République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université Larbi Tébessi - Tébessa
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie
Département : Mathématiques et Informatique

Mémoire de fin d'étude
Pour l'obtention du diplôme de MASTER
Domaine : Mathématiques et Informatique
Filière : Informatique
Option : Systèmes d'information

Thème

Une méthode intelligente pour la détection et la classification des opinions

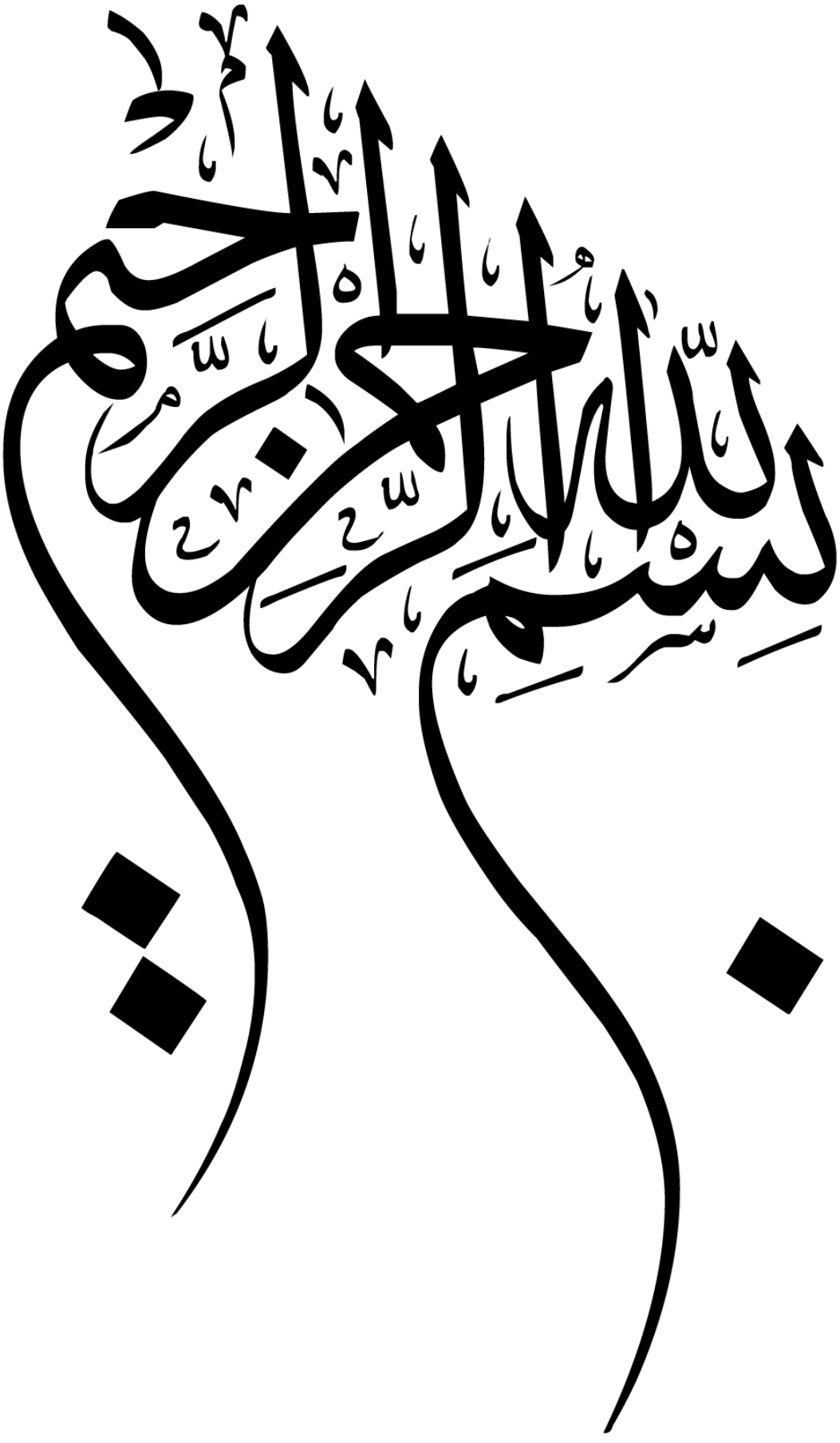
Présenté Par

Boualleg Leyla

Devant le jury

Dr. Bendib Issam	*MCB*	*Univérsité Larbi Tébessi*	*Président*
Dr. Ali Widad	*MCB*	*Univérsité Larbi Tébessi*	*Examinateur*
Dr. Amroune Mohamed	*MCA*	*Univérsité Larbi Tébessi*	*Encadreur*

Date de soutenance : 11/ 07 / 2021


بسم الله الرحمن الرحيم

REMERCIEMENTS

Tout d'abord, je remercier Dieu

Je voudrais remercier dans un premier temps, mon encadreur

***Dr. Amroune Mohamed**, pour sa patience, sa disponibilité et ses bons conseils qui ont alimenté ma réflexion.*

*Je voudrais remercier les membres de jury qui ont accepté de juger ce modeste travail **Dr.Widad Ali** .et **Dr .Issam Bendib** et tous les enseignants de département mathématique & informatique de l'université de Tébessa.*

Enfin Je tiens à remercier à nos familles et amis respectifs, pour leur soutien de tous les instants, et pour la patience dont ils ont fait preuve tout au long de cette période.

