimentale est généralement exprimée en phrases consécutives. Les changements de sentiment sont indiqués par des expressions adverses telles que MAIS et CEPENDANT. Certains critères ont également été proposés pour déterminer s'il fallait ajouter un mot au lexique positif ou négatif. Cette étude était basée sur du texte japonais et a été utilisée pour trouver des mots de sentiment dépendant du domaine et leurs orientations.

Jiaoa et Zhoua [87] utilisent la méthode des champs aléatoires conditionnels (CRF) qui a été utilisée comme technique d'apprentissage séquentiel pour extraire des expressions d'opinion. Les auteurs ont également utilisé cette méthode afin de discriminer la polarité des sentiments par un algorithme d'appariement de motifs à plusieurs chaînes. Leur algorithme a été appliqué sur les revues en ligne chinoises. Ils ont établi de nombreux dictionnaires émotionnels. Ils ont travaillé sur des revues en ligne de voitures, d'hôtels et d'ordinateurs. Leurs résultats ont montré que leur méthode a atteint des performances élevées.

Xu et Liao [88] ont utilisé un modèle CRF à deux niveaux avec des interdépendances non fixées pour extraire les relations comparatives. Cela a été fait en utilisant les dépendances complexes entre les relations, les entités et les mots et les interdépendances non fixées entre les relations. Leur objectif était de créer un modèle graphique pour extraire et visualiser les relations comparatives entre les produits à partir des avis clients. Ils ont affiché les résultats sous forme de cartes de relations comparatives pour l'aide à la décision dans la gestion des risques d'entreprise. Ils ont travaillé sur les avis des clients mobiles d'Amazon.com, epinions.com, des blogs, du SNS et des e-mails. Leurs résultats ont montré que leur méthode peut extraire des relations comparatives plus précisément que d'autres méthodes, et leur carte de relation comparative est potentiellement un outil très efficace pour soutenir la gestion des risques d'entreprise et la prise de décision.

Cruz et Troyano [89] ont proposé une approche basée sur la taxonomie pour extraire les opinions au niveau des caractéristiques et les mapper dans la taxonomie des caractéristiques. Leur cible principale était une analyse des sentiments orientés domaine. Ils ont défini un ensemble de ressources spécifiques au domaine qui capturent des connaissances précieuses sur la façon dont

les gens expriment leurs opinions sur un domaine donné. Ils ont utilisé des ressources qui ont été automatiquement induites à partir d'un ensemble de documents annotés. Ils ont travaillé sur trois domaines différents (casques, hôtels et critiques de voitures) sur epinions.com. Ils ont comparé leur approche à d'autres techniques indépendantes du domaine. Leurs résultats ont prouvé l'importance du domaine afin de construire des systèmes d'extraction d'opinion précis, car ils ont conduit à une amélioration de la précision, par rapport aux techniques indépendantes du domaine.

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

2.2.2.1 Approche statistique

La recherche de modèles de cooccurrence ou de mots d'opinion peut être effectuée à l'aide de techniques statistiques. Cela pourrait être fait en dérivant les polarités postérieures en utilisant la cooccurrence d'adjectifs dans un corpus, comme proposé par Fahrni et Klenner [90]. Il est possible d'utiliser l'ensemble complet 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 assez grand [7]. La polarité d'un mot peut être identifiée en étudiant la fréquence d'occurrence du mot dans un grand corpus annoté de textes. Si le mot apparaît plus fréquemment parmi les textes positifs, alors sa polarité est positive. S'il apparaît plus fréquemment parmi les textes négatifs, alors sa polarité est négative. S'il a des fréquences égales, alors c'est un mot neutre.

Les mots d'opinion similaires apparaissent fréquemment ensemble dans un corpus. Par conséquent, si deux mots apparaissent fréquemment ensemble dans le même contexte, ils sont susceptibles d'avoir 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. Cela pourrait être fait en utilisant PMI [7].

Cao et Duan [91] ont utilisé une approche statistique d'analyse sémantique latente (LSA) qui est utilisée pour analyser les relations entre un ensemble de documents et les termes mentionnés

dans ces documents afin de produire un ensemble de modèles significatifs liés aux documents et termes. Les auteurs ont utilisé LSA pour trouver les caractéristiques sémantiques des textes de revue afin d'examiner l'impact des différentes fonctionnalités. L'objectif de leur travail est de comprendre pourquoi certains avis reçoivent de nombreux votes utiles, tandis que d'autres reçoivent peu ou pas de votes. Par conséquent, au lieu de prédire un niveau utile pour les avis qui n'ont pas de votes, ils ont étudié les facteurs qui déterminent le nombre de votes pour l'utilité qu'un avis particulier reçoit (incluez les votes «oui» et «non»). Ils ont travaillé sur les commentaires des utilisateurs de logiciels de CNET Download.com. Ils ont montré que les caractéristiques sémantiques ont plus d'influence que d'autres caractéristiques sur le nombre de critiques de votes d'utilité reçues.

2.2.2.2 Approche sémantique

L'approche sémantique donne directement des valeurs de sentiment et s'appuie sur différents principes pour calculer la similitude entre les mots. Ce principe donne des valeurs de sentiment similaires aux mots sémantiquement proches. Par exemple, WordNet fournit différents types de relations sémantiques entre les mots utilisés pour calculer les polarités des sentiments. WordNet pourrait également être utilisé pour obtenir une liste de mots de sentiments en élargissant de manière itérative l'ensemble initial avec des synonymes et des antonymes, puis en déterminant la polarité des sentiments pour un mot inconnu par le nombre relatif de synonymes positifs et négatifs de ce mot [92].

L'approche sémantique est utilisée dans de nombreuses applications pour construire un modèle de lexique pour la description des verbes, noms et adjectifs à utiliser en analyse des sentiments comme le travail présenté par Maks et Vossen [93]. Leur modèle décrit les relations de subjectivité détaillées entre les acteurs dans une phrase exprimant des attitudes distinctes pour chaque acteur. Ces relations de subjectivité sont étiquetées avec des informations concernant à la fois l'identité du porteur d'attitude et l'orientation (positive ou négative) de l'attitude. Leur modèle comprenait une catégorisation en catégories sémantiques pertinentes pour l'analyse des sentiments. Il a fourni des moyens pour identifier le titulaire de l'attitude, la polarité de l'attitude et aussi la description des émotions et des sentiments des différents acteurs impliqués dans le texte. Ils ont utilisé WordNet dans leur travail. Leurs résultats ont montré que la subjectivité du locuteur et parfois la subjectivité de l'acteur peuvent être identifiées de manière fiable.

Une autre approche proposée par Pai et Chu [94]. Les auteurs ont extrait des évaluations positives et négatives et ont aidé les consommateurs dans leur prise de décision. Leur méthode peut être utilisée comme un outil pour aider les entreprises à mieux comprendre les évaluations de produits ou de services et à traduire en conséquence ces opinions en intelligence d'affaires à utiliser comme base pour l'amélioration des produits/services. Ils ont travaillé sur des critiques de restauration rapide taïwanaise. Leurs résultats ont montré que leur approche est efficace pour fournir des évaluations liées aux services et produits.

2.2.2.3 Approche mixte

Les méthodes sémantiques peuvent être mélangées avec les méthodes statistiques pour effectuer la tâche d'analyse des sentiments comme le travail présenté par Zhang et Xu [95] qui ont utilisé les deux méthodes pour trouver la faiblesse du produit à partir des critiques en ligne. Leur chercheur de faiblesse a extrait les fonctionnalités et regroupé les fonctionnalités explicites en utilisant une méthode basée sur le morphème pour identifier les mots de fonctionnalité à partir des avis. Ils ont utilisé une mesure de similarité basée sur HowNet pour trouver les caractéristiques explicites fréquentes et peu fréquentes qui décrivent le même aspect. Ils ont identifié les caractéristiques implicites avec la méthode de sélection basée sur les statistiques de collocation PMI. Ils ont regroupé des produits comportant des mots dans des aspects correspondants en appliquant des méthodes sémantiques. Ils ont utilisé la méthode d'analyse des sentiments basée sur les phrases pour déterminer la polarité de chaque aspect dans les phrases en tenant compte de l'impact des adverbes de degré. Ils pouvaient trouver les faiblesses du produit, car c'était probablement l'aspect le plus insatisfait dans les avis des clients, ou l'aspect qui était plus insatisfait par rapport aux avis sur les produits de leurs concurrents. Leurs résultats ont exprimé la bonne performance du détecteur de faiblesse [25].

2.3 Approches hybrides

Dans cette section des techniques de classification des sentiments seront discutées, où les auteurs ont utilisé plus d'une technique d'apprentissage automatique et technique basée sur un Lexique, liées les unes aux autres. Ce type de techniques combinées pour effectuer la classification est appelé approche hybride. Quelques-unes sont mises en évidence comme ci-dessous:

Filho et al. [96] ont proposé un processus de classification hybride avec trois classificateurs différents, classificateur basé sur des règles, classificateur basé sur un lexique et classificateur de

machines à vecteurs de support (SVM). Ils ont utilisé une approche par pipeline pour la classification qui fonctionne par modèle d'interruption. Dans leur approche, chaque classificateur classe une revue jusqu'à ce qu'un certain niveau de confiance dans l'exactitude soit obtenu. Si le niveau est atteint, la classe de sentiment final est affectée à la révision; sinon l'examen est fourni au classificateur suivant. Si le niveau de précision n'est toujours pas atteint; puis le mécanisme de vote est utilisé pour la classification. Les auteurs ont utilisé SVM basé sur le noyau linéaire pour la classification ainsi que l'approche basée sur des règles et lexique. Ils ont utilisé cinq ensembles de données différents pour tester leur approche proposée, à savoir Twitter2013, SMS2013, Twitter2014, LiveJournal2014 et Twitter2014 Sarcasme. Leur approche proposée a montré une valeur d'exactitude de l'ordre de 53,31% et 65,39% sur l'ensemble de données Twitter2013 et Twitter2014 respectivement.

Zhao et Jin [97] ont proposé une approche hybride basée sur des étiquettes sémantiques qui combine une méthode basée sur la sémantique avec la technique SVM. Ils ont considéré chaque texte de révision (critique) comme une séquence de phrases sémantiques et ont obtenu deux étiquettes de sentiment potentiel pour chaque revue (avis, critique). Ils ont attribué le label hybride comme nouvelle fonctionnalité pour améliorer les performances de l'approche. Ils ont collecté les critiques des sites de critiques de films chinois tels que Mtive et DouBan film. Leur ensemble de formation contient 2000 critiques de films chinois tandis que l'ensemble de test contient 1000 critiques de films chinois.

Nandi et Agrawal [98] ont proposé une classification hybride des sentiments combinant l'approche basée sur le dictionnaire du lexique avec le résultat du classificateur SVM. L'approche lexicale repose sur le dictionnaire de mots, c'est-à-dire le sac de mots pour l'analyse et fonctionne sur le principe que la polarité du document est la somme de la polarité des mots ou des phrases individuels. Ils ont pris en compte les tweets de Twitter pour la classification. Ils ont rassemblé les tweets liés à la politique indienne pour la classification.

Desai et Mehta [99] ont proposé des algorithmes de classification hybrides pour l'analyse des problèmes et des avantages des étudiants. Ils ont combiné à la fois l'approche basée sur les connaissances et l'apprentissage automatique pour traiter le tweet. Ils ont collecté les tweets avec #engineringProblem ainsi que les hashtags #engineeringPerks et les ont considérés comme les ensemble de données pour leur analyse. Afin d'effectuer l'analyse basée sur les connaissances, un corpus est créé avec la collection de tous les tweets collectés. Ensuite, les lexiques sont trouvés

avec des valeurs d'opinion plus élevées à la fois pour la polarité positive et négative. Ces mots sont connus comme les mots semences, et tous les synonymes et antonymes possibles pour ces mots semences sont rassemblés pour former un dictionnaire de lexiques. Ce dictionnaire de lexiques est alimenté par l'approche d'apprentissage automatique qui le considère comme une entrée et en fonction de ces entrées, les tweets sont classés.

3 Les lexiques des sentiments

Comme l'indique Liu et Bing [100], les mots / expressions ont deux types de sentiments : absolu et relatif. Le sentiment absolu signifie que le sentiment reste le même, étant donné le bon mot/phrase et la bonne signification. Par exemple, le mot «beau» est un mot positif. Le sentiment relatif signifie que le sentiment change en fonction du contexte. Par exemple, le mot «augmenté» ou «alimenté» a un sentiment positif / négatif basé sur l'objet du mot.

Il existe une troisième catégorie de sentiment: le sentiment implicite. Le sentiment implicite est différent du sentiment absolu. Le sentiment implicite est le sentiment qui est couramment invoqué dans l'esprit d'un lecteur lorsqu'il lit ce mot/cette phrase. Prenons l'exemple des «parcs d'attractions». Un lecteur éprouve généralement un sentiment positif en lisant ce mot. De même, l'expression «se réveiller au milieu de la nuit» implique un sentiment négatif implicite.

Actuellement, la plupart des lexiques de sentiment se limitent aux mots de sentiment absolus. L'extraction de sentiments implicites dans des phrases forme une autre branche de travail. Nous nous en tenons également à cette définition et discutons des lexiques de sentiment qui capturent le sentiment absolu. Le développement précoce de lexiques de sentiments s'est concentré sur la création de dictionnaires de sentiments.

Stone et al. [101] présentent un lexique appelé «General Inquirer» qui a été largement utilisé pour l'analyse des sentiments. Finn [102] présente un lexique appelé AFINN. Comme General Inquirer, c'est aussi un lexique généré manuellement. Pour montrer la méthodologie générale sous-jacente aux lexiques de sentiment, nous décrivons quelques sentiments populaires lexiques dans les sous-sections à venir.

3.1 SentiWordNet

SentiWordNet, décrit en premier par Esuli et Sebastiani [3], est un lexique des sentiments qui complète WordNet [103] avec des informations sur les sentiments. L'étiquetage est flou et se fait

en ajoutant trois scores de sentiment à chaque synset dans WordNet comme suit. Chaque synsets a trois scores:

1. Pos (s): le score positif des synsets
2. Neg (s): le score négatif des synsets
3. Obj (s): le score objectif des synsets

Ainsi, dans SentiWordNet, le sentiment est associé à la signification d'un mot plutôt qu'au mot lui-même. Cette représentation permet à un mot d'avoir plusieurs sentiments correspondant à chaque sens. Puisqu'il y a trois scores, chaque signification en soi peut être à la fois positive et négative, ou ni positive ni négative.

3.2 Treebank des sentiments de Stanford

Ce Treebank a été introduit dans Socher et Richard [104]. Afin de créer le Treebank, le travail est également venu avec un lexique appelé le sentiment Treebank, qui est un lexique composé d'arbres d'analyse partielle annotés avec des sentiments. Le lexique a été créé comme suit. Un corpus de critiques de films a été obtenu à partir de www.rottentomatoes.com, composé de 10 662 phrases. Chaque phrase a été analysée à l'aide de l'analyseur de Stanford (Stanford Parser). Cela a donné un arbre d'analyse pour chaque phrase. Les arbres d'analyse ont été divisés en phrases, c'est-à-dire que chaque arbre d'analyse a été divisé en ses composants, chacun d'eux étant ensuite sorti sous forme de phrase. Cela a donné lieu à 215 154 phrases. Chacune de ces phrases a été étiquetée pour le sentiment à l'aide de l'interface d'Amazon Mechanical Turk. La sélection des étiquettes est également décrite dans le papier original. Au départ, la granularité des valeurs de sentiment était de 25, c'est-à-dire que 25 valeurs possibles pouvaient être données pour le sentiment, mais il a été observé à partir des données de l'expérience Mechanical Turks que la plupart des réponses contenaient l'une des 5 valeurs seulement. Ces 5 valeurs étaient alors appelées «très positives», «positives», «neutres», «négatives» et «très négatives».

3.3 SO-CAL

Le système SO-CAL (Sentiment Orientation CALculator) [105] est basé sur une ressource à faible couverture construite manuellement et composée de mots bruts. Contrairement à SentiWordNet, aucune information de sens n'est associée à un mot. SO-CAL utilise comme base une ressource de sentiment lexical composée d'environ 5000 mots. (En comparaison, SentiWordNet a plus de 38 000 mots polaires et plusieurs autres mots strictement objectifs.)