DÉDICACE

Je dédie ce modeste travail

A ma chère mère, *Qui m'a soutenu et encouragé durant ces années d'études.*

Sa prière et sa bénédiction m'ont été d'un grand secours pour mener à bien mes études.

A mon cher père, *Qui a été à mes côtés pour me soutenir et m'encourager.*

Il a su m'inculquer le sens de la responsabilité, de l'optimisme et de la confiance en soi face aux difficultés de la vie. Ses conseils ont toujours guidé mes pas vers la réussite.

A mes chers frères *« Ibrahim» « Farhate » « Said » « Wahid»*

A mes cher sœur *« Nasira » et A ma grand-mère*

Aux femmes de mes frères *«Sana» et «Djamila »*

A mes oncles*« Likhmissi » et «Elhedj »*

Aux enfants de mon frère*« Imane» et «Mohamed elamine»*

A ma chère amie*« Souhayla»*

A mes amis*« Hadjer» «Salwa» «Khouloud» «Marwa» «Siradj»*

« Marwa» «Khaira» «Imane» «Sabrina » «Fairouz».

Résumé

La fouille d'opinion ou l'analyse des sentiments est un domaine qui étudie la polarité des opinions à partir de textes. C'est l'un des domaines de recherche actif à se focalise sur l'heure actuelle, et ce en raison de l'essor des médias sociaux, où il est devenu facile d'exprimer et d'échanger des opinions. A travers se support cependant, le problème demeure de savoir comment analyser et évaluer ces opinions. Vu que les analyses manuellement prennent du temps, voire est impossible, les chercheurs ont essayé le trouver des solutions qui aident à analyser ces opinions automatiquement et parmi ces solutions L'apprentissage profond. L'apprentissage profond joue un rôle important dans l'avancement de ces recherches. Dans ce travail, nous proposons un modèle d'apprentissage en profondeur basé sur des réseaux de neurones convolutifs pour analyser les opinions arabes à l'aide d'un ensemble de données d'opinion d'Hôtel, où Les résultats ont atteint une précision de 66%.

Les Mots clés : La fouille d'opinion**,** Analyse des sentiments, Traitement du langage naturel, Apprentissage profond, Réseaux de neurones convolutifs, Réseaux de neurones récurrents

Abstract

Opinion mining or Sentiment analysis is a field that studies the polarity of opinions from texts. It is one of the active research's fields at present time, and this due to the rise of social media, where it became easy to express and exchange opinions. However the problem remains in how to analyze and evaluate these opinions. Since analyzing them manually is time consuming or even impossible, researchers have tried to find solutions that help in analyzing those opinions automatically and among those solutions Deep learning. The deep learning

plays an important role in the progress of these researches. In this work, we propose a deep learning model based on convolutional neural networks to analyze Arabic opinions using Hotel reviews dataset, where the results achieved an accuracy of 66%.

Keywords: Opinion mining, Sentiment analysis, Natural Language processing, Deep learning, Convolutional neural networks, Recurrent neural networks.

الملخص

البحث عن الرأي أو تحليل المشاعر هو مجال يدرس قطبية الآراء في النصوص. وهو أحد مجالات البحث النشط في الوقت الحاضر، ويرجع ذلك إلى ظهور وسائل التواصل الاجتماعي، حيث أصبح من السهل التعبير عن الآراء وتبادلها. لكن المشكلة تبقى في كيفية تحليل وتقييم هذه الآراء. نظرًا لأن تحليلها يدويًا يستغرق وقتًا طويلاً أو حتى مستحيلًا، فقد حاول الباحثون إيجاد الية تساعد في تحليل تلك الآراء تلقائيًا، و من بين تلك الحلول التعلم العميق. فهو يلعب دورًا مهمًا في تقدم هذه الأبحاث. في هذا العمل ، نقترح نموذجًا للتعلم العميق يعتمد على الشبكات العصبية التلافيفية لتحليل الآراء العربية باستخدام مجموعة بيانات للآراء حول الفنادق ، حيث حققت النتائج دقة 66٪.

الكلمات المفتاحية: البحث عن الراي، تحليل المشاعر، معالجة اللغة الطبيعية، التعلم العميق، الشبكات العصبية التلافيفية، الشبكات العصبية المتكررة

Sommaire

CHAPITRE02 : L'APPRENTISSAGE PROFOND & LE TRAITEMENT DE LANGAGE NATUREL

CHAPITRE 3 : CONCEPTION & REALISATION

Liste des tableaux

Liste des figures

Liste d'abréviations

- **NLP:** Natural Language processing
- **OM:** Opinion Mining
- **SA :** Sentiment Analysis
- **TAL :** Traitement du Langage Naturel
- **MOOC:** Massiv Open Online Courses.
- **IA :** Intelligence Artificielle.
- **SVM :** Support Victor Machine
- **ML:** Machine Learning
- **DL:** Deep Learning
- **CNN:** Convolutional Neural Networks
- **ReLu:** Rectified Linear Unit
- **LSTM:** Long Short-Term Memory
- **RNN:** Recurrent Neural Network
- **NER :** Named Entity Recognition
- **SRL :** Semantic Role Labeling
- **TF-IDF:** Term Frequency & Inverse Document Frequency
- **AOA :** Analyse d'Opinions en Arabe
- **DNN** : Deep Neural Networks

Introduction générale

1. Introduction

Avec l'essor d'Internet et la révolution des médias sociaux, un grand nombre de personnes peuvent exprimer leurs points de vue et leurs sentiments sur des entités, des produits, des personnes, etc. La croissance dégage énorme volume de données d'opinion disponibles sur le Web. En fait, 2,5 milliards d'octets de données sont créés chaque jour. Ces dernières années, 90 % des données mondiales ont été générées [1].

Dans ce contexte, l'analyse d'opinion automatisée suscite un intérêt croissant de la part des entreprises et de la communauté scientifique et les recherches sur l'analyse d'opinion se multiplient. Le problème de l'analyse des opinions est complexe. Il combine plusieurs tâches telles que : la détection de la subjectivité, la détection de la polarité et de son intensité, l'identification de l'entité sur qui est l'opinion et ses différents aspects, détermination de la polarité par aspect, identification du preneur d'opinion et de son profil, étude de l'évolution des opinions sur une entité donnée dans le temps. A la variété des tâches s'ajoute la nature du support utilisé pour exprimer les opinions. On distingue trois natures :

- ✓ Énoncé textuel,
- ✓ Oral,
- ✓ Audiovisuel.

Dans ce sujet, l'analyse d'opinion (AO) se réduit à détecter la polarité d'un énoncé textuel donné. Exprimé langue Arabe.

Les travaux réalisés dans le domaine de l'OA pour la détection de la polarité d'un énoncé textuel peuvent être classés selon trois approches. Le premier est symbolique, elle utilise des lexiques et règles linguistiques. La seconde consiste en une approche digitale basée sur des méthodes machine et/ou deep learning. La troisième consiste en une approche hybride qui est une combinaison des deux précédentes : il utilise à la fois des lexiques et des algorithmes d'apprentissage automatique. Jusqu'à récemment, les classificateurs Support Vector Machines (SVM) et Naive Bayes (NB) étaient les classificateurs les plus populaires dans ce domaine. Suivant la tendance actuelle, les études les plus récentes qui suivent la tendance actuelle utilisent l'apprentissage en profondeur et les réseaux de neurones.

[1] *https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1*

2. Objective

Dans ce travail, il s'est concentré sur la détermination de la polarité pour la langue arabe avec des méthodes basées sur les réseaux de neurones. Et il y a eu de nombreuses études sur ce type d'apprentissage, mais très peu sur la langue Arabe.

La langue arabe est l'une des langues les plus utilisées dans le monde. Trois catégories d'arabe peuvent être distinguées : l'arabe classique (AC) utilisé dans les textes religieux, l'Arabe standard moderne (ASM) comme langue officielle et l'arabe dialectal (AD) utilisé par les locuteurs arabes dans leurs communications informelles quotidiennes.

Un mot en arabe se définit, au sens graphique, comme étant une suite de caractères délimitée par deux séparateurs (vide ou autre marqueur de séparation, comme la ponctuation). Ou la langue Arabe se caractérise par son agglutination et sa richesse morphologique Ces caractéristiques apparaissent dans la structure des mots et conduisent à un vocabulaire clairsemé. Une phrase en arabe peut être composé d'un seul mot, reflétant la complexité des mots arabes.