Chaque mot dans SO-CAL a une étiquette de sentiment qui est un entier dans [-5, +5] à part 0 car les mots objectifs sont simplement exclus. Les points forts de SO-CAL résident dans sa précision, car il est annoté manuellement, et l'utilisation de fonctionnalités détaillées qui gèrent les sentiments dans divers cas d'une manière conforme aux phénomènes linguistiques.

SO-CAL utilise plusieurs «fonctionnalités» pour modéliser différentes catégories de mots et leurs effets sur le sentiment. De plus, quelques fonctionnalités spéciales opèrent en dehors du champ d'application du lexique afin d'affecter le sentiment au niveau du document.

4 Utilisation de la sémantique dans l'analyse des sentiments

En raison de la nature de la tâche, la sémantique est évidemment un ingrédient crucial de tout système d'analyse des sentiments. Cependant, en fonction de la complexité du système, et aussi en fonction de la tâche spécifique qui est entreprise, des informations sémantiques de différents types peuvent être consultées et incorporées de différentes manières.

4.1 Informations lexiques

Indépendamment de l'approche (basée sur un lexique ou d'apprentissage automatique), pratiquement tous les systèmes d'analyse des sentiments reposent sur des informations dérivées de ressources lexicales. Dans l'apprentissage automatique, ces informations sont utilisées comme fonctionnalités, généralement en combinaison avec d'autres fonctionnalités. Certains systèmes reposent sur une seule ressource; par exemple, Günther et Furrer [106] n'utilisent que SentiWordNet, tandis que d'autres systèmes essaieront d'utiliser et de combiner les informations de toutes les sources disponibles [107].

La plupart des approches statistiques prennent en charge des modèles SVM qui implémentent de telles fonctionnalités et des fonctionnalités similaires, et Taboada et al. [108] soulignent que les informations des unigrammes de base semblent être les plus utiles. Bien que cela montre que les mots en eux-mêmes sont très instructifs pour cette tâche, les limites intrinsèques à l'utilisation des informations issues d'une simple recherche dans le dictionnaire sont tout à fait évidentes si l'on pense que les mots sont utilisés dans un contexte (syntaxique) spécifique et que leur polarité peut changer considérablement en fonction de la manière dont ils se rapportent aux autres mots du texte.

Une autre limitation cruciale de la non-prise en compte du contexte est la négation [109]. Pour faire face à cet aspect, Taboada et al. [108] incorporent des informations de ce qu'ils appellent des décaleurs de valence contextuels, montrant une augmentation des performances. Il s'agit d'un premier pas vers un traitement linguistique plus approfondi, considéré de plus en plus nécessaire, même pour des textes courts tels que les tweets.

4.2 La sémantique distributionnelle

Dans le contexte de l'analyse des sentiments, l'idée d'exploiter l'hypothèse distributionnelle - à savoir, l'hypothèse selon laquelle les mots qui apparaissent dans les mêmes contextes ont tendance à avoir des significations similaires [110] - se résume simplement au fait que les modèles de similarité qui prédisent pour, par exemple, le fait que «incroyable» et «merveilleux» soient similaires pourrait être étendu pour prédire que si «incroyable» a une valeur positive, il en sera de même pour « merveilleux », et « terrible » sera à l'autre bout du spectre. Cependant, les modèles de similitude générale prennent généralement en compte le contexte lexical d'un mot, mais pas nécessairement la polarité du texte. Par conséquent, la similarité est potentiellement exacte au niveau syntaxique et sémantique, mais pas nécessairement d'une manière sensible au sentiment, de sorte que «bon» et «mauvais» finissent par être très similaires. Ceci est vrai pour les modèles de similitude distributionnelle classiques ainsi que pour les représentations vectorielles distribuées plus récentes et réussies connues sous le nom de plongements de mots (Word embeddings), au moins dans leur formulation standard [111]. Comme première tentative d'incorporer directement le sentiment dans l'apprentissage du contexte distributionnel d'un mot, de sorte que les mots exprimant un sentiment similaire finissent par avoir des représentations vectorielles similaires, Maas et al. [112] ont développé un modèle où ils donnent une fonction de prédiction de sentiment à chaque mot. En ce qui concerne la classification positive/négative des tweets, leur système s'avère plus performant que les modèles qui intègrent des intégrations entraînées de manière non sensible aux sentiments. Un modèle encore plus puissant et récent, qui surpasse celui de Maas et al. [112], a été proposé par Tang et al. [113], qui forment un réseau de neurones en associant chaque n-gramme à la polarité d'une phrase, et montrer que les plongements de mots spécifiques au sentiment distinguent efficacement les mots avec une polarité de sentiment opposée. Ce modèle fonctionne mieux que les autres modèles qui utilisent des plongements généralement appris.

4.3 Entités, propriétés et relations

L'intérêt pour une analyse des sentiments plus fine a également nécessairement conduit à la nécessité d'une analyse sémantique plus fine, et donc d'un traitement du langage plus approfondi. Par exemple, l'analyse des sentiments basée sur les aspects doit reposer sur l'identification d'entités spécifiques et/ou de propriétés d'entités dans des critiques ou des tweets. Pour ce faire, des techniques standard de détection et de classification d'entités sont employées, telles que des marqueurs séquentiels, éventuellement recyclés pour des domaines spécifiques.

De plus, les relations entre les entités et les événements impliqués doivent être identifiées, afin de savoir ce qui est dit de quelle entité. Un moyen évident de le faire est d'exploiter les relations de dépendance, bien qu'un traitement plus approfondi des tweets ne soit pas si simple en raison du langage idiosyncratique et souvent non grammatical que contiennent ces textes courts (bien que des travaux récents basés sur l'apprentissage de l'intégration de graphes de connaissances neuronales montrent une réduction des erreurs de plus de 26% dans l'analyse sémantique des tweets [114]). Des problèmes similaires se posent lorsque l'on développe des systèmes pour détecter une position, comme pour évaluer l'opinion de quelqu'un envers une cible donnée, toutes les relations entre les entités impliquées doivent être correctement identifiées et associées au sentiment exprimé. Cependant, une analyse linguistique plus approfondie du texte est également bénéfique, sinon nécessaire, pour l'analyse standard des sentiments au niveau du message ou du texte, car elle aide à traiter la question des décaleurs de valence contextuels mentionnés dans la section 4.1 en tenant également compte de l'ordre des mots et de la structure de phrase.

À cette fin, le Natural Language Processing Group de l'université de Stanford a développé une banque d'arbres des sentiments. Cette banque d'arbres a été utilisée pour former un réseau de neurones récursif construit au-dessus de structures grammaticales [104], obtenant une augmentation de 5 points de pourcentage sur la classification de la polarité des phrases. Sur un niveau de sentiment à grain fin, ils ont obtenu une amélioration de 9,7% par rapport à une ligne de base de sac de mots, et ont globalement montré la capacité de capturer avec précision les effets de la négation et sa portée à différents niveaux dans les structures arborescentes.

5 Word embeddings

Word embeddings ou le plongement de mots sont des types d'algorithmes qui visent à représenter la signification des mots sous la forme de vecteurs, où les mots ayant une

signification et un contexte similaires sont représentés par des vecteurs similaires. Les plongements de mots sont considérés comme des ingrédients importants dans l'analyse des sentiments ainsi que dans d'autres tâches PNL, et ils servent de première couche de traitement de données dans les approches DL. Par conséquent, cette sous-section présente différente incorporations de mots couramment utilisées dans l'analyse des sentiments.

Les plongements de mots récents suivent une hypothèse de distribution, où les mots ayant le même contexte ont des significations similaires. Ainsi, les mots avec le même contexte ou une sémantique similaire créent des caractéristiques similaires et sont classés dans une classe.

Bengio et al. [115] a lancé l'intégration de mots en concevant un modèle de langage, qui apprend en même temps la représentation distribuée pour chaque mot et la fonction de vraisemblance pour les séquences de mots. Suite au succès de ce modèle, des approches extensives ont été suggérées pour améliorer les résultats et capturer des informations sémantiques et syntaxiques. Collobert et Weston [116] ont construit un modèle d'incorporation de mots pré-entraîné basé sur le modèle DL, qui peut apprendre les caractéristiques nécessaires à une tâche spécifique lorsque les informations antérieures ne sont pas suffisantes. Ces derniers travaux ont inspiré et jeté les bases de nombreuses études récentes dans le domaine.

Il existe différents algorithmes pour générer des vecteurs d'incorporation de mots qui ont été proposés et sont accessibles au public. Trois algorithmes parmi les plus célèbres sont discutés dans le paragraphe suivant.

5.1 Word2vec

Mikolov et al. [117] proposent word2vec pour la première fois en 2013, qui est le résultat de la combinaison du skipgramme et du continu bag-of-word (CBOW). Le modèle Skipgram prédit un mot cible en utilisant les mots apparaissant à proximité. En revanche, le modèle CBOW prédit le mot cible en fonction du contexte environnant. Le contexte est représenté à l'aide de la méthode du sac de mots qui sont contenus dans une fenêtre de taille spécifiée autour du mot cible.

5.2 GloVe

Un autre modèle de plongement de mots populaire GloVe signifie que le vecteur global a été développé par le groupe NLP de l'Université de Stanford [118]. C'est un algorithme d'apprentissage non supervisé basé sur un réseau neuronal profond pour l'apprentissage des vecteurs de mots. GloVe a été proposé après le Word2vec et a subi quelques changements par

rapport au modèle Skipgram. Le premier changement est que l'algorithme GloVe a adopté la perte de carré au lieu de la perte d'entropie croisée comme fonction objective pour former le modèle pour générer le mot incorporé. Deuxièmement, l'algorithme GloVe considère les informations statistiques globales des mots anglais en fonction de l'ensemble de données, tandis que l'algorithme Word2vec ne considère que les informations à l'intérieur des fenêtres de taille de correction.

5.3 FastText

Joulin et al. [119] ont proposé une méthode d'incorporation de mots FastText, qui est un autre modèle de réseau neuronal courant pour générer des vecteurs de mots précis. Une caractéristique du modèle FastText est que la vitesse d'entraînement est rapide. Dans les expériences de l'auteur, le modèle FastText n'a eu besoin que de 4 secondes pour compléter une époque pour effectuer la classification des sentiments sur l'ensemble de données de révision complète de Yelp. En revanche, d'autres modèles ont eu besoin de plus de 30 minutes pour terminer une époque. Cela montre que l'architecture du modèle du FastText était simple, qui contient une couche pour le mot d'incorporation, une couche cachée et une couche pour la sortie.

6 Conclusion

Dans ce chapitre, nous avons présenté la revue de littérature pour la classification des sentiments. Les approches utilisées sont basées sur l'apprentissage machine, sur un lexique ou hybride. Les travaux de la première approche se différencient par le type du classificateur utilisé, et par le choix des features. Les résultats de ces travaux montrent que pour que cette approche soit efficace, il faut que la base d'apprentissage soit conséquente et que le choix des features soit pertinent. Les travaux de la deuxième approche se basent sur des lexiques externes (généraux tel que SentiWorNet ou construit manuellement, tel que le lexique LF), ou des lexiques internes construits à partir du corpus (tel que le lexique HF). Ces lexiques contiennent des mots subjectifs. Certains travaux ne s'intéressent qu'aux adjectifs, d'autres ajoutent des adverbes, des noms et des verbes.

L'analyse des sentiments à l'aide d'approches d'apprentissage en profondeur a attiré un grand nombre de chercheurs. Par conséquent, une pléthore de modèles d'apprentissage en profondeur ont été proposés et se sont avérés donner de bons résultats sur diverses tâches d'analyse des

sentiments. Le succès des approches mentionnées est attribué à leur capacité d'apprentissage automatique des fonctionnalités et au succès des modèles d'incorporation de mots. Bien que les approches traditionnelles très populaires comme Naïve Bayes et SVM se soient révélés utiles pendant si longtemps, le potentiel du modèle d'apprentissage profond ne peut être négligé. En fait, ce dernier promet de bien mieux fonctionner que le premier, avec des contraintes minimales sur la tâche ou les données pour l’analyse des sentiments. Afin d’améliorer les performances, certains travaux combinent l’utilisation du lexique avec l’apprentissage profond.

Chapitre 3

Analyse des sentiments dans les réseaux sociaux

Sommaire

1 Introduction

La croissance exponentielle de l'utilisation des appareils numériques, associée à un accès en ligne omniprésent, offre un terrain sans précédent pour la connectivité constante des personnes et offre d'énormes capacités pour exprimer publiquement des opinions, des attitudes ou des réactions concernant de nombreux aspects des activités humaines quotidiennes. Les médias sociaux, tels que les blogs, les forums et les plateformes de réseaux sociaux (par exemple, Facebook, LinkedIn, Twitter, Instagram, YouTube) deviennent rapidement une partie intégrante de la vie des gens, les espaces virtuels où les individus quotidiens partagent des opinions et des informations et maintiennent et / ou étendent leur réseau relationnel.

De nos jours, le perfectionnement constant des outils analytiques offre un éventail plus riche d'opportunités pour analyser ces données à de nombreuses fins différentes. Les différences dans les fonctionnalités et les caractéristiques des réseaux sociaux se reflètent dans l'énorme quantité de statistiques et de mesures différentes qu'il est possible de suivre et d'analyser. Les métriques les plus adoptées sont numériques, relativement faciles à obtenir et disponibles gratuitement, comme les métriques d'engagement et d'influence. Cependant, les métriques de ce type sont souvent définies comme des «métriques de vanité», puisqu'elles n'interprètent pas les données collectées. Pour cette raison, d'autres types de méthodes d'analyse ont été introduits. Parmi celles-ci, l'une des plus utilisées est celle de l'analyse des sentiments (AS), qui consiste à analyser les sentiments (les opinions, les émotions et les attitudes) derrière les mots à l'aide d'outils de traitement du langage naturel. L'AS est considérée comme une métrique de qualité, qui regarde derrière les chiffres pour comprendre comment les informations sur les émotions et les attitudes sont véhiculées dans le langage. Compte tenu de l'intérêt croissant pour l'application de l’analyse des sentiments aux données des réseaux sociaux, la recherche dans ce domaine a reconnu les limites liées à la gestion des caractéristiques complexes du langage naturel sans considérer les données collectées via les réseaux sociaux. La plupart des travaux de SA sont basés simplement sur des informations textuelles exprimées dans des publications et des commentaires en ligne. Les premières approches pour surmonter cette limitation importante émergent dans la littérature récente, essayant, par exemple, de tirer parti des informations sur les relations d'amitié entre les individus, car les utilisateurs connectés peuvent être plus susceptibles d'avoir des opinions similaires. Cependant, ces caractéristiques ne ressemblent qu’aux riches structures de relation

encodée dans un réseau social. Parmi les méthodes analytiques complémentaires possibles qui commencent à être introduites dans l'analyse des données collectées via les réseaux sociaux, l'une des plus intéressantes est l'analyse des réseaux sociaux (SNA), qui, grâce à une approche quantitative et relationnelle, permet de considérer des données relationnelles (ie, les connexions et liens existants entre utilisateurs sur les réseaux sociaux). Dans ce contexte, ce chapitre définira d'abord les réseaux sociaux et décrira brièvement leur histoire, en soulignant les différences et les caractéristiques spécifiques qui les caractérisent. Ensuite, l'analyse des réseaux sociaux sera discutée dans le cadre de l'analyse de sentiments avec un accent particulier sur des techniques d'analyse du sentiment dans les réseaux sociaux. Enfin, le chapitre soulignera comment l'analyse des réseaux sociaux peut être efficacement intégrée dans les approches de l'analyse de sentiments pour renforcer leur fiabilité et validité.

2 Définition des réseaux sociaux

Les réseaux sociaux sont des services basés sur le web dont la fonctionnalité principale est de connecter des personnes ou des entités. Ils sont utilisés pour référer à des différents services en ligne qui sont associés à une catégorie générale de situations d'interactions sociale et professionnelle.

Selon Garton et al. [120], un réseau social est définit comme "un ensemble d'individus, d'organisations ou d'entités entretenant des relations sociales fondées sur l'amitié, le travail collaboratif et l'échange d'information".

Danah et al. [121] définit le réseau social comme un système dans lequel ; (a) les utilisateurs sont des entités de premier ordre avec un profil semi-public, (b) les utilisateurs peuvent créer des liens explicites vers d'autres utilisateurs ou éléments de contenu, et (c) les utilisateurs peuvent naviguer sur le réseau social en parcourant les liens et les profils d'autres utilisateurs.

Les réseaux sociaux servent un certain nombre d'objectifs, mais trois rôles principaux se démarquent comme étant communs à tous les sites.