Pour le traitement automatique du langage naturel écrit, la majorité des réseaux de neurones prennent en entrée des représentations vectorielles continues (encastrements) de mots. L'espace de projection est un espace continu censé préserver les similitudes sémantiques et syntaxiques des mots. Les L'imbrication de mots s'est avérée être un atout fondamental pour plusieurs tâches de traitement du langage naturel, y compris l'analyse d'opinion. La complexité de la langue arabe peut avoir un impact sur la qualité de l'espace d'intégration.

Actuellement, les plongements pré-entraînés existants représentent un mot arabe quelles que soient les caractéristiques d'agglutination et la richesse morphologique de l'arabe.

3. Structure de mémoire

Ce document est organisé en deux parties principales. Nous présentons dans un premier temps l'état de l'art de l'analyse d'opinion. Nous développons ensuite les travaux menés au cours de ce thème, et L'apport de l'intelligence artificielle et de ses branches à l'analyse des opinions, et enfin l'évaluation et les résultats.

Nous présentons généralement le domaine de l'analyse d'opinion dans le premier chapitre : définition, types, champs d'application de l'AO. Ainsi les approches d'AO et le processus général de classification des opinions.

Dans le deuxième chapitre, nous présentons les techniques d'apprentissage profond et apprentissage automatique et les différences entre les deux, ainsi nous présentons les réseaux de neurones, enfin nous présentons le NLP et les travaux associés dans le domaine AOA.

Le troisième chapitre traite des expérimentations menées pour mettre en œuvre le modèle d'analyse des sentiments pour la langue Arabe basé sur l'apprentissage en profondeur en plus des résultats obtenus.

Et enfin, la conclusion résume tout ce que nous avons abordé dans ce travail.

Chapitre 1

Analyse d'opinion

1. Introduction

La détection l'opinion ou l'analyse des sentiments (AS) est un domaine de recherche actif dans le traitement du langage naturel Il vise à étudier les attitudes et les opinions des gens sur des entités ou des sujets exprimés dans un texte.

La détection des opinions se concentre sur la découverte de la polarité de positif, négatif ou neutre des avis exprimés.

Les systèmes de la détection les opinions sont appliqués dans presque tous les domaines d'affaires et sociaux parce que les opinions sont au centre de presque toutes les activités humaines et sont des influenceurs clés de nos comportements. Nos croyances et perceptions de la réalité, et les choix que nous faisons, sont largement conditionnés par la façon dont les autres voient et évaluent le monde. Pour cette raison, lorsque nous devons prendre une décision, nous recherchons souvent les opinions des autres.

2. Analyse des sentiments : définition des concepts

2.1. Opinion

Il y a plusieurs définitions :

- **Selon la NLP (Natural Language processing)**

Dans la NL**P** (Natural Language processing) l'information textuelle dans le monde peut être classée en deux catégories principales, les **faits** et les **opinions**. Les faits sont un énoncé objectif sur les entités et les événements dans le monde. Les opinions sont subjectives et reflètent les sentiments des gens ou des leur perceptions au sujet des entités et des événements (positive, négative et neutre).

- **Selon le dictionnaire Larousse en ligne**

« Jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense : Exprimer son opinion au cours du débat. L'opinion des critiques. »[1]

[1] https://www.larousse.fr/dictionnaires/francais/sentiment/72138

« Ensemble des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc.

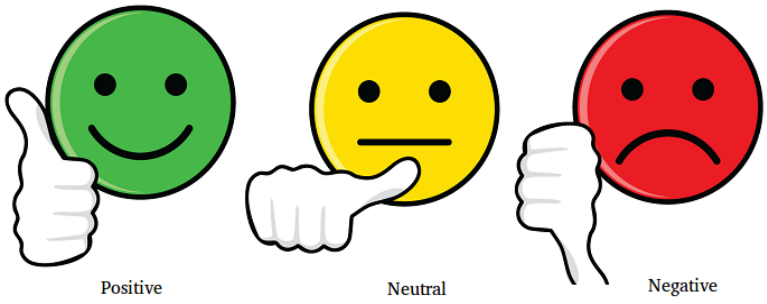


Figure 1. 1:Représentation des différentes catégories d'opinion [1]

2.2.Sentiment

- Une attitude, une pensée ou un jugement suscité par le sentiment : la prédilection
- Un point de vue ou une notion spécifique : opinion

2.3.Emotion

- Sentiment raffiné : sensibilité délicate notamment exprimée dans une œuvre d'art
- Dédalisme émotionnel
- Un sentiment romantique ou nostalgique à la limite de la sentimentalité

2.4.Information

- D'après le dictionnaire d'informatique Morvan [**2]** l'information" est un objet à la base de la communication des connaissances.
- D'après le GUINGUAY-LAURET **[3]**, l'information est la signification que l'on attribue à une expression conventionnelle ou "donnée" de telle sorte qu'elle constitue pour l'observateur un élément de connaissance **[4].**

- Le Microblogging : Le Microblogage (microblogging en anglais) permet à des internautes de publier des messages courts pour exprimer une opinion, donner un avis ou encore partager un contenu ou une information en temps réel. L'ensemble de ces messages constitue un flux**[5]** .
- **Phrase Subjective :** est un adjectif, signifiant basé sur ou influencé par des sentiments ou des émotions personnels **[6].**
- **Phrase objective :** est une phrase non basée sur ou influencé des sentiments ou des émotions personnels, une phrase objective indique des faits et des informations connues propos du monde.