- Premièrement, les réseaux sociaux sont utilisés pour maintenir et renforcer les liens sociaux existants ou pour établir de nouveaux liens sociaux. Les sites permettent aux utilisateurs «d'articuler et de rendre visibles leurs réseaux sociaux», et ainsi de «communiquer avec des personnes qui font déjà partie de leur réseau social étendu ».

- Deuxièmement, les réseaux sociaux sont utilisés par chaque membre pour télécharger son propre contenu. Notez que le contenu partagé varie souvent d'un site à l'autre et qu'il ne s'agit parfois que du profil de l'utilisateur lui-même.
- Troisièmement, les réseaux sociaux sont utilisés pour trouver du nouveau contenu intéressant en filtrant, recommandant et organisant le contenu téléchargé par les utilisateurs.

Afin de comprendre le fonctionnement du monde social au sein de ces plateformes, l'analyse des réseaux sociaux, est fondée sur une vision structurale s'attachant à décrire les interdépendances entre les acteurs afin de simplifier leur représentation. La capacité à représenter de façon simplifiée la complexité d'un système social représente la force de cette analyse structurale. En effet, les réseaux sociaux sont des systèmes complexes ayant de nombreux éléments en interaction qui sont essentiellement les utilisateurs, les communautés et les contenus générés.

3 Un bref historique

Nous donnons maintenant un bref historique des réseaux sociaux. Le site Classmates.com est considéré comme le premier site Web permettant aux utilisateurs de se connecter à d'autres utilisateurs. Il a commencé en 1995 en tant que site permettant aux utilisateurs de se reconnecter avec d'anciens camarades de classe et compte actuellement plus de 40 millions d'utilisateurs enregistrés. Cependant, Classmates.com n'a pas permis aux utilisateurs de créer des liens vers d'autres utilisateurs; il permettait plutôt aux utilisateurs de se connecter les uns aux autres uniquement via les écoles qu'ils avaient fréquentées. En 1997, le site SixDegrees.com est le premier site qui répond à la définition d'un réseau social en ligne d'en haut qui a été créé. Il était le premier site de réseau social qui permettait aux utilisateurs de créer des liens directement vers d'autres utilisateurs.

Les réseaux sociaux en ligne ont commencé à gagner en popularité à mesure que de plus en plus d'utilisateurs se connectaient à Internet. Au début des années 2000, un certain nombre de sites à usage général pour trouver des amis ont été créés, dont le plus notable est Friendster. Friendster se concentrait sur le fait de permettre aux amis d'amis de se rencontrer, en commençant par un rival du site de rencontres en ligne Match.com. D'autres sites similaires créés dans le même laps de temps incluent CyWorld, Ryze et LinkedIn.

En 2003, MySpace a été créé comme une alternative à Friendster et les autres. MySpace a permis aux utilisateurs de personnaliser fortement l'apparence de leur profil, qui s'est avéré très populaire auprès des utilisateurs, faisant de MySpace rapidement de devenir le plus grand réseau social en ligne.

En 2004 Facebook, commence de devenir en quelques années le réseau social le plus connu et le plus utilisé au monde, a été initialement créé par Mark Zuckerberg comme un outil pour connecter les étudiants de l'Université Harvard. L'idée était de développer un réseau social pour soutenir une communauté fermée, en créant la version en ligne de l'annuaire universitaire. Ensuite, Facebook a été étendu pour connecter toutes les universités américaines, et a finalement été ouvert à tous, jusqu'à sa propagation en tant que phénomène de masse, alors qu'il en comptait maintenant plus d'un milliard d'utilisateurs.

Un autre jalon dans l'histoire des réseaux sociaux en ligne est le 14 février 2005: Chad Hurley, Steve Chen et Jawked Karim, trois jeunes employés de PayPal, ont enregistré le nom de domaine YouTube et le 23 avril de la même année, la première vidéo - Moi au zoo - a été téléchargée. Seulement 10 mois après le lancement officiel, cette plateforme avait déjà établi un record: en 2006, YouTube avait en moyenne 65 000 téléchargements de vidéos par jour et 20 millions d'accès uniques. Le succès n'est pas passé inaperçu et, en octobre 2009, YouTube a été vendu à Google pour 1,65 milliard de dollars. Un an seulement après la création de YouTube, en 2006, Twitter a été créé. Ses créateurs Jack Dorsey, Evan Williams et Biz Stone ont voulu créer un système capable de permettre aux gens de communiquer par SMS avec un petit nombre d'amis. Twitter a donc été développé comme un système de microblogage, un service d'échange d'informations, qui permet aux utilisateurs d'envoyer des messages (un tweet) de 140 caractères. La principale raison de la publication d'un tweet étant le partage, Twitter a développé un système pour élargir le public cible: le hashtag. Cette fonctionnalité facilite la possibilité de suivre des sujets et des fils d'intérêt: si un mot est précédé du symbole # (en anglais «hash»), alors cliquer dessus conduit au résultat.

Tous ces sites ont des objectifs différents mais utilisent la stratégie commune d'exploiter le réseau social pour améliorer leurs sites. La liste ci-dessus ne se veut pas exhaustive, car de nouveaux sites sont régulièrement créés.

4 L'analyse des réseaux sociaux

L'analyse des réseaux sociaux (SNA) consiste essentiellement en une série de techniques mathématiques et informatiques qui en utilisant les théories des réseaux et des graphes, peuvent être utilisées pour comprendre la structure et la dynamique de réseaux réels ou artificiels [122].

L'analyse des réseaux sociaux adopte une approche quantitative-relationnelle, plutôt que de s'appuyer sur les caractéristiques et attributs des individus (par exemple, le nombre de messages envoyés et reçus), et est basée sur des données relationnelles (des liens ou des contacts) qui caractérisent un groupe de personnes ou un ensemble d'organisations de complexité variable (par ex., familles, groupes d'amis, associations). Les relations sont représentées par des interactions de diverses natures (amitié, bisness,..). La potentialité de l'analyse des réseaux sociaux est essentiellement double: l'application de la théorie des graphes aux relations de données et, par conséquent, la description de la structure de l'interaction au moyen d'indices mathématiques-algébriques [123].

Les réseaux sociaux sont généralement représentés par des graphes, qui ont l'avantage de donner une image claire et immédiate de la structure sociale. Les graphes sont la structure mathématique d'un sociogramme, exprimée visuellement sous la forme d'un réseau composé de nœuds connectés. Les graphes sont utiles car ils représentent graphiquement les relations sociales entre les individus et surtout en fournissent une représentation formelle (Figure 21). De plus, il est possible de calculer un indice pour décrire des dimensions structurelles spécifiques, telles que la densité, l'inclusion et la cohésion.

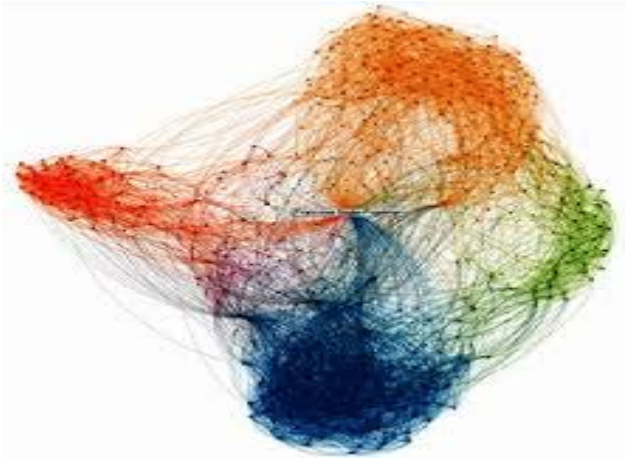


Figure 21- Réseau des liens LinkedIn Un ensemble de sous-composants est clairement observable

5 Analyse du sentiment dans les réseaux sociaux

Depuis le début des années 2000, l'analyse des sentiments, également appelée extraction d'opinion, est l'un des domaines de recherche les plus actifs dans le traitement du langage naturel. L'analyse des sentiments est définie comme le processus d'identification des opinions exprimées dans un morceau de texte comme positives, négatives ou neutres [124] .

La tendance générale des recherches sur l'analyse des sentiments dans les réseaux sociaux (SNA) est d'appliquer les techniques héritées de l'analyse traditionnelle des sentiments. Cependant, compte tenu de l'évolution des sources où les opinions sont exprimées, les stratégies disponibles dans l'état actuel de la technique ne sont plus efficaces pour les opinions minières dans ce nouvel environnement difficile.

En réalité, L'analyse des sentiments des réseaux sociaux, en plus d'hériter d'une multitude de problèmes de l'analyse traditionnelle des sentiments et du traitement du langage naturel, introduit d'autres complexités (messages courts, contenu bruyant, métadonnées telles que le sexe, l'emplacement et l'âge) et de nouvelles sources d'informations non exploitées dans les approches traditionnelles.

En particulier, les réseaux sociaux ont clairement un impact sur le langage, les défis quotidiens en matière d'analyse des sentiments se concentrent principalement sur l'évolution constante du langage utilisé en ligne dans les contenus générés par les utilisateurs: les mots qui nous entourent chaque jour influencent les mots que nous utilisons. Le langage utilisé dans les réseaux sociaux pour communiquer entre nous a tendance à être plus malléable que l'écriture formelle, la combinaison de la communication informelle et personnelle et l'audience de masse offerte par les réseaux sociaux qui est une recette pour un changement rapide. Prenant en considération sérieusement la révolution continue du langage, les systèmes d'analyse des sentiments devraient pouvoir s'y adapter nativement, ou bien être adaptés par les chercheurs. En tant qu'effet secondaire, cette évolution du langage influence fortement la manière dont l'ironie et le sarcasme sont prononcés.

6 Techniques d'analyse du sentiment dans les réseaux sociaux

Les propriétés du lien entre les individus sur les réseaux sociaux en ligne sont essentielles pour comprendre le processus d'influence sociale à travers eux. Ce concept est bien représenté par la construction sociologique de la force des liens, qui représente la force des relations interpersonnelles dyadiques dans le contexte des réseaux sociaux [125], et qui a un impact sur les flux d'information. Dans les scénarios de réseaux sociaux en ligne, des fonctionnalités spécifiques sont associées à des bords entre deux personnes, telles que les commentaires qu'ils ont faits l'un sur l'autre ou les messages qu'ils se sont échangés. Ces caractéristiques comportementales peuvent contenir un signal de sentiment fort, ce qui est utile pour prédire les signes de bord et peut être utilisé pour s'adapter à un modèle de sentiment conventionnel. Un modèle de sentiment purement basé sur les fonctionnalités de bord ne peut pas prendre en compte la structure du réseau car il raisonne sur les bords comme étant indépendants les uns des autres.

Des études récentes ont tenté de considérer conjointement l'analyse de sentiments et l'analyse des réseaux sociaux afin de donner de meilleures prédictions que l'une ou l'autre ne peut à elle seule donner. Nous en discutons dans la suite quelques-uns.

Thomas et al. [126], en particulier, ont utilisé l'affiliation à un parti et les mentions dans les discours pour prédire les modèles de vote à partir des transcriptions des débats au Congrès Américain. Ils ont montré que l'intégration d'informations, même très limitées, concernant les relations inter-documents peut considérablement augmenter la précision de la classification soutien/opposition. L'intégration des informations sur les accords ne présente un avantage supplémentaire que lorsque les documents d'entrée sont relativement difficiles à classer individuellement.

Bermingham et al. [127] ont essayé de combiner AS et SNA pour explorer le potentiel de radicalisation violente en ligne. En particulier, grâce à une analyse détaillée d'un véritable ensemble de données YouTube, ils ont développé un modèle qui synthétise des informations textuelles et des réseaux sociaux pour prédire conjointement la polarité (positive ou négative) des évaluations de personne à personne. Plus précisément, ils ont incorporé dans leur modèle des mesures de travail intranet (c.-à-d. Centralité et entre-deux) et des analyses de l'ensemble du réseau (densité et vitesse de communication moyenne). En adoptant leur méthode de notation de

polarité basée sur un dictionnaire pour attribuer des scores de positivité et de négativité aux profils et aux commentaires YouTube, ils ont pu caractériser les utilisateurs et les groupes d'utilisateurs par leur sentiment envers un ensemble de concepts qui présentaient un intérêt particulier pour les terroristes.

Tan et al. [128] ont utilisé des informations sur les relations entre les utilisateurs de Twitter (par exemple, les suivis et les mentions) pour améliorer l'AS afin de prédire les attitudes à l'égard des événements politiques et sociaux. Travailler au sein d'un cadre semi-supervisé, ils ont proposé des modèles induits d'un réseau (Twitter par exemple) formé par des utilisateurs se référant les uns aux autres à l'aide de @ - mentions. Leurs résultats ont révélé que l'incorporation d'informations sur les réseaux sociaux peut en effet conduire à des améliorations statistiquement significatives de la classification des sentiments par rapport à la performance d'une approche basée sur des machines vectorielles de support n'ayant accès qu'à des fonctionnalités textuelles. Plus en détail, ils ont constaté que :

(1) la probabilité que deux utilisateurs partagent la même opinion est en effet corrélée avec leur connexion au réseau social.

(2) l'utilisation de modèles graphiques intégrant des informations sur les réseaux sociaux peut conduire à des améliorations statistiquement significatives de classification de la polarité des sentiments au niveau de l'utilisateur par rapport à une approche utilisant uniquement des informations textuelles.

Ma et al. [129], ont proposé deux méthodes de recommandation sociale qui utilisent l'information sociale pour améliorer la précision des prévisions des systèmes de recommandation traditionnels. Plus précisément, les informations du réseau social ont été utilisées dans la conception de deux termes de régularisation sociale pour contraindre la fonction d'objectif de factorisation matricielle. De plus, les amis aux goûts différents ont été traités différemment dans les termes de régularisation sociale afin de représenter la diversité gustative des amis de chaque utilisateur. L'analyse expérimentale sur deux grands ensembles de données (un ensemble de données contenait un réseau d'amis sociaux, tandis que l'autre ensemble de données contenait un réseau de confiance sociale) a montré que ces méthodes proposées surpassent d'autres algorithmes de pointe.

Sperious et al. [130] ont exploré la possibilité d'exploiter le graphe d'abonnés Twitter pour améliorer la classification de polarité, sous l'hypothèse que les gens s'influencent les uns les autres ou ont des affinités partagées par rapport aux sujets. Plus précisément, ils ont proposé d'incorporer des étiquettes d'un classificateur d'entropie maximale, en combinaison avec le graphe des abonnés Twitter. Les utilisateurs (abonnés) ont été utilisés comme fonctionnalités distinctes et combinés avec la matrice de contenu. Ils ont construit un graphe qui a des utilisateurs, des tweets, des unigrammes de mots, des bigrammes de mots, des hashtags et des émoticônes comme nœuds; les utilisateurs sont connectés sur la base du graphe d'abonnés Twitter aux tweets qu'ils ont créés, et les tweets sont connectés aux unigrammes, bigrammes, hashtags et émoticônes qu'ils contiennent. Sperious et al. ont comparé l'approche de propagation d'étiquettes avec le classificateur supervisé bruyamment lui-même et avec une méthode standard basée sur le lexique utilisant des rapports positifs / négatifs sur plusieurs ensembles de données de tweets qui avaient été annotés pour la polarité. Ils ont montré qu'un classificateur d'entropie maximale formé avec une supervision à distance fonctionne mieux qu'un prédicteur de rapport basé sur le lexique, améliorant la précision de la classification de polarité de 58,1% à 62,9%. En utilisant les prédictions de ce classificateur en combinaison avec un graphe qui incorpore des tweets et des caractéristiques lexicales, ils ont obtenu une précision encore meilleure de 71,2%.

De même, Hu et al. [131] ont proposé une formulation d'optimisation mathématique intégrant la cohérence des sentiments et les théories sociologiques de la contagion émotionnelle pour la classification des sentiments. Ils ont utilisé une méthode appelée « approche sociologique de la gestion des textes bruyants et courts (SANT) », qui a extrait les relations de sentiment entre base des théories sociales, et modélisé les relations en utilisant une matrice de graphe laplacienne. Ils ont signalé que la méthode proposée peut utiliser les relations de sentiment entre les messages pour faciliter la classification des sentiments et gérer efficacement les données Twitter bruyantes. Une étude empirique de deux ensembles de données Twitter dans le monde réel a montré les performances supérieures du cadre adopté dans la gestion des tweets bruyants et courts, et SANT atteint des performances cohérentes pour différentes tailles de données d'entraînement.

Sur la base d'un autre principe sociologique fondamental, celui de l'homophilie, des travaux récents ont tenté d'inclure des caractéristiques de connexion des utilisateurs pour prédire les attitudes face aux événements politiques et sociaux à l'aide de méthodes et d'index SNA.

Pozzi et al. [132] ont déclaré que la prise en compte des liens d'amitié est une hypothèse faible pour modéliser l'homophilie: en ligne, comme hors ligne, deux amis pourraient ne pas partager la même opinion sur un sujet donné. Partant de cette critique, ils ont proposé une méthode alternative pour représenter l'homophilie; c'est-à-dire, un utilisateur qui approuve quelque chose (par exemple, par «aime»). Un cadre semi-supervisé a été utilisé pour estimer les polarités des utilisateurs sur un sujet donné en combinant le contenu des articles et les relations d'approbation pondérées sur les microblogs (Twitter). L'étude a montré que l'intégration des relations d'approbation de manière significative a surpassé l'approche basée uniquement sur le texte, conduisant à des améliorations significatives par rapport aux performances des classificateurs supervisés complexes basés uniquement sur des caractéristiques textuelles.

7 Conclusion

De nos jours, les réseaux sociaux ont émergé en supportant un large éventail d'utilisateurs et d'intérêts dans le monde entier et en incitant les chercheurs à exploiter les données sociales numériques. Plusieurs travaux ont profité du phénomène d'homophilie dans la résolution des problèmes d'analyse des sentiments. Cependant, peut de travaux ont exploité le phénomène d'influence sociale qui est aussi crucial dans les réseaux sociaux. C'est l'objet du chapitre suivant qui va nous permettre d'aborder ces différentes méthodes.

Chapitre 4
Les communautés de sentiments

Sommaire

1 Introduction

Dans ce chapitre, nous prendrons un nouveau concept, appelé «communauté de sentiments», pour étudier les sentiments et les relations des utilisateurs sur les réseaux sociaux. Une communauté de sentiments représente un groupe d'utilisateurs étroitement connectés et partageant les mêmes idées. Les utilisateurs qui appartiennent à la même communauté sentimentale non seulement se connectent étroitement les uns aux autres, mais ont également un sentiment presque constant sur un produit / service spécifique. La détection de la communauté des sentiments est un outil efficace pour les gestionnaires pour découvrir les préférences des clients et les interactions sociales pour la segmentation des clients. Par rapport aux communautés basées uniquement sur la structure de périphérie, les communautés de sentiments peuvent aider à identifier des groupes d'utilisateurs hautement connectés ayant des sentiments similaires sur un produit. Ces communautés de sentiments détectés représentent souvent des segments de marché de clients ayant des sentiments communs et des structures cohérentes. En exploitant ces résultats, les entreprises peuvent identifier des communautés de sentiments positifs comme des clients fidèles pour le marketing cible sous la forme de nouvelles recommandations de produits et de promotions. Des études supplémentaires devraient être menées pour étudier les caractéristiques démographiques de ces clients «hardcore» et les plus précieux afin de guider la conception future de produits / services. Identifier les communautés de sentiments négatifs permet également aux entreprises de mieux planifier leur campagne de marketing en offrant des rabais ou des fonctionnalités supplémentaires. Les utilisateurs de ces communautés négatives ont tendance à exprimer leurs sentiments honnêtes. Pour les produits achetés une seule fois, comme un film, bien qu'il ne soit pas réaliste de repenser le même produit et de persuader les consommateurs de l'acheter à nouveau, découvrir les communautés de sentiments a encore de nombreuses valeurs.

(1) Les communautés de sentiments représentent les différents goûts des consommateurs, et les vendeurs peuvent recommander des produits similaires aux communautés ayant des opinions positives et éviter de recommander des produits similaires à des communautés avec des opinions opposées.

(2) Sur la base des communautés de sentiments découvertes, les entreprises peuvent concevoir et publier différents nouveaux produits pour différentes communautés de sentiments.

2 Définitions et concepts des problèmes

Dans cette section, pour utiliser pleinement le facteur de sentiment pour trouver des structures communautaires plus significatives et plus précieuses, nous présentons les nouvelles définitions comme suit.

Définition 1 (Communauté): Dans une représentation graphique des réseaux sociaux, une communauté est un sous-graphe de nœuds dont les liens internes ont une densité plus élevée, tandis que les liens entre communautés ont une densité plus faible. Parfois, une communauté est également appelée clique ou cluster [135].

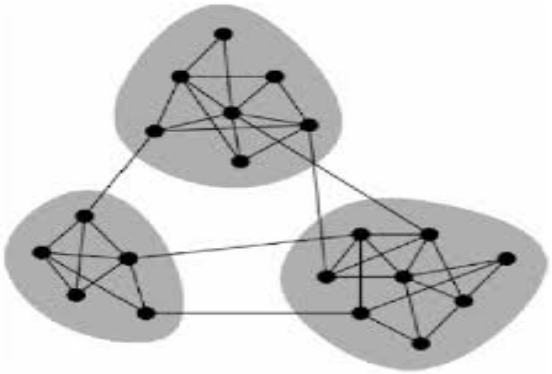


Figure 22- Un exemple de graphe de communautés dans un réseau social [134]

Définition 2 (Communauté de sentiments): Une communauté de sentiments est composée d'un groupe de personnes, qui sont directement ou indirectement connectées, et partagent des sujets de sentiment avec certains membres de ce groupe [133].

En fait, certaines communautés ont des sous-groupes qui peuvent partager des sujets de sentiment même s'il n'y a pas de lien réel entre eux, ce qui est très courant sur les réseaux sociaux. En fait, on peut considérer qu'il existe des liens virtuels entre ces sous-groupes. La **figure 23** illustre un exemple de réseau avec des communautés sociales.

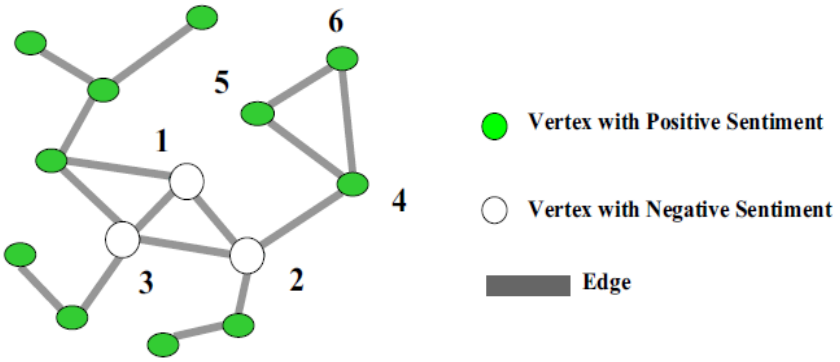


Figure 23- Un exemple de graphe de communautés de sentiments [135]

Définition 3 (Détection de communauté): La détection de communauté est le processus par lequel les nœuds des réseaux sont regroupés en fonction des connexions entre eux. En appliquant un algorithme de détection de communauté à un graphe représentant un tel réseau, on obtient un ensemble de communautés que nous appelons la structure de communauté du graphe [136].

***Problématique* :**

De nombreux algorithmes de détection communautaire ont été développés, basés à la fois sur les structures de réseau et les attributs de nœud [150], qui peuvent compléter les structures de réseau pour trouver des communautés plus précises. Ces méthodes existantes tentent de trouver des clusters de nœuds avec une connectivité interne élevée et une connectivité externe faible. Cependant, ces méthodes existantes n'ont pas pris en compte le sentiment des utilisateurs dans la détection des communautés. En analysant uniquement la connectivité des réseaux sans tenir compte des sentiments des utilisateurs, ces méthodes ne sont pas en mesure de séparer les utilisateurs qui interagissent activement les uns avec les autres mais qui ont des sentiments opposés. En revanche, l'identification de groupes d'utilisateurs qui sont étroitement liés à la fois par des relations sociales et des sentiments sera d'une plus grande valeur pour les entreprises pour la segmentation de la clientèle et le marketing cible sur SNS.

3 Utilisation de la détection de communauté pour l'analyse des sentiments

Dans les réseaux sociaux, les communautés sont considérées comme des structures importantes sous la forme de groupes d'entités (personnes). Il existe de nombreuses applications du monde réel pour la découverte de communautés. Par exemple, de nombreux supermarchés et centres commerciaux proposent des plateformes de communication client en ligne. Les clients peuvent non seulement publier leurs propres avis sur les produits et services, mais également échanger leurs opinions et leurs sentiments avec les autres. Afin de comprendre ce qui intéresse les clients et leur sentiment, les responsables aimeraient identifier des groupes d'utilisateurs, en particulier certains groupes avec des sentiments envers certains sujets. Plus précisément, ils aimeraient avoir une vue d'ensemble de leurs profils sociaux, des communautés auxquelles ils appartiennent, des sujets dont ils ont discuté, du sentiment envers certains sujets et des principales personnes avec lesquelles ils ont contacté dans chaque communauté. Aussi, l'identification de groupes

d'utilisateurs qui sont étroitement liés à la fois par des relations sociales et des sentiments sera d'une plus grande valeur pour les entreprises pour la segmentation de la clientèle, ce qui leur permet de connaître l'évolution de leurs communautés et des personnes qu'elles ont aliénées.

4 Classifications des méthodes de détection de communautés

Dans la littérature d'analyse des réseaux sociaux, de nombreux travaux ont été consacrés à la recherche de communautés sociales dans des réseaux sociaux. Pour détecter des communautés à partir des réseaux sociaux, y compris des informations sur les sentiments, un modèle de détection efficace est nécessaire. Les méthodes de détections des communautés ont été largement étudiées pendant de nombreuses années et deux catégories les plus importantes sont les méthodes traditionnelles qui reposent uniquement sur les liens au sein des réseaux. Cependant, le contenu sémantique peut en outre être utilisé pour partitionner les réseaux en communautés. Ces dernières années, l'identification des communautés en intégrant des liens sociaux et du contenu sémantique a été étudiée .Un état de l'art des méthodes de détection de la communauté des sentiments est discuté dans cette section :

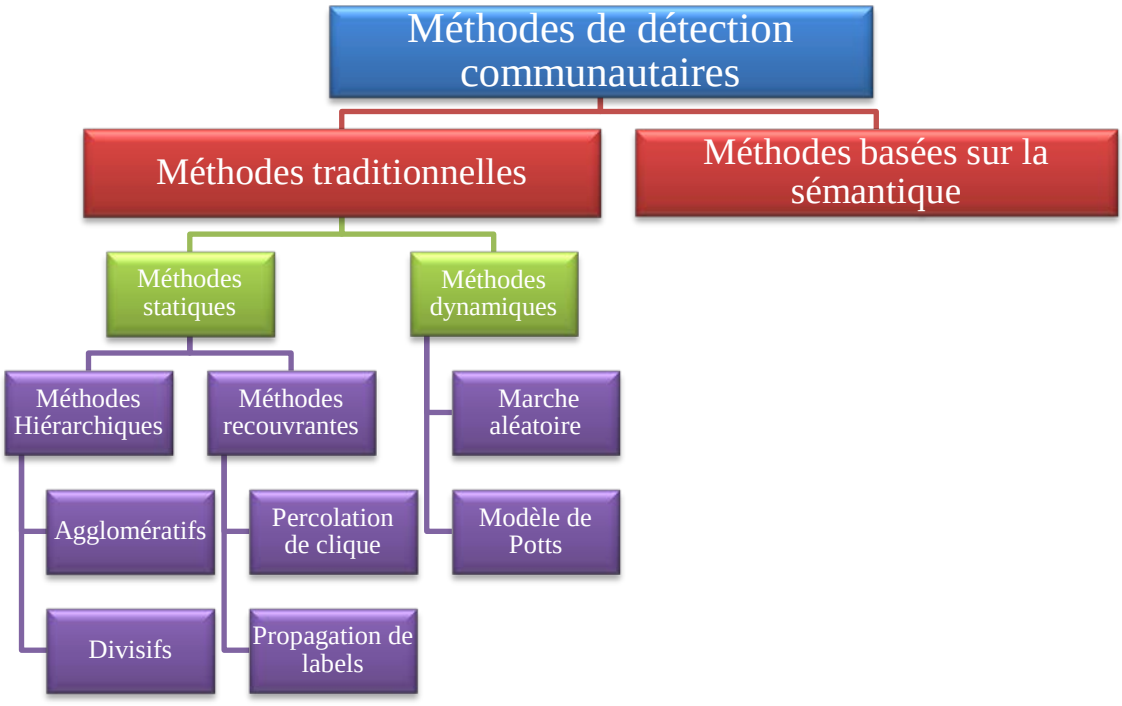


Figure 24- Méthodes de détection de communautés

4.1 Méthodes traditionnelles

Les méthodes traditionnelles peuvent être globalement classées en méthodes statiques et dynamiques. Les méthodes statiques tentent de révéler ou d'étudier la structure d'une communauté à partir de l'état actuel du réseau. Aucun effort n'est fait pour étudier les caractéristiques évolutives ou les changements de comportement qui peuvent survenir avec le temps. Les méthodes dynamiques, quant à elles, visent à étudier comment les communautés se forment, évoluent et meurent.

4.1.1 Méthodes statiques

Les méthodes de détection de communauté statique tentent de découvrir les communautés à partir de l'instantané actuel du réseau et ne prennent pas en compte la perspective historique ou évolutive des communautés. Ces méthodes ne sont pas évolutives car elles ne peuvent pas prédire avec précision le comportement / l'évolution des membres de la communauté car elles souffrent de la limitation des informations évolutives manquantes. Les méthodes statiques peuvent être globalement classées en deux classes :

- Méthodes Hiérarchiques
- Méthodes recouvrantes

4.1.1.1 Méthodes Hiérarchiques

Cette méthode tente de révéler une structure à plusieurs niveaux / hiérarchique dans un graphe. Le point de départ de l'algorithme de regroupement hiérarchique est la définition d'une mesure de similarité et les sommets sont regroupés en fonction de la présence/absence de cette mesure de similarité. Si les communautés sont fusionnées de manière itérative, sur la base d'une mesure de similarité à partir d'un seul sommet, l'algorithme est dit agglomératif. Cependant, si l'algorithme tente de trouver des communautés, en supprimant les arêtes entre sommets qui sont moins similaires, on parle d'algorithme divisif.

i. **Les algorithmes agglomératifs :** ces classes d'algorithmes regroupent les sommets itérativement en communautés qui sont fusionnés deux à deux en fonction de la présence d'une mesure de similarité jusqu'à avoir une grande communauté représentant l'ensemble des sommets du graphe. L'algorithme agglomératif glouton de Newman [151] a été le premier algorithme suggéré pour l'optimisation de la modularité, où initialement, chaque

sommet appartient à un module distinct, puis ils sont fusionnés de manière itérative en fonction du gain de modularité. Il a une complexité temporelle de $O(n)^3$ sur des réseaux clairsemés. Une version rapide de l'algorithme de Newman [151], implémentée par des structures de données efficaces avec une complexité de calcul de $O (n\log^2 n)$ sur des réseaux clairsemés.

Un autre algorithme glouton rapide pour la détection de communautés est proposé par Blondel et al. [152] qui est une méthode heuristique, repose également sur l'optimisation de la modularité. Il assigne différentes communautés à chaque sommet; un par sommet. Il fusionne de manière itérative les sommets en fonction du gain de modularité. Si aucun gain, alors le sommet reste dans sa propre communauté. La procédure est répétée jusqu'à ce que plus aucune amélioration ne soit possible. Il reconstruit ensuite le réseau de la manière dont les communautés identifiées dans la première phase sont remplacées par des super sommets. Sa complexité temporelle est O (n log n).

ii. **Les algorithmes de divisifs:** ces classes d'algorithmes partent du graphe d'un réseau et supprime récursivement les arêtes sur la base d'une faible similitude. Certains des principaux algorithmes de la catégorie sont l'algorithme de Girvan-Newman [153] où les arêtes sont supprimées de manière itérative en fonction du score entre les arêtes appelée centralité d'intermédiarité et l'algorithme Radicchi et al. [154], où la fonction de score d'une arête est calculée en divisant le nombre de triangles construits par cette arête sur le nombre maximum de triangles possibles et les arêtes sont supprimées de manière itérative en fonction du coefficient le plus faible.