3. La fouille d'opinion

L'exploration d'opinions (OM) ou l'analyse des sentiments (SA) peuvent être définies comme la tâche de détecter, d'extraire et de classer les opinions sur quelque chose. C'est un type de traitement du langage naturel (TLN) pour suivre l'humeur du public à une certaine loi, politique ou marketing, etc. Cela implique un moyen de développement pour la collecte et l'examen des commentaires et des opinions sur la législation, les lois, politiques, etc., qui sont publiées sur les médias sociaux. Le processus d'extraction d'informations est très important car c'est une technique très utile mais aussi une tâche difficile. Cela signifie que pour extraire le sentiment d'un objet sur le Web, il faut automatiser les systèmes d'exploration d'opinion pour le faire **[7].**

4. L'analyse des sentiments

L'analyse des sentiments (également appelée analyse des opinions) fait référence à l'utilisation du traitement du langage naturel (TLN) et de l'apprentissage automatique pour identifier et caractériser les états affectifs et les opinions à partir de textes ou de données vocales. L'analyse des sentiments peut être appliquée aux avis des clients, aux billets de blogue, ou micro-blogues aux réponses aux enquêtes et aux médias sociaux **[8].**

Voici une image qui illustre comment la popularité de l'analyses des sentiment s'est accrue au cours des derniéres années .Figure2

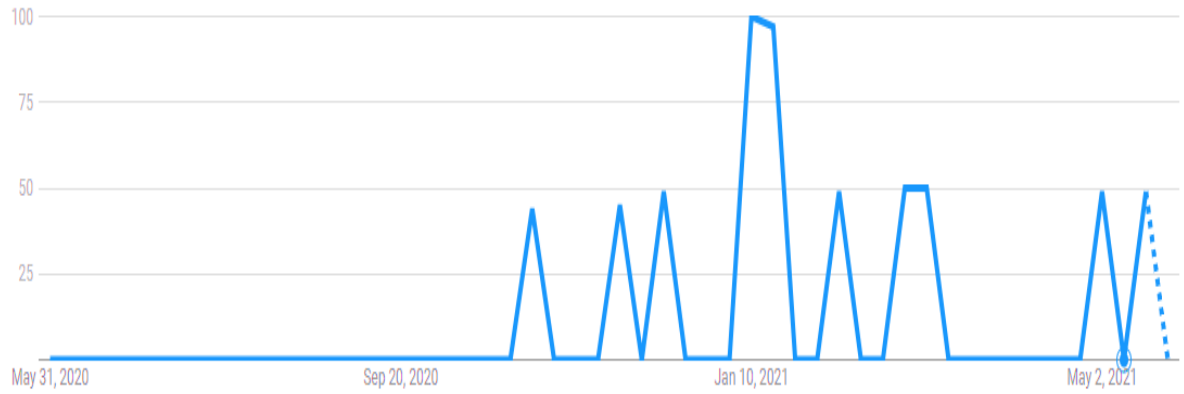


Figure 1. 2: tendance de l'analyse des sentiments au cours des dernières années « source des données : Google trends1) »

5. Différents types d'opinions :

On peut distinguer deux types d'opinions l'un s'appelle opinion régulière selon **[9]**.et l'autre dite l'opinion comparative selon [**jindal&liu**], il existe aussi des opinions fondées sur la façon dont ils sont exprimés dans le texte :

1. Opinions explicites
2. Opinions implicites

5.1. Opinion régulière versus opinion comparative :

Une opinion peut prendre différentes nuances et peut être assignée à l'un des groupes suivants :

❖ **Opinion régulière**

Une opinion régulière est souvent désignée dans la littérature comme une opinion standard et elle à deux sous-types principaux :

1. **Opinion directe**

Une opinion directe fait référence à une opinion exprimée directement sur une entité (par exemple, "La luminosité de l'écran de l'iPhone est impressionnante").

2. Opinion indirecte

Une opinion indirecte est une opinion qui est exprimée indirectement sur une entité sur la base de ses effets sur d'autres entités.

Par exemple :

La phrase : "Après être passé à l'iPhone, j'ai perdu toutes mes données»

décrit un effet indésirable du passage à l'iPhone sur "les données", ce qui donne indirectement un sentiment négatif à l'iPhone.

❖ **Opinion comparative**

Une opinion comparative exprime une relation de similitude ou de différence entre deux ou plusieurs entités et/ou une préférence du détenteur d'opinion basée sur certains aspects communs des entités **[11].** Par exemple, les phrases "iOS est plus performant qu'Android" et "iOS est le système d'exploitation le plus performant" expriment deux opinions comparatives. Une opinion comparative est habituellement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe.

5.2. Opinions explicites versus opinions implicites

Parmi les différentes nuances qu'une opinion peut prendre, nous distinguons les opinions explicites et les opinions implicites :

❖ **Opinion explicite**

Une opinion explicite est une déclaration subjective qui donne une opinion régulière ou comparative.

Par exemple

"La luminosité de l'écran de et l'iPhone est impressionnant ".

❖ **Opinion implicite**

Une opinion implicite est un énoncé objectif qui implique une opinion régulière ou comparative qui exprime habituellement un fait désirable ou indésirable.

Par exemple

- « Samedi soir, j'irai au cinéma pour regarder 'Lone Survivor'. J'ai hâte de le regarder! »
- « 'Saving Private Ryan' est plus violent que 'Lone Survivor' ».

Le premier exemple suggère qu'il y a de bonnes attentes à propos du film, bien qu'il ne soit pas expliqué en mots, alors que la compréhension de l'opinion cachée dans le second exemple est difficile même pour les humains. Pour certaines personnes, la violence dans les films de guerre pourrait être une bonne caractéristique qui rend le film plus réaliste, alors qu'elle pourrait être une caractéristique négative pour d'autres.

Il est clair que les opinions explicites sont plus faciles à détecter et à classer que les opinions implicites. Une grande partie de la recherche actuelle s'est concentrée sur des opinions explicites. Relativement moins de travail a été fait sur les opinions implicites [**10].**

6. Les niveaux d'analyse des sentiments

En général, il existe trois niveaux d'analyse : le niveau du document (Message level ou Document level), le niveau de la phrase (Sentence level) et le niveau des aspects (Entity and Aspect level). Comme le montre la figure 1.3.

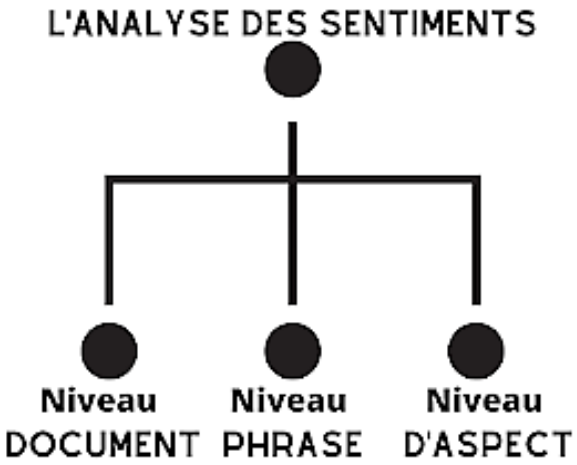


Figure 1. 3:les niveaux d'analyse du sentiment [12]

6.1. Niveau document

Dans ce niveau, la tâche exprime une polarité générale de document traité, sentiment tantôt positive tantôt négative .Par exemple dans un revu d'opinion sur un produit le système détecte l'avis globale sur le produit. Cette dernière connue sous le nom classification de sentiment au niveau du document.

L'inconvénient majeur de cette méthode, est celui des avis précis sur les différents composants des produits **[12].**

6.2. Niveau de la phrase

Le but est de déterminer la polarité de chaque phrase contenue dans un message texte. L'hypothèse est que chaque phrase, dans un message donné, dénote une seule opinion sur une seule entité **[13].**

6.3. Niveau des aspects

Les deux analyses au niveau document et phrase ne précisent pas exactement ce que les personnes aiment ou détestent. L'analyse au niveau aspect est plus fine. Cette granularité s'intéresse à un aspect donné et détermine la polarité relative à cet aspect. Il s'agit donc d'attribuer une polarité (positive, négative ou neutre) à chaque aspect évoqué dans un document d'opinion **[14].** Ceci nécessite, dans un premier temps, une extraction d'aspects et une identification de polarité pour chaque aspect dans un deuxième temps.

Par exemple, la phrase «L'iPhone est très bon, mais il faut encore travailler sur la durée de vie de la batterie et les problèmes de sécurité» évalue trois aspects : iPhone (positif), la durée de vie de la batterie (négatif) et la sécurité (négative).