4.1.1.2 Méthodes recouvrantes

Dans les réseaux sociaux, la plupart des sommets peuvent appartenir simultanément à plusieurs communautés. Les techniques hiérarchiques de détection des communautés ne parviennent pas à identifier les communautés qui se chevauchent. La percolation de clique et la propagation de label sont les techniques les plus connues et utilisées pour l'identification des communautés qui se chevauchent dans les réseaux.

i. **Techniques de percolation de cliques:** L'idée de ces algorithmes est de créer des communautés à partir de k-cliques qui correspondent à un sous-graphe complet (entièrement connecté) de **k** sommets. Deux cliques sont considérées comme adjacentes

si elles partagent des **k-1** sommets. Cfinder proposé par Palla [155] repose sur la construction d'un graphe à partir de l'ensemble de cliques de taille **k** ou chaque clique est représentée par un sommet. Une limite de cet algorithme est qu'il nécessite un paramétrage : la valeur de **k** qui est la taille des communautés à considérer. EAGLE est un algorithme hiérarchique agglomératif basé sur des cliques maximales a été proposé par Shen et al. [156]. Il commence à identifier toutes les cliques maximales qui sont les communautés initiales. Ensuite, les communautés ayant le plus fort taux de similarité sont fusionnées, formant de nouvelles communautés, qui à leur tour, pourront être fusionnées avec des communautés semblables. La complexité estimée de cet algorithme est de $O(n^2+ (h+n) s)$**,** avec **s** est le nombre de cliques maximal, et **h** est le nombre de paires des cliques maximales en voisins.

ii. **Techniques de propagation de labels:** La propagation de label dans un réseau est la propagation d'une étiquette vers différents sommets existant dans le réseau. Chaque sommet atteint le label possédé par un nombre maximum de sommets voisins. L'algorithme de Raghavan et al. [157] est le premier qui implante cette idée. C'est un algorithme itératif où à chaque itération un sommet envoie son label à ses voisins directs, et reçoit ceux de ses voisins. Chaque sommet détermine le label majoritaire qu'il adopte pour l'itération suivante. Ce processus itératif mène à un accord sur un label précis pour chaque groupe de sommets. L'avantage de cet algorithme est qu'il est le plus performant en pratique, avec une complexité de O(n), mais il peut y avoir un problème de convergence lié à un échange infini de label entre deux sommets. L'algorithme COPRA de Gregory [158] propose une adaptation de la méthode propagation de labels aux cas avec recouvrement. Il propose, de ne plus choisir seulement le label le plus courant chez ses voisins, mais de maintenir une liste des labels les plus courants dans son entourage. Un paramètre de l'algorithme fixe le nombre maximum de labels qu'un sommet peut retenir (sans quoi chaque label s'étendrait à l'infini). La limite de cet algorithme est que le choix de sommet à traiter est aléatoire et avec une condition d'arrêt (non pas une mesure), plusieurs résultats finaux peuvent être obtenus.

4.1.2 Méthodes dynamiques

Les réseaux du monde réel évoluent continuellement et afin d'analyser ces réseaux dynamiques, il devient important d'étudier comment ces communautés se forment, évoluent et meurent avec le temps. Dans le cas de réseaux sociaux, où l'évolution est un phénomène courant, la dynamique du réseau peut donner lieu à une transformation significative de la structure de la communauté du réseau. Les techniques de détection de communautés dans les réseaux dynamiques, tels que Twitter, Facebook, LinkedIn, etc. révisent l'attribution à la communauté des sommets modifiés ou nouveaux lors des mises à jour temporelles dans le réseau.

i. **Marche aléatoire :** Une marche aléatoire peut être utilisée pour détecter des clusters dans un graphe en passant au-dessus des sommets de manière aléatoire afin de fusionner différents groupes en utilisant une approche ascendante. Ces algorithmes basés sur l'idée qu'un marcheur aléatoire partant d'un sommet a plus de probabilité à rester piégé pendant un certain temps dans la communauté du sommet de départ. L'algorithme WalkTrap [159] et Infomap [160] sont des exemples des techniques les plus populaires basées sur des marches aléatoires.

ii. **Modèle de Potts** : Le modèle de Potts est l'un des modèles bien connus utilisés en physique statistique [161]. Il démontre un système de spins qui peuvent être dans **q** états différents. Les variables de spin de Potts peuvent être mappées aux sommets du graphe ayant une structure de communauté. À partir des interactions entre les spins voisins, il est plausible que la structure de la communauté puisse être identifiée à partir de clusters de spin de même valeur du système, car il y aura plus d'interactions dans la communauté et moins d'interactions en dehors de la communauté. Inspirés par cette idée, Reichardt et al. [162] ont suggéré une technique de détection communautaire basée sur le modèle q-Potts avec des interactions entre voisins les plus proches.

4.2 Méthodes basée sur la sémantique

Le contenu sémantique et les relations entre les utilisateurs dans un réseau social peuvent être utilisés pour partitionner les sommets en communautés. Le contexte ainsi que la relation des sommets sont pris en compte dans le processus de détection de communauté de sentiments basée sur la sémantique. Quelques méthodes de détection de communauté des sentiments sont discutées par la suite.

L'une des premières tentatives d'intégration des liens sociaux et d'information de contenus dans le but de découvrir la communauté qui a été développée par Pathak et al. [137] est le modèle CART. Ce modèle a été proposé pour extraire des communautés en utilisant le contenu sémantique d'un réseau social.

Xu et al. [138], proposent l'incorporation des sentiments dans la détection des communautés, pour créer des communautés basées sur les sentiments. Ils constituent la pierre angulaire de la modélisation de tels problèmes en utilisant non seulement les propriétés des utilisateurs, mais aussi les sentiments. Cependant, une fois de plus, ils se heurtent au problème de l'optimisation d'un problème NP-hard non convexe. Néanmoins, en raison d'un modèle simple, ils utilisent une approche SDP traditionnelle basée sur l'arrondi pour résoudre ce problème.

Julian et Jure [139], ont développé une approche qui ne combine pas les propriétés structurelles avec le score de sentiment. Néanmoins, ce travail met en évidence l'approche innovante qui, en s'éloignant des propriétés structurelles du réseau, tente de modéliser la communauté en utilisant des fonctionnalités supplémentaires (dans ce cas, caractéristique du sentiment textuel). Même si cette approche utilise des fonctionnalités différentes pour la détection de la communauté, elle donne une idée que des fonctionnalités supplémentaires en plus de la structure du réseau peuvent améliorer les performances de détection et l'analyse de la communauté.

Paul et al. [140] utilisent un système qui effectue la détection communautaire des entités en fonction de leurs opinions et de leurs données sociales. Ils ont utilisé l'algorithme de détection de la communauté Infomap pour identifier les structures communautaires. Sur la base des opinions, plusieurs structures de communauté peuvent être obtenues, une pour chaque cible. L'algorithme de détection de communauté sera également exécuté sur le graphe obtenu via l'intégration de réseau. Pour les données sociales, l'algorithme de détection de communauté est appliqué de la même manière que pour les opinions, sur chaque graphe obtenu à l'étape de construction du graphe. Cependant, les structures communautaires qui en résultent sont combinées à l'aide d'une méthode de regroupement par consensus pour la stratégie d'intégration de partition. Ils ont implémenté deux méthodes décrites dans [141]: l'algorithme de partitionnement de similarité basé sur les clusters (CSPA) et l'algorithme de méta-clusters (MCLA).

Baoguo et Suresh [142] proposent deux approches communautaires, l'une est une approche Sentiment-Topic basé sur l'auteur pour la découverte communautaire, appelé ASTC, l'autre est appelée ASTCx (l'extension du modèle ASTC). La principale différence entre elles est que

ASTCx représente les mots de sentiment et les mots de sujet séparément, alors que ces mots sont mélangés dans l'approche ASTC. Dans chaque modèle génératif, les liens sociaux, les sujets et les sentiments sont systématiquement combinés. Ces trois éléments sont très importants pour l'identification de structures communautaires significatives. Cependant, cela n'indique pas que plus il y a d'informations supplémentaires incorporées dans le modèle, meilleur est le résultat que nous pouvons obtenir. Lorsque les informations ne sont pas importantes, les facteurs redondants peuvent rendre le modèle plus complexe et inefficace. Notez que toutes les communautés ne disposent pas d'informations sur les sentiments. L'objectif principal est d'obtenir certaines communautés incluant des polarités de sentiments distinctes sur certains sujets.

5 Conclusion

Dans ce chapitre, nous prenons un nouveau concept appelé «communauté de sentiments», pour étudier les sentiments et les relations des utilisateurs sur les réseaux sociaux. Nous avons aussi abordé le problème lié à la communauté de sentiments. En effet, de nombreux problèmes de recherche complexes liés à ce domaine sont apparus. Plus précisément, la détection de communauté de sentiments qui revêt une grande importance et peut avoir de nombreuses applications concrètes. Une conclusion remarquable est que la plupart des méthodes existantes se concentrent sur la structure topologique des réseaux. Cependant, les travaux récents s'intéressent de plus en plus à la sémantique des réseaux sociaux. De plus, les applications sociales présentées manquent de capacités de personnalisation. Par conséquent, l'exploit des informations disponibles dans les réseaux sociaux pour la personnalisation des systèmes interactifs sociaux utilisent les communautés de sentiments.

Chapitre 5
Contributions

Sommaire

1 Introduction

Dans le chapitre précédent, nous avons discuté des différentes approches liées à la détection de communauté dans les réseaux sociaux. En règle générale, toutes les approches sont soit basées sur les liens uniquement, soit basées sur la sémantique des réseaux. Dans ce chapitre, nous avons construit un système qui peut analyser à la fois les interactions sociales et les sentiments pour trouver des personnes similaires et voir comment cette similitude est liée à leurs interactions sociales. Afin d'atteindre notre objectif, nous utiliserons une nouvelle approche de détection communautaire pour regrouper les personnes en communautés d'un point de vue social et en utilisant les sentiments qu'elles ont exprimés. Cela nous permettra d'identifier les communautés de personnes partageant des sentiments, les communautés de personnes ayant de fortes interactions sociales et comment leurs interactions sont liées à leurs sentiments.

2 Contribution : Approche de détection de communautés

Dans cette section, nous présentons notre approche pour répondre à la problématique de détection de communautés dans les réseaux sociaux. D'abord, nous commençons par décrire l'approche générale proposée. Ensuite, nous détaillons chaque étape et la manière avec laquelle elles sont intégrées dans les réseaux sociaux. Notre contribution vise à s'appuyer sur les travaux existants dans l'analyse des réseaux sociaux pour répondre à la problématique énoncée.

2.1 Principe

Notre approche proposée est applicable sur des réseaux non orientés. La détection de communautés se fait en deux phases et le principe de l'algorithme peut être résumé de la manière suivante :

- Durant la première phase, nous détectons tous les circuits afin de décomposer le réseau initial en petits groupes élémentaires.
- Dans la deuxième phase, nous proposons une procédure itérative ayant pour objectif l'identification des différentes communautés en fusionnant les différents sous graphes issus de la première phase via un principe de fusion utilisé dans les méthodes basées sur des cliques.

2.2 Procédure d'implémentation

2.2.1 Détails de la première phase:

L'objectif de la première phase est de décomposer le réseau initial en petits groupes élémentaires en détectant tous les circuits dans ce réseau. La détection des circuits se fait par l'exploration en profondeur du graphe cible en parcourant les sommets un par un et à chaque fois qu'on tombe dans un sommet déjà visité, tous les sommets situés dans ce parcours appartenant au même circuit.

2.2.2 Algorithme de la première phase:

L'algorithme de cette phase nécessite l'utilisation d'une pile qui est capable de déterminer tous les sommets formant un circuit : chaque fois que l'on parcourt un lien allant du sommet de pile ***A*** à un autre sommet ***B*** appartenant à la pile, tous les sommets situés dans la pile depuis ***A*** jusqu'à ***B*** sont sur le même circuit. L'algorithme suivant décrit les étapes de la première phase.

Algorithme 1

***Entrées* : *G* = (*V* , *E*) : le graphe initial**

Début

Empiler (racine); // **empiler le premier sommet**

Répéter

Pour chaque lien *Nij* faire

Empiler (*N)*;

Si *N Pile* alors :

Ajouter *Circuit* (*N)* dans *list_circuit* ;

Dépiler (*N)* ;

Fin si

Fin pour

Dépiler (*N)* ; // **dépiler le sommet après visité tous ces liens**

Jusqu'à *Pile* (***vide***);

Fin Début

***Sorties* : *G*1 , *G*2 , *G*3 , ...,*GN* : des sous graphes représentant les différents** circuits détectés.

Puisque chaque sommet est empilé exactement une fois. La complexité de cet algorithme est O(n) avec **n** le nombre de sommets.

2.2.3 Détails de la deuxième phase:

Dans la deuxième phase, nous proposons une procédure itérative ayant pour objectif l'identification des différentes communautés en fusionnant les différents sous graphes issus de la première phase via un principe de fusion utilisé dans les méthodes basées sur des cliques. Pour cela, chaque sous graphe ayan **n - 1** sommets en commun avec un autre, doit être fusionné, même si les tailles des sous graphes sont différentes.

2.2.4 Algorithme de la deuxième phase :

La mise en œuvre de cette phase nécessite l'utilisation d'une liste afin de mémoriser tous les circuits détectés. Une autre procédure lancée s'occupe du raffinement des circuits détectés et de la suppression de redondances. L'algorithme suivant décrit les étapes de la deuxième phase.

Algorithme 1

***Entrées* : G_1 , G_2 , G_3 , ...,G_N : des sous graphes obtenus durant la première phase**

Début :

Pour *i de 1 à n* faire // pour chaque sous graphe

Pour *j de i à n* faire

Si *Test_graphe* (G_i , G_j) alors :

C = Fusionné (G_i , G_j) ;

Fin si

Fin pour

Liste_comm (C) // ajouter communauté dans la liste

Fin pour

Fin début

***Sorties* : C_1, C_2 , C_3 , ...,C_N : des sous graphes représentant les différentes communautés détectées**

Dans ce cas, on ne peut pas faire le calcul de complexité exact puisque les sous graphes détectés dépendent du nombre de liens (densité) dans le réseau initial. Par expérimentation, on peut dire que la complexité est de O(m – n) où **m** est le nombre de liens et **n** celui des sommets.

La complexité globale estimée de notre méthode est égale à la somme de complexité des deux phases. La complexité de l'algorithme est donc : O(n) + O(m-n) = O(n+m-n) = O(m).

3 Expérimentations

Nous avons testé notre approche sur des réseaux synthétiques générés par l'ordinateur et sur certains réseaux du monde réel couramment utilisés dans la littérature dont la structure des communautés est connue d'avance. Nous avons comparé notre approche avec les algorithmes Lauvin [152], Edge Betweenness [153], Infomap [160], Label Propagation [157] et Walktrap [159].

3.1 Réseaux synthétiques générés par ordinateur

Dans cette section, nous avons généré des réseaux synthétiques aléatoires. Ensuite, nous avons comparé les structures communautaires générées par notre méthode proposée avec certaines méthodes connues. Six réseaux ont été générés avec différentes valeurs de **n** (nombre de sommets) 100, 500, 1000, 2000, 5000 et 10000. Pour chaque valeur de **n**, le paramètre de densité **d** variait entre 0,05, 0,1, 0,2 et 0,3. Le nombre de réseaux est de 24 réseaux aléatoires au total. En effet, la détection des communautés devient très facile si le paramètre de densité d est égal à 0,05 (peu de liens intercommunautaires). Lorsque **d** est augmenté, le nombre de liens intercommunautaires augmente, la détection de la communauté devient difficile à trouver. Les figures 1 montrent la structure de la communauté trouvée par notre approche.

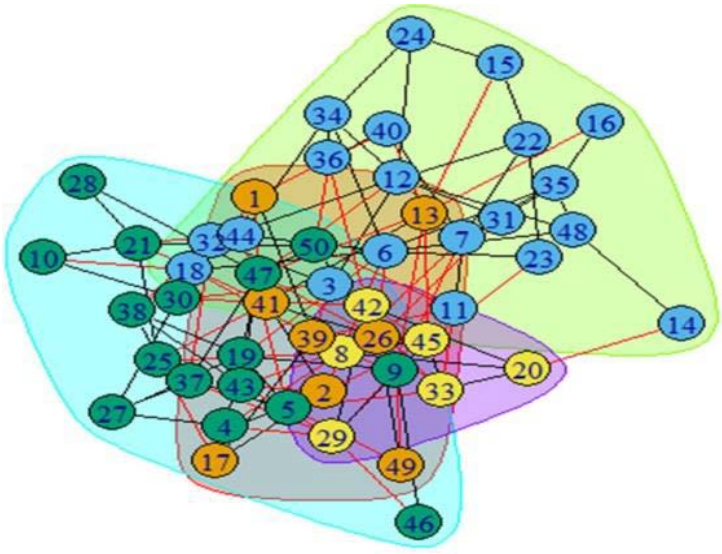


Figure 25- Structure communautaire trouvée par notre approche sur un réseau synthétique (n=50, d=0,1)

De nombreuses mesures de similarité peuvent être utilisées pour comparer la structure de la communauté qui sera trouvée par les méthodes comme Rand Index, Jaccard, Adjusted Rand Index etc. Nous avons choisi d'utiliser la métrique Rand Index pour estimer la similitude entre deux communautés de différentes structures. L'indice Rand Index donne une valeur comprise entre 0 et 1. La valeur 1 indique que les deux structures communautaires sont identiques.

Tableau 1- Rand Index de différentes méthodes sur réseau aléatoire avec d = 0.05,0.1,0.2,0.3

n=100	*d=0.05*	*d=0.1*	*d=0.2*	*d=0.3*
Infomap	0.9321212	0.8982828	0.8757576	0.8404724
EdgeBetweenness	0.9587879	0.9054545	0.8846465	0.8480808
LabelPropagation	0.2216162	0.2682828	0.2957576	0.3344464
Lauvain	0.9274747	0.8811111	0.8593939	0.8128990
Walktrap	0.9717172	0.9301010	0.9091619	0.8836364
Notre approche	0.9678435	0.9178043	0.894541	0.8702392

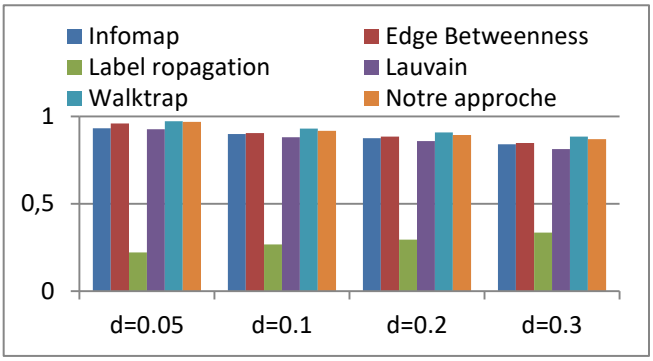


Tableau 2- Rand Index de différentes méthodes sur réseau aléatoire avec d = 0.05,0.1,0.2,0.3

n=500	*d=0.05*	*d=0.1*	*d=0.2*	*d=0.3*
Infomap	0.8677194	0.8322145	0.8104561	0.7841472
EdgeBetweenness	0.7332114	0.6995512	0.6786123	0.6496330
LabelPropagation	0.2977194	0.3265812	0.3692198	0.4047729
Lauvain	0.8777956	0.8574108	0.8287419	0.8055143
Walktrap	0.9173788	0.8845130	0.8544850	0.7963254
Notre approche	0.9052351	0.8589344	0.8301227	0.7928938

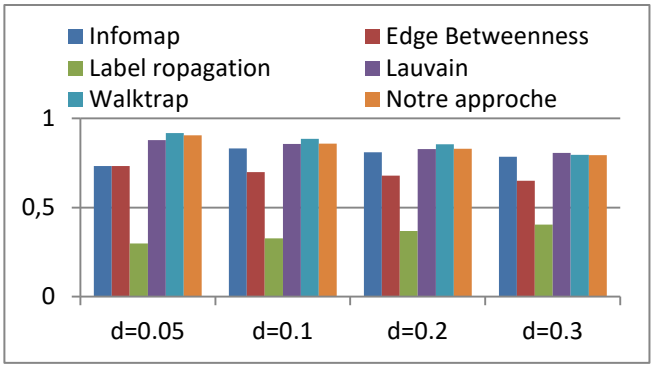


Tableau 3- Rand Index de différentes méthodes sur réseau aléatoire avec d = 0.05,0.1,0.2,0.3

n=1000	*d=0.05*	*d=0.1*	*d=0.2*	*d=0.3*
Infomap	0.8236444	0.8041263	0.7779411	0.7544728
EdgeBetweenness	-	-	-	-
LabelPropagation	0.3595315	0.3812127	0.4247817	0.4627758
Lauvain	0.8371812	0.8188746	0.7865411	0.7532392
Walktrap	0.8688649	0.8444756	0.8199874	0.7851722
Notre approche	0.8320397	0.8173082	0.7903568	0.7627616

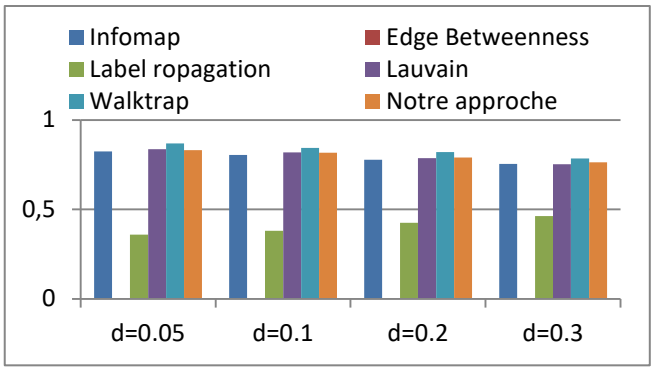


Tableau 4- Rand Index de différentes méthodes sur réseau aléatoire avec d = 0.05,0.1,0.2,0.3

n=2000	*d=0.05*	*d=0.1*	*d=0.2*	*d=0.3*
Infomap	0.7829896	0.7543321	0.7388743	0.7041432
EdgeBetweenness	-	-	-	-
LabelPropagation	0.4188954	0.4473694	0.4789114	0.5103221
Lauvain	0.8173537	0.7841322	0.7478555	0.7133577
Walktrap	0.8283807	0.8074126	0.7696336	0.7404752
Notre approche	0.8045673	0.7850932	0.7498977	0.7254221

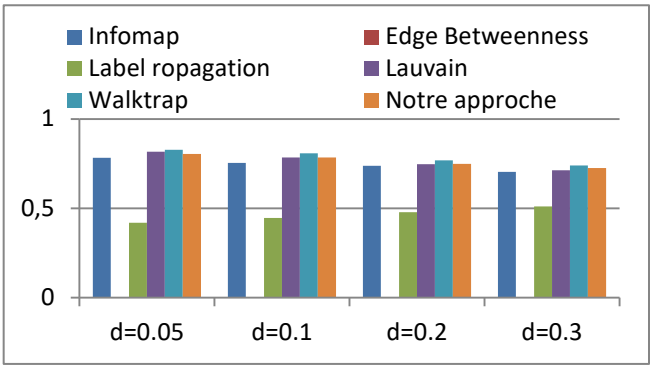


Tableau 5- Rand Index de différentes méthodes sur réseau aléatoire avec d = 0.05,0.1,0.2,0.3

n=5000	*d=0.05*	*d=0.1*	*d=0.2*	*d=0.3*
Infomap	0.7466327	0.7192734	0.6923154	0.6696259
EdgeBetweenness	-	-	-	-
LabelPropagation	0.4559153	0.4894537	0.5207328	0.5596259
Lauvain	0.7648769	0.7474123	0.7074122	0.6841246
Walktrap	0.7926148	0.7657417	0.7398913	0.7163724
Notre approche	0.7709564	0.7456933	0.7126549	0.6875432

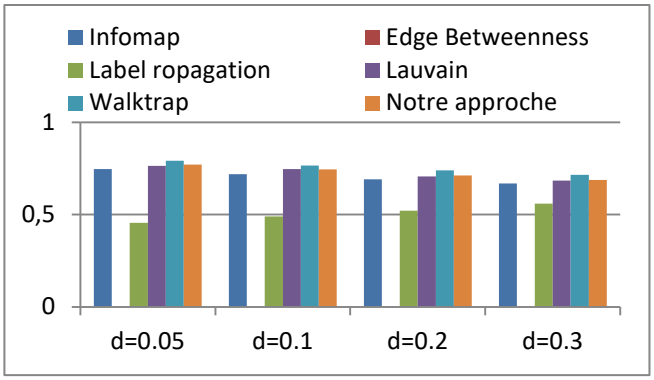
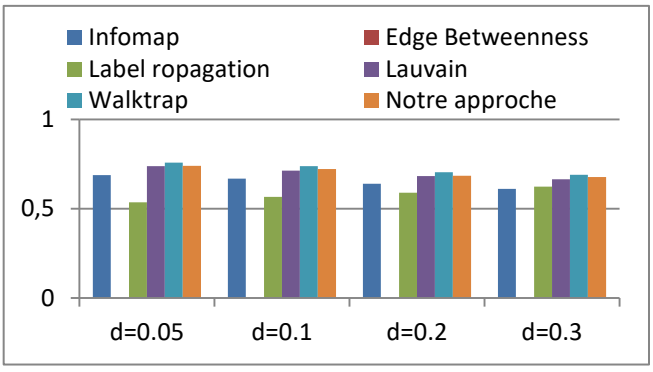


Tableau 6- Rand Index de différentes méthodes sur réseau aléatoire avec d = 0.05,0.1,0.2,0.3

n=10000	*d=0.05*	*d=0.1*	*d=0.2*	*d=0.3*
Infomap	0.6878312	0.6687565	0.6398741	0.610653
EdgeBetweenness	-	-	-	-
LabelPropagation	0.5352169	0.5668026	0.5889307	0.6229781
Lauvain	0.7390056	0.7127116	0.6828765	0.6642326
Walktrap	0.7584490	0.7380359	0.7045641	0.6905723
Notre approche	0.7404335	0.7223987	0.6849212	0.6783487



3.2 Réseaux du monde réel

Nous avons comparé notre méthode sur des réseaux réels. Le tableau 7 donne le nombre de sommets (n), de liaisons (m) et la référence pour chaque réseau.

Ces réseaux sont très populaires et largement utilisés par plusieurs algorithmes pour tester leurs performances puisque sa structure communautaire était connue auparavant. Comme le montrent les figures 3 et 4, notre méthode libère des structures communautaires qui sont très proches des structures initiales.

Tableau 7- Benchmarks des réseaux réels du monde

Réseaux	Description	n	m	Références
Karate Club	réseau des clubs de karaté de Zachary	34	78	[163]
Dolphins	réseau d'associations de dauphins	62	159	[164]
Political Books	réseau de livres politiques achetés	105	441	[165]
Football American	réseau de matchs entre équipes de football collégial	115	613	[166]

Nous avons testé le temps de calcul sur l'ensemble de données synthétiques ou réel sur un ordinateur équipé de Windows Server 2008, de deux processeurs Intel 2,67 GHz et de 24 G de RAM.

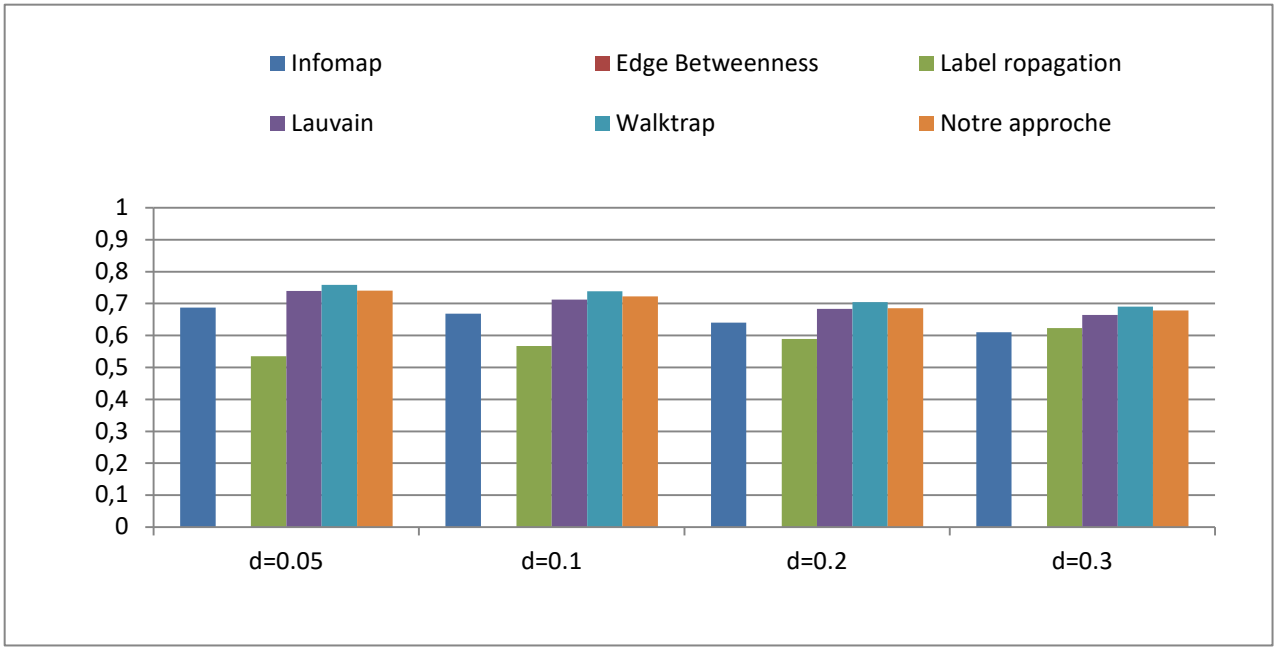
Tableau 8- Résultats de l'exécution de notre approche et les autres avec les Benchmarks choisis

	Karate Club		*Dolphins*		*Political Books*		*Football*	
	Nbr C	**Tps(ms)**	**Nbr C**	**Tps(ms)**	**Nbr C**	**Tps(ms)**	**Nbr C**	**Tps(ms)**
Infomap	3	43	6	69	6	109	11	123
EdgeBetweenness	5	49	5	97	5	621	10	1189
Label Propagation	2	27	5	31	3	39	11	47
Lauvain	4	30	5	37	4	46	10	51
Walktrap	5	40	4	49	4	54	9	62
Notre approche	3	31	4	39	4	48	9	53

Les résultats de notre approche ont démontré ses performances, et elle est toujours parmi les plus rapides par rapport aux autres algorithmes.

Tableau 9- Valeurs de Rand Index de notre approche et les autres méthodes

	Karate Club	*Dolphins*	*Political Books*	*Football*
Infomap	0.6878312	0.6687565	0.6398741	0.610653
EdgeBetweenness	-	-	-	-
LabelPropagation	0.5352169	0.5668026	0.5889307	0.6229781
Lauvain	0.7390056	0.7127116	0.6828765	0.6642326
Walktrap	0.7584490	0.7380359	0.7045641	0.6905723
Notre approche	0.7404335	0.7223987	0.6849212	0.6783487



Sur la base des résultats présentés dans le tableau 8, nous constatons que notre approche est efficace par rapport aux méthodes existantes. Il offre de très bons résultats en se rapprochant des meilleurs résultats pour tous les réseaux en termes de temps d'exécution. Les résultats du tableau 9 prouvent que la qualité des communautés obtenue dans notre méthode est plus proche de la réalité. Un exemple de structure communautaire sur un réseau réel est représenté par les figures 2 et 3.