7. Types d'analyse des opinions

Il existe de nombreux types d'analyses d'opinion allant des systèmes qui se concentrent sur la classification de la polarité (positif, négatif, neutre) aux systèmes qui détectent des émotions (en colère, heureux, triste, etc.) ou identifient des intentions (par exemple, intéressé, pas intéressé). Dans la section suivante, nous aborderons les types les plus importants.

7.1. Analyse fine des opinions

Au lieu de parler de phrases positives, négatives ou neutres, nous considérons les catégories suivantes :

- ✓ Très positive
- ✓ Positive
- ✓ Neutre
- ✓ Négative
- ✓ Très négative

Certains systèmes offrent également différentes classifications de polarité en identifiant si le sentiment positif ou négatif est associé à un sentiment particulier, tel que la colère, la tristesse ou des inquiétudes (sentiments négatifs) ou du bonheur, de l'amour ou de l'enthousiasme (sentiments positifs).

7.2. Les détections des opinions

L'opinion (ou le sentiment) peut être exprimée de manière très variée et subtile et donc il est difficile de la déterminer par la recherche traditionnelle thématique à l'aide des mots clés seulement. 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.

La fouille d'opinion se compose de plusieurs tâches :

- Détection de la présence ou non de l'opinion.
- Classification de l'axiologie de l'opinion (positif, négatif, neutre).
- Classification de l'intensité de l'opinion.
- Identification de l'objet de l'opinion (ce sur quoi porte l'opinion).
- Identification de la source de l'opinion (qui exprime l'opinion**) [15].**

7.3. La classification des opinions

La classification d'opinions a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime, en considérant généralement les classes : positive et négative, ou encore positive, négative et neutre. Deux grands types de méthodes sont utilisés pour cette tâche **.**Il y a tout d'abord les approches plutôt linguistiques qui consistent à répertorier le vocabulaire porteur d'opinions, puis à établir des règles de classification selon la présence ou l'absence des mots appartenant à ce vocabulaire. Il existe également les approches d'apprentissage automatique.

8. Les approches d'analyse des opinions

8.1. Approches basées sur lexique

Appelé aussi symbolique ou linguistique, jusqu'à maintenant, la plupart des études de l'analyse des sentiments se sont basées sur cette méthode. Elle permet d'identifier la polarité d'un texte à l'utilisation de deux ensembles de mots, ceux qui expriment un sentiment positif et ceux qui expriment un sentiment négatif.

Le modèle compte dans le texte le nombre de mots positifs et le nombre de mots négatifs, la somme donne une évaluation globale du sentiment de texte, si le nombre de mots positifs l'emporte sur celle de mots négatifs, le texte considéré comme positif, inversement, le texte est considéré comme négatif, éventuellement neutre si les nombres sont égaux **[16].**

8.2. Approche Basée Sur Le Corpus

Dans Bing Liu **[17]** indique que l'approche basée sur le corpus peut être appliquée dans deux cas. Le premier cas est une identification des mots d'opinion et de leurs polarités dans le corpus de domaine en utilisant un ensemble donné de mots d'opinion. Le second cas concerne la construction d'un nouveau lexique dans un domaine particulier à partir d'un autre lexique utilisant un corpus de domaine. Les résultats suggèrent que même si les mots d'opinion dépendent du domaine, il peut arriver que le même mot ait une orientation opposée selon le contexte.

8.3. Approches hybrides

Cette approche est appelée aussi classification semi-supervisée, elle combine les points forts de deux approches précédentes, il y a trois façon de faire. La première est d'exploiter les outils linguistiques pour élaborer le corpus puis classer les textes par un outil d'apprentissage supervisé. La deuxième façon est d'utiliser l'apprentissage automatique pour établir le corpus d'opinion nécessaire à l'approche basée sur lexique. La troisième façon est le conjointement des deux approches précédentes et la combinaison de leurs résultats soit par un système de vote soit par un algorithme d'apprentissage **[18].**

9. La classification des opinions

La classification d'opinions a pour but d'attribuer une étiquette au texte selon l'opinion qu'il exprime, en considérant généralement les classes : positive et négative, ou encore positive, négative et neutre.

Deux grands types de méthodes sont utilisés pour cette tâche. Il y a tout d'abord les approches plutôt linguistiques qui consistent à répertorier le vocabulaire porteur d'opinions, puis à établir des règles de classification selon la présence ou l'absence des mots appartenant à ce vocabulaire. Il existe également les approches d'apprentissage automatique.

Figure 1. 4:Le processus général de classification d'opinions [10].

9.1. Acquisition et prétraitement du corpus

Dans cette phase, les textes sont prétraités linguistiquement. Une élimination des mots vides et des mots qui n'apportent aucune information n'est faite, ainsi qu'une analyse lexicale pour enlever les mots qui ont un sens commun (redondant). et un étiquetage grammatical est fait (pour reconnaître l'adjectif, l'adverbe,
Le verbe, etc).