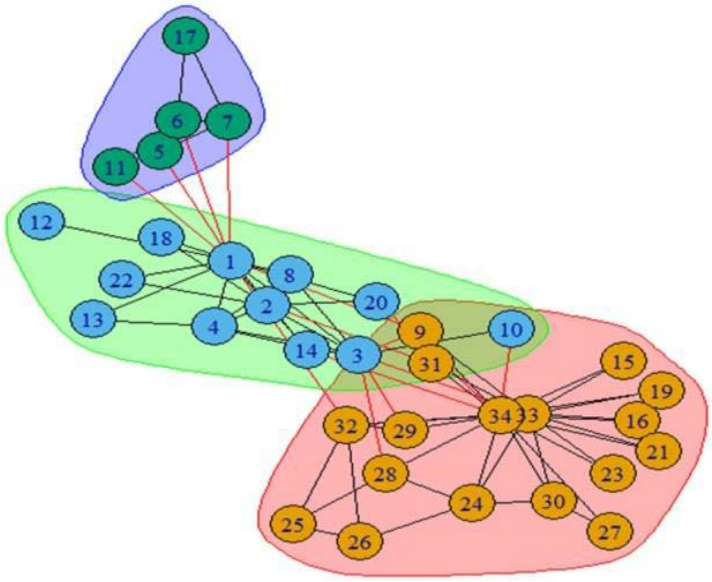


Figure 26- Structure communautaire trouvée par notre approche sur le club de Karaté Zachary

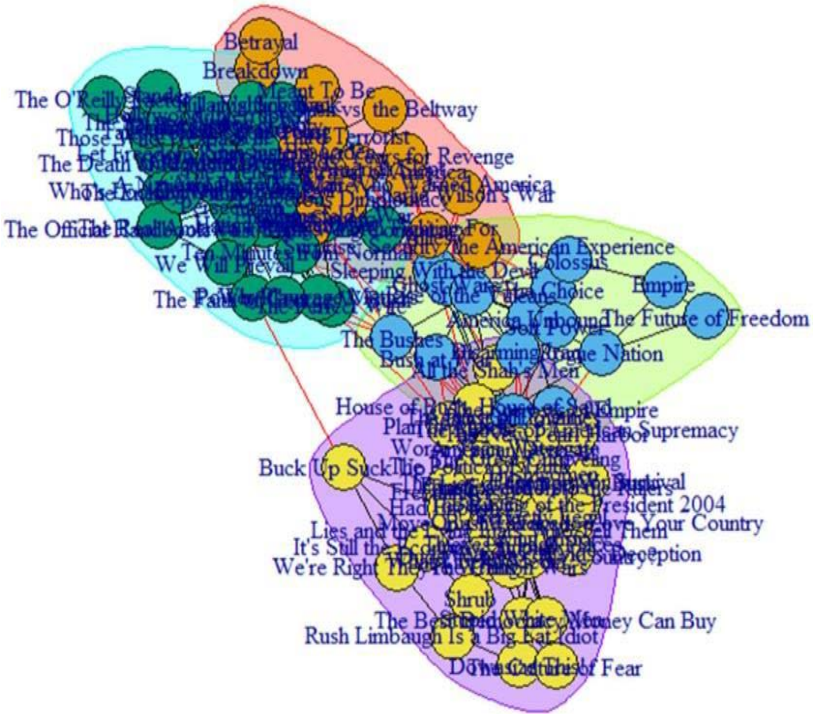


Figure 27- Structure communautaire trouvée par notre approche sur les livres politiques

3 Conclusion

Dans ce chapitre, nous avons proposé une approche de détection de communautés basées sur des cliques. De plus, nous avons intégré le concept sentimental dans la détection de la communauté afin de détecter les communautés de sentiments dans les réseaux sociaux. Cette dernière se concentre sur la découverte de communautés d'utilisateurs interconnectés qui partagent des sentiments communs sur les réseaux sociaux. Contrairement aux méthodes de détection de communauté classiques existantes qui ne prennent en compte que la connectivité dans la structure du réseau, notre approche prend en compte le nouveau concept de communauté de sentiments, qui prend en compte à la fois les relations des utilisateurs dans les réseaux sociaux et leurs sentiments.

Conclusion générale

L'analyse des sentiments est devenue un domaine de recherche très populaire. De nombreuses recherches ont été menées dans ce domaine, mais il existe encore de nombreux problèmes, car l'analyse des sentiments traite des données non structurées basées sur du texte.

L'analyse des sentiments a attiré un grand nombre de chercheurs. Par conséquent, une pléthore de modèles ont été proposés et se sont avérés donner de bons résultats sur diverses tâches de l'analyse des sentiments. Le succès des approches mentionnées est attribué à leur capacité d'apprentissage automatique des fonctionnalités et au succès des modèles d'incorporation de mots. Par conséquent, dans cette thèse, nous avons présentés d'abord le processus de l'analyse des sentiments, y compris ses applications, ses tâches et les défis d'analyse des sentiments. De même, nous avons présenté les techniques d'apprentissage automatique, y compris les techniques traditionnelles et les techniques d'apprentissage en profondeur. Nous discutons aussi des techniques basées sur le dictionnaire (lexique) et les techniques hybrides.

De plus, nous avons abordé les différents réseaux sociaux, en soulignant leur historique, leurs définitions et leurs techniques d'analyses. En outre, nous avons discuté les différentes techniques d'analyse des sentiments dans les réseaux sociaux. Enfin, des techniques de détection de communautés pour l'analyse des sentiments dans les réseaux sociaux ont été explorées.

Après avoir analysé toutes ces études, nous constatons que les approches basées sur le dictionnaire prennent moins de temps de traitement que les approches d'apprentissage automatique, mais la précision n'est pas à la hauteur. Les approches d'apprentissage automatique offre une meilleure précision. De cette enquête, on peut conclure que les approches d'apprentissage automatique offrent une meilleure précision par rapport à l'approche basée sur un dictionnaire.

Comme l'analyse des sentiments est utilisée pour prédire les points de vue des utilisateurs et que les modèles d'apprentissage en profondeur concernent tous la prédiction ou l'imitation de l'esprit humain, les modèles d'apprentissage en profondeur offrent plus de précision que les modèles peu profonds. Les techniques d'apprentissage en profondeur sont meilleures que les techniques traditionnelles telles que les SVMs et les réseaux de neurones normaux car elles ont plus de couches cachées par rapport aux réseaux de neurones normaux qui ont une ou deux couches

cachées. Les réseaux d'apprentissage en profondeur sont capables de fournir une formation de manière supervisée / non supervisée. Les réseaux d'apprentissage en profondeur effectuent une extraction automatique des fonctionnalités et n'impliquent aucune intervention humaine, ils peuvent donc gagner du temps car l'ingénierie des fonctionnalités n'est pas nécessaire.

Afin d'enrichir cette thèse, nous souhaitons quelques perspectives à notre travail et nous citons :

- Adapter notre approche de détection des communautés de sentiments dans les réseaux sociaux en tenant compte des liens logiques, pour l'identification de groupes d'utilisateurs avec des sentiments.
- Concevoir une base de données convenable pour évaluer l'approche sur différents scénarios.
- L'attribution des sentiments aux utilisateurs a été faite manuellement dans notre travail. Nous proposons dans le futur une attribution automatique des sentiments.

Bibliographies

[1] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2) (2016) 102–107.

[2] M. Tsytsarau and T. Palpanas, « Survey on mining subjective data on the web,» Data Mining and Knowledge Discovery, pp. 478-514, 2012.

[3] Andrea Esuli and Fabrizio Sebastiani (2006). « SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining ». In: Proceedings of the fifth International Conference on Language Resources and Evaluation.

[4] Bo Pang and Lillian Lee. «A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts». In Proceedings of the 42nd annual meeting on Association for Computational Linguistics, page 271. Association for Computational Linguistics, 2004.

[5] Hussein, D.M.E.-D.M. « A survey on sentiment analysis challenges». J. King Saud Univ. Eng. Sci. **2018**, 30, 330–338.

[6] B. Pang and L. Lee, « Opinion Mining and Sentiment Analysis » Vols. %1 of %22 (1-2), nº 1-135, 2008.

[7] Nhan Cach Dang , María N. Moreno-García and Fernando De la Prieta, Sentiment Analysis Based on Deep Learning: A Comparative Study, 2020

[8] P. Beineke, T. Hastie, and S. Vaithyanathan, « The sentimental factor: Improving review classification via human-provided information » in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Forum Convention Centre, Barcelona. Association for Computational Linguistics, 2004, pp. 263–270.

[9] P. D. Turney, « Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews » in Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, Pennsylvania. Association for Computational Linguistics, 2002, pp. 417–424.

[10] Ion Smeureanu, Cristian Bucur, « Applying Supervised Opinion Mining Techniques On Online User Reviews » , Informatica Economică, 2012

[11] A.Jeyapriya, C.S.Kanimozhi Selvi, « Extracting Aspects And Mining Opinions In Product Reviews Using Supervised Learning Algorithm », IEEE, 2015

[12] Cheng J, Greiner R, Kelly J, Bell D, Liu W (2002) « Learning Bayesian networks from data: an information- theory based approach ». Artif Intell 137:43–90

[13] R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett. 2003. « Bioinformatics applications of information extraction from journal articles ». Journal of Bioinformatics, 19(1):135–143.

[14] Ortigosa-Hernandez Jonathan, Rodrıguez Juan Diego, Alzate Leandro, Lucania Manuel, Inza Inaki, Lozano Jose A. « Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers ». Neurocomputing 2012;92:98–115.

[15] Kaufmann JM. JMaxAlign: « A Maximum Entropy Parallel Sentence Alignment Tool ». In: Proceedings of COLING'12: Demonstration Papers, Mumbai; 2012. p. 277–88.

[16] Jasakaran Kaur, Sheveta Vashisht, « Analysis And Indentifying Variation In Human Emotion Through Data Mining » , Int.J.Computer Technology & Applications, 2012

[17] Cortes C, Vapnik V. « Support-vector networks, presented at the Machine Learning »; 1995.

[18] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R., 2011, June. « Sentiment analysis of twitter data ». In Proceedings of the workshop on languages insocial media (pp. 30-38). Association for Computational Linguistics. http://www.cs.columbia.edu/ ~julia/ papers/ Agarwaletal11.pdf

[19] T. Mullen and N. Collier, « Sentiment analysis using support vector machines with diverse information sources ». in Proceedings of EMNLP, Barcelona, Spain, vol. 4, 2004, pp. 412–418.

[20] Wikipedia, machine à vecteurs de support ,<https://fr.wikipedia.org/wiki/ machine_%c3%a0_vecteurs_de_support >, 2017.

[21] Chin Chen Chien, Tseng You-De. « Quality evaluation of product reviews using an information quality framework ». Decis Support Syst 2011;50:755–68.

[22] Li Yung-Ming, Li Tsung-Ying. « Deriving market intelligence from microblogs». Decis Support Syst 2013.

[23] Balahur, A., 2013, June. « Sentiment analysis in social media texts». In 4th work-shop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 120-128).http://www.aclweb.org/anthology/W13-1617.

[24] RAJARAJAN, Jagadeesh, 2015. What is « Multilayer perceptrons using backpropagation algorithm », in simple words? - Disponible à l'adresse : https://www.quora.com/What-is-Multilayer-perceptrons-using-backpropagation-algorithm-in-simple-words.

[25] Medhat, W., Hassan, A., & Korashy, H., « Sentiment Analysis Algorithms and Applications : A Survey ». Ain Shams Engineering Journal, pp. 1093-1113, 2014.

[26] DELAUNAY, David, 2016. Mathématiques Sup et spé. Disponible à l'adresse : http://mp.cpgedupuydelome.fr.

[27] Ng Hwee Tou, Goh Wei, Low Kok. « Feature selection, perceptron learning, and a usability case study for text categorization». In: Presented at the ACM SIGIR conference; 1997.

[28] Ruiz M, Srinivasan P. « Hierarchical neural networks for text categorization ». In: Presented at the ACM SIGIR conference; 1999.

[29] Freund Y, Schapire R (1999). « Largemargin classification using the perceptron algorithm » .Mach Learn 37:277– 296.

[30] ZHU Jian , XU Chen, WANG Han-shi, « Sentiment classification using the theory of ANNs », The Journal of China Universities of Posts and Telecommunications, July 2010, 17(Suppl.): 58–62.

[31] Duncan, B., & Zhang, Y. 2015 . « Neural networks for sentiment analysis on twitter ». In Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on (pp. 275-278). IEEE.

[32] Liu Bing, Hsu Wynne, Ma Yiming. « Integrating classification and association rule mining ». In: Presented at the ACM KDD conference; 1998.

[33] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, and W. Zuo, « Sentiment analysis of chinese documents: From sentence to document level », Journal of the American Society for Information Science and Technology,vol. 60, no. 12, pp. 2474–2487, 2009.

[34] Kai Gao, Hua Xu, Jiushuo Wanga, « A Rule-Based Approach To Emotion Cause Detection For Chinese Micro-Blogs », ELSEVIER, 2015

[35] Ahmed, Shoiab and Ajit Danti, « Effective sentimental analysis and opinion mining of web reviews using rule based classifiers ». In Computational Intelligence in Data Mining, vol. 1, pp. 171-179. Springer, New Delhi, 2016.

[36] Jasakaran Kaur, Sheveta Vashisht, « Analysis And Indentifying Variation In Human Emotion Through Data Mining », Int.J.Computer Technology & Applications, 2012

[37] Jeevanandam Jotheeswaran, Dr. Y. S. Kumaraswamy, « Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure », Journal of Theoretical and Applied Information Technology, 2013

[38] Quinlan JR. Induction of decision trees. Machine Learn 1986;1:81–106.

[39] D. E. Rumelhart, G. E. Hinton, et R. J. Williams, 1985. « Learning internal representations by error propagation ». Rapport technique.

[40] Diederik P. Kingma and Max Welling. 2013. « Auto-encoding variational bayes ». *CoRR* abs/1312.6114 (2013). Retrieved from http://arxiv.org/abs/1312.6114.

[41] H. Yanagimoto, M. Shimada, and A. Yoshimura, « Document similarity estimation for sentiment analysis using neural network» , 2013 IEEE/ACIS 12th Int. Conf. Comput. Inf. Sci., pp. 105110, 2013.

[42] C. Baecchi, T. Uricchio, M. Bertini, and A. Del Bimbo, « A multimodal feature learning approach for sentiment analysis of social network multimedia, Multimed ». Tools Appl., vol. 75, no. 5, pp. 25072525, 2016.

[43] Abburi Harika, Rajendra Prasath, Manish Shrivastava and Suryakanth V. Gangashetty, « Multimodal sentiment analysis using deep neural networks », In International Conference on Mining Intel-ligence and Knowledge Exploration, pp. 58-65. Springer, Cham, 2016.

[44] Sun Xiao, Chengcheng Li, and Fuji Ren, « Sentiment analysis for Chinese microblog based on deep neural networks with convolution-al extension features », Neurocomputing, vol. 210, pp. 227-236, 2016.

[45] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, « Sentiment Analysis Using Convolutional Neural Network », Comput. Inf. Technol. Ubiquitous Comput. Commun. Dependable, Auton. Secur. Comput. Pervasive Intell. Comput. (CIT/IUCC/DASC/PICOM), 2015 IEEE Int. Conf., pp. 23592364,2015.

[46] A. Severyn and A. Moschitti, « Twitter Sentiment Analysis with Deep Convolutional Neural Networks », Proc. 38th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR 15, pp. 959962, 2015.

[47] L. Yanmei and C. Yuda, « Research on Chinese Micro-Blog Sentiment Analysis Based on Deep Learning », 2015 8th Int. Symp. Comput. Intell. Des., pp. 358361, 2015.

[48] J. Islam and Y. Zhang, « Visual Sentiment Analysis for Social Images Using Transfer Learning Approach », 2016 IEEE Int. Conf. Big Data Cloud Comput. (BDCloud), Soc. Comput. Netw. (SocialCom), Sustain. Comput. Commun., pp. 124130, 2016.

[49] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang, « Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN » , Expert Systems with Applications, 2016.

[50] Lee Gichang, Jaeyun Jeong, Seungwan Seo, CzangYeob Kim, and Pilsung Kang, « Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network », Knowledge-Based Systems, vol. 152, pp. 70-82, 2018.

[51] Hochreiter S, Schmidhuber J. « Long short-term memory. Neural Comput », 1997, 9: 1735–1780

[52] Cho K, Merrienboer B V, Gulcehre C, et al. « Learning phrase representations using RNN encoder-decoder for statistical machine translation ». In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, 2014. 1724–1734.

[53] Rong Wenge, Baolin Peng, Yuanxin Ouyang, Chao Li, and Zhang Xiong, « Structural information aware deep semi-supervised recurrent neural network for sentiment analysis », Frontiers of Computer Science, vol. 9, no. 2, pp. 171-184, 2015.

[54] T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang, « Using a Sequence Model for Sentiment Analysis », no. August, pp. 3444, 2016.

[55] Hu Fei, Li Li, Zi-Li Zhang, Jing-Yuan Wang, and Xiao-Fei Xu, « Emphasizing essential words for sentiment classification based on recurrent neural networks », Journal of Computer Science and Technology, vol. 32, no. 4, pp. 785-795, 2017.

[56] Goller C, K¨uchler A. « Learning task dependent distributed representations by back propagation through structure ». IEEE Trans Neur Netw, 1996, 1: 347–352.

[57] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, AndrewY. Ng, and Christopher Potts, « Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank », EMNLP (2013)

[58] Richard Socher, Cliff C. Lin, Chris Manning, and Andrew Y. Ng. 2011. « Parsing natural scenes and natural language with recursive neural networks ». In International Conference on Machine Learning. Omnipress, 129–136.

[59] C. Li, B. Xu, G. Wu, S. He, G. Tian, and H. Hao, « Recursive deep learning for sentiment analysis over social data, Proc. - 2014 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. - Work. WI IAT 2014,vol. 2, pp. 13881429, 2014.

[60] W. Li and H. Chen, « Identifying top sellers in underground economy using deep learning-based sentiment analysis », Proc. - 2014 IEEE Jt. Intell. Secur. Informatics Conf. JISIC 2014, pp. 6467, 2014.

[61] T. Mikolov, K. Chen, G. Corrado, and J. Dean, « Efficient Estimation of Word Representations in Vector Space », Arxiv, no. 9, pp. 112, 2013.

[62] Zhou Shusen, Qingcai Chen, and Xiaolong Wang, « Fuzzy deep be-lief networks for semi-supervised sentiment classification », Neuro-computing, vol. 131, pp. 312-322, 2014.

[63] Patrawu Ruangkanokmas, T. Achalakul, and K. Akkarajitsakul, Deep Belief Networks with Feature Selection for Sentiment Classification, Uksim.Info, pp. 16, 2016.

[64] Jin, Yong, Harry Zhang, and Donglei Du, « Incorporating positional information into deep belief networks for sentiment classification », In Industrial Conference on Data Mining, pp. 1-15. Springer, Cham, 2017.

[65] R. Ghosh, K. Ravi, and V. Ravi, « A novel deep learning architecture for sentiment classification », 3rd IEEE Int. Conf. Recent Adv. Inf. Technol., pp. 511516, 2016.

[66] L. Arnold, S. Rebecchi, S. Chevallier, and H. Paugam-Moisy, « An Introduction to Deep Learning », Esann, no. April, p. 12, 2011.

[67] Zhou Shusen, Qingcai Chen, and Xiao-long Wang, « Active deep learning method for semi-supervised sentiment classification », Neurocomputing, vol. 120, pp. 536-546, 2013.

[68] P. Vateekul and T. Koomsubha, « A Study of Sentiment Analysis Using Deep Learning Techniques on Thai Twitter Data », 2016.

[69] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. WordNet: An on-line lexical database1990: Oxford Univ. Press.

[70] Turney, Peter D. and Micharel L. Littman. « Measuring praise and criticism : Inference of semantic orientation from association ». ACM Transactions on Information Systems, 2003.

[71] Kamps, Jaap, Maarten Marx, Robert J. Mokken, and Maarten De Rijke. « Using WordNet to measure semantic orientation of adjectives ». in Proc. Of LREC-2004. 2004.

[72] Esuli, Andrea and Fabrizio Sebastiani. « Determining term subjectivity and term orientation for opinion mining ». in Proceedings of Conf. of the European Chapter of the Association for Computational Linguistics (EACL-2006). 2006.

[73] Esuli, Andrea and Fabrizio Sebastiani. SentiWordNet: « A publicly available lexical resource for opinion mining. in Proceedings of Language Resources and Evaluation » (LREC-2006). 2006.

[74] Andreevskaia, Alina and Sabine Bergler. Mining WordNet for fuzzy sentiment: « Sentiment tag extraction from WordNet glosses ». in Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL-06). 2006.

[75] Kaji, Nobuhiro and Masaru Kitsuregawa. « Automatic construction of polarity-tagged corpus from HTML documents ». in Proceedings of COLING/ACL 2006 Main Conference Poster Sessions (COLING-ACL-2006). 2006.

[76] Kaji, Nobuhiro and Masaru Kitsuregawa. « Building lexicon for sentiment analysis from massive collection of HTML documents ». in Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-2007). 2007.

[77] Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, and Jeff Reynar. « Building a sentiment summarizer for local service reviews ». in Proceedings of WWW-2008 workshop on NLP in the Information Explosion Era. 2008.

[78] Zhu, Xiaojin and Zoubin Ghahramani. « Learning from labeled and unlabeled data with label propagation ». School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CALD-02-107, 2002.

[79] Rao, Delip and Deepak Ravichandran. « Semi-supervised polarity lexicon induction ». in Proceedings of the 12th Conference of the European Chapter of the ACL (EACL-2009). 2009.

[80] Qiu Guang, He Xiaofei, Zhang Feng, Shi Yuan, Bu Jiajun, Chen Chun. « DASA: dissatisfaction-oriented advertising based on sentiment analysis ». Expert Syst Appl 2010;37:6182–91.

[81] Hassan, Ahmed and Dragomir Radev. « Identifying text polarity using random walks ». in Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010). 2010.

[82] Velikovich, Leonid, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. The viability of web-derived polarity lexicons. in Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (HAACL-2010). 2010.

[83] Dragut, Eduard C., Clement Yu, Prasad Sistla, and Weiyi Meng. « Construction of a sentimental word dictionary ». in Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2010). 2010.

[84] Richa Sharma,Shweta Nigam and Rekha Jain, « Mining Of Product Reviews At Aspect Level », International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.3, May 2014.

[85] Hatzivassiloglou, Vasileios and Kathleen R. McKeown. « Predicting the semantic orientation of adjectives ». in Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997). 1997.

[86] Kanayama, Hiroshi and Tetsuya Nasukawa. « Fully automatic lexicon expansion for domain-oriented sentiment analysis ». in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2006). 2006.

[87] Jiao Jian, Zhou Yanquan. « Sentiment Polarity Analysis based multi-dictionary ». In: Presented at the 2011 International Conference on Physics Science and Technology (ICPST'11); 2011.

[88] Xu Kaiquan, Liao Stephen Shaoyi, Li Jiexun, Song Yuxia. « Mining comparative opinions from customer reviews for competitive intelligence ». Decis Support Syst 2011;50:743–54.

[89] Cruz Fermı´n L, Troyano Jose A, Enrıquez Fernando, Javier Ortega F, Vallejo Carlos G. « Long autonomy or long delay?' The importance of domain in opinion mining ». Expert Syst Appl 2013;40:3174–84.

[90] Fahrni A, Klenner M. Old wine or warm beer : « target-specific sentiment analysis of adjectives ». In: Proceedings of the symposium on affective language in human and machine, AISB; 2008. p. 60–3.

[91] Cao Qing, Duan Wenjing, Gan Qiwei. « Exploring determinants of voting for the ''helpfulness'' of online user reviews: a text mining approach ». Decis Support Syst 2011;50:511-21.

[92] Kim S, Hovy E. « Determining the sentiment of opinions. In: Proceedings of interntional conference on Computational Linguistics (COLING'04) »; 2004.

[93] Maks Isa, Vossen Piek. « A lexicon model for deep sentiment analysis and opinion mining applications. Decis Support Syst 2012;53:680–8.

[94] Pai Mao-Yuan, Chu Hui-Chuan, Wang Su-Chen, Chen Yuh-Min. « Electronic word of mouth analysis for service experience ». Expert Syst Appl 2013; 40 : 1993–2006.

[95] Zhang Wenhao, Hua Xu, Wan Wei. Weakness finder: « find product weakness from Chinese reviews by using aspects based sentiment analysis ». Expert Syst Appl 2012;39:10283–91.

[96] B. P. P. Filho, L. Avanço, T. A. S. Pardo, and M. d. G. V. Nunes, Nilc_usp: « an improved hybrid system for sentiment analysis in twitter messages." in in Proceeding of 8th International Workshop on Semantic Evaluation, Dublin, Ireland. Association of Computational Linguistics Special Interest Group on the Lexicon-SIGLEX, 2014.

[97] K. Zhao and Y. Jin, « A hybrid method for sentiment classification in chinese movie reviews based on sentiment labels, » in 2015 International Conference on Asian Language Processing (IALP),Suzhou,China. IEEE, 2015, pp. 86–89.

[98] V. Nandi and S. Agrawal, « Political sentiment analysis using hybrid approach, » International Research Journal of Engineering and Technology (IRJET), vol. 03, 2016.

[99] M. Desai and M. A. Mehta, « A hybrid classification algorithm to classify engineering students'problems and perks, » Computers and Society, pp. 21–35, 2016.

[100] Liu, Bing. 2010. « Sentiment analysis and subjectivity ». In Handbook of natural language processing, vol. 2, 627–666.

[101] Stone, Philip J., Dexter C. Dunphy, andMarshall S. Smith. 1966. The general inquirer : « A computer approach to content analysis ». Cambridge, MA: MIT Press.

[102] Finn, Arup. 2011. *AFINN*. Informatics and Mathematical Modelling, Technical University of Denmark.

[103] Miller, George A. 1995.WordNet: « A lexical database for English. Communications of the ACM 38» (11): 39–41.

[104] Socher et Richard . 2013. « Recursive deep models for semantic compositionality over a sentiment Treebank ». In Proceedings of the conference on empirical methods in natural language processing (EMNL*P),* vol. 1631.

[105] Brooke, Julian, Milan Tofiloski, and Maite Taboada. 2009. « Cross-linguistic sentiment analysis : From English to Spanish ». In: *RANLP*.

[106] T. Günther, L. Furrer, GU-MLT-LT: sentiment analysis of short messages using linguistic features and stochastic gradient descent, in: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh InternationalWorkshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, GA, 2013, pp. 328–332.

[107] Y. Miura, S. Sakaki, K. Hattori, T. Ohkuma, Teamx: a sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data, in: Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 628–632.

[108] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) 267–307.

[109] R. Saurí, A Factuality Profiler for Eventualities in Text, ProQuest, Ann Arbor, MI, 2008.

[110] M. Sahlgren, The distributional hypothesis, Ital. J. Linguist. 20 (1) (2008) 33–54.

[111] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems (NIPS), 2013, pp. 3111–3119.

[112] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1, HLT '11, Association for Computational Linguistics, Stroudsburg, PA, 2011, pp. 142–150.

[113] D. Tang, F.Wei, N. Yang, M. Zhou, T. Liu, B. Qin, Learning sentiment-specific word embedding for Twitter sentiment classification, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1, 2014, pp. 1555–1565.

[114] L. Heck, H. Huang, Deep learning of knowledge graph embeddings for semantic parsing of Twitter dialogs, in: Proceedings of the Second IEEE Global Conference on Signal and Information Processing (DRAFT), IEEE Institute of Electrical and Electronics Engineers, 2014, pp. 597–601.

[115] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. J Mach Learn Res, 2003, 3: 1137–1155

[116] Collobert R, Weston J. A unified architecture for natural language processing. In: Proceedings of International Conference on Machine Learning, Helsinki, 2008. 160–167.

[117] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. In: Proceedings of International Conference on Learning Representations (ICLR), Scottsdale, 2013.

[118] Pennington J, Socher R, Manning D C. Glove: global vectors for word representation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, 2014. 1532–1543

[119] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017. 427–431.

[120] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. Journal of Computer-Mediated Communication, 3(1) :0–0, 1997. ISSN 1083-6101. doi : 10.1111/j.1083-6101.1997.tb00062.x.

[121] Danah Boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 2007.

[122] B. Wellman, Structural analysis: from method and metaphor to theory and substance, Contemp. Stud. Social. 15 (1997) 19–61.

[123] R. West, H.S. Paskov, J. Leskovec, C. Potts, Exploiting social network structure for person-to-person sentiment analysis, Trans. Assoc. Comput. Linguist. 2 (2014) 297–310.

[124] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool, San Rafael, CA, 2012.

[125] R.B. Money, M.C. Gilly, J.L. Graham, Explorations of national culture and word-of-mouth referral behavior in the purchase of industrial services in the United States and Japan, J. Market. 62 (4) (1998) 76–87.

[126] M. Thomas, B. Pang, L. Lee, Get out the vote: determining support or opposition from congressional floor-debate transcripts, in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2006, pp. 327–335.

[127] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, A.F. Smeaton, Combining social network analysis and sentiment analysis to explore the potential for online radicalisation, in: Proceedings of the International Conference on Advances in Social Network Analysis and Mining, 2009, ASONAM'09, IEEE, 2009, pp. 231–236.

[128] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, P. Li, User-level sentiment analysis incorporating social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2011, pp. 1397–1405.

[129] H. Ma, D. Zhou, C. Liu, M.R. Lyu, I. King, Recommender systems with social regularization, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 287–296.

[130] M. Sperious, N. Sudan, S. Upadhyay, J. Baldridge, Twitter polarity classification with label propagation over lexical links and the follower graph, in: Proceedings of the First Workshop on Unsupervised Learning in NLP, Association for Computational Linguistics, 2011, pp. 53–63.

[131] X. Hu, L. Tang, J. Tang, H. Liu, Exploiting social relations for sentiment analysis in microblogging, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013,pp. 537–546.

[132] F.A. Pozzi, D. Maccagnola, E. Fersini, E. Messina, Enhance user-level sentiment analysis on microblogs with approval relations, in: AI*IA 2013: Advances in Artificial Intelligence, Springer, New York, NY, 2013,pp. 133–144.

[133] Baoguo Y. and Suresh M., Community Discovery Using Social Links and Author-Based Sentiment Topics in 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014).

[134] Rushed Kanawati . "Détection de communautés dans les réseaux sociaux". A3 - LIPN UMR CNRS,2010

[135] Dong Wang , Jiexun Li, Kaiquan Xu, and Yizhen Wu. Sentiment community detection: exploring sentiments and relationships in social networks. In Electron Commer Res DOI 10.1007/s10660-016-9233-8, 2016

[136] Parau, P., Stef, A., Lemnaru, C., Dinsoreanu, M., & Potolea, R. (2013). Using community detection for sentiment analysis. 2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP).doi:10.1109/iccp.2013.6646080

[137] N. Pathak, C. Delong, A. Banerjee, and K. Erickson. " Social topic models for community extraction". 2008.

[138] Kaiquan Xu, Jiexun Li, and Stephen Shaoyi Liao. "Sentiment community detection in social networks". In: iConference. 2011.

[139] Julian J. McAuley and Jure Leskovec. "Learning to Discover Social Circles in Ego Networks". In: NIPS. 2012.

[140] URL http ://www.asimovinstitute.org/neural-network-zoo/

[141] A. Strehl and J. Ghosh, Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions, The Journal of Machine Learning Research, Volume 3, pp. 584-617 (2003)

[142] Baoguo Y. and Suresh Manandhar. "Community Discovery Using Social Links and Author-Based Sentiment Topics". 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014).

[143] Le site Web https://www.merriam-webster.com/dictionary/college

[144] Dave K., Lawrence S. et Pennock D. M. : Mining the peanut gallery: opinion extraction and semantic classification of product reviews. p. 519–528, 2003.

[150] Nedioui Med Abdelhamid. " Fouille et apprentissage automatique dans les réseaux sociaux dynamique" . 2014 Mémoire magistère

[151] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69 :066133, 2004.

[152] V.D. Blondel, J.L. Guillaume, R. Lambiotte et E. Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics : Theory and Experiment, vol. 2008, page P10008, 2008.

[153] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12) :7821{7826, 2002.

[154] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. PNAS, 101(9):2658–2663, 2004.

[155] G. Palla, A.L. Barabasi, and T. Vicsek. Quantifying social group evolution. Nature, 446(7136) :664{667, 2007.

[156] H. Shen, X. Cheng, K. Cai, and M.B. Hu. Overlapping and hierarchical community structure in networks. Physica A : Statistical Mechanics and its Applications, 388(8) :1706{1712, 2009.

[157] U.N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 76(3) :036106, 2007.

[158] S. Gregory. Finding overlapping communities in networks by label propagation. New Journal of Physics, 12(10) :103018, 2010.

[159] P. Pons and M. Latapy. Computing communities in large networks using random walks. Computer and Information Sciences-ISCIS 2005, pages 284{293, 2005.

[160] M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences,105(4) :1118{1123, 2008.

[161] Wu, F.-Y. (1982). The potts model. *Reviews of modern physics*, *54*(1), 235.

[162] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. Physical Review Letters, 93(21) :218701, 2004.

[163] W.W. Zachary. An information flow model for conflict and fission in small groups. Journal of anthropological research, 1977.

[164] David Lusseau, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten, and Steve M Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4) :396–405, 2003.

[165] KREBS V., A network of books about recent US politics sold by the online bookseller amazon.com, http ://www.orgnet.com, 2008.

[166] J. Park and M E J Newman. A network-based ranking system for us college football. *Journal of Statistical Mechanics : Theory and Experiment*, 2005(10) :P10014, 2005.

[167] Yip, Y., Cheung, W., & Ng, K. (2004). HARP: A practical projected clustering algorithm. IEEE Transactions on Knowledge and Data Engineering, 16(11), 1387–1397.