9.2. Extraction des caractéristiques

La conversion d'un morceau de texte en un vecteur de caractéristiques ou autre représentation qui rend ces caractéristiques les plus saillantes et les plus importantes disponibles, est une partie importante des approches d'apprentissage supervisée. Il existe plusieurs travaux qui traitent la sélection de caractéristiques pour les approches d'apprentissage supervisée en général, ainsi que pour les approches d'apprentissage adaptées aux problèmes spécifiques de la catégorisation classique de texte et de l'extraction d'information.

9.3. Classification

La classification d'opinions Plusieurs méthodes ont été utilisées pour la classification d'opinions. Leur but est de réordonner les documents pertinents selon un score d'opinion. Ainsi les documents qui contiennent le plus d'opinions sont classés parmi les premiers. Apres ces étapes, une évaluation des résultats est faite. Les résultats sont souvent confrontés à la perception humaine de l'opinion. La comparaison est faite grâce à des mesures de similarité. Plusieurs campagnes d'évaluation ont vu le jour, permettant aux chercheurs de présenter leurs travaux et de les évaluer sur des collections test élaborées par ces campagnes **[19].**

10. Domaines d'application de l'analyse d'opinions

Le domaine d'analyse d'opinions connait un intérêt croissant. Les domaines d'application d'analyse d'opinions sont nombreux. Dans cette section, nous présentons les domaines d'application de l'analyse d'opinions suivants : domaine, commercial, éducatif, politique et le domaine de la santé.

10.1.Domaine commercial

Les entreprises analysent les commentaires de clients et leurs retours afin d'améliorer la qualité des produits ou modifier la stratégie du marketing. En effet, l'analyse des commentaires dans les médias sociaux est de plus en plus considérée comme un outil d'aide à la décision pour la compréhension du marché, la catégorisation de la clientèle et la gestion de produits. Le client ou le consommateur se renseigne, avant d'acheter un produit ou un service, en analysant les retours des autres utilisateurs. Cette analyse de retours permet au client d'avoir une idée générale sur l'objet en question et le comparer éventuellement avec d'autres objets disponibles sur le marché. Ainsi, l'analyse d'opinions s'avère utile sur le plan commercial pour l'étude de marché, et donc très utilisée dans les entreprises pour adopter les bonnes stratégies dont le but ultime est d'augmenter leurs chiffres d'affaires.

10.2.Domaine éducatif

L'analyse de sentiments est aussi appliquée dans le domaine de l'éducation [**20,21 ,22]**. En effet, plusieurs sites web et plateformes MOOC proposent des cours en ligne. L'apprentissage en ligne est de plus en plus utilisé. Les MOOC sont suivies par plusieurs apprenants du monde entier avec seulement un ordinateur et une connexion internet. Ces plateformes offrent la possibilité de suivre les cours sans contrainte de temps. Chaque apprenant suit son cours à son rythme et selon ses disponibilités. Des certifications délivrées par ces plateformes sont également fournies pour justifier et prouver la validation de modules sous certaines conditions. Au cours de leur apprentissage, les apprenants interagissent via des forums et des espaces de discussion pour poser ou répondre aux questions, s'exprimer et donner des avis, etc. Et à la fin du module d'apprentissage, les apprenants remplissent des formulaires pour donner leurs points de vue sur le cours suivi. Afin d'évaluer leurs contenus, les MOOC mesurent la satisfaction des apprenants via leurs interactions dans les forums en cours d'apprentissage et leurs évaluations du module d'apprentissage par formulaire.

10.3.Domaine de la santé

L'analyse d'opinions est aussi appliquée dans le domaine de la santé **[23].** En effet, l'identification ou la compilation des commentaires relatifs à la santé (ou un de ses aspects : cigarettes électroniques par exemple comme alternative au tabac **[24]** dans des certains sites

web ou réseaux sociaux est utile à la fois pour les fournisseurs de soins de santé et les professionnels de la réglementation en santé publique. L'analyse de contenus relatifs à la santé permet aussi de mesurer l'impact d'une maladie sur la personne touchée par ce problème et son entourage. Des études dans ce sens ont été menées pour des maladies chroniques **[25]** comme le diabète **[26]**, le cancer **[27],** etc.

10.4.Domaine politique

Les politiciens ont recours aux réseaux sociaux pour s'adresser au public et rendre leurs programmes plus accessibles aux électeurs. L'analyse d'opinions s'applique dans plusieurs domaines. Elle intervient pratiquement afin de développer d'outils pour extraire, identifier, synthétiser et comparer des opinions.

10.5.Économie

Avant d'acheter un produit, la majorité des clients demandent conseil sur un produit ou un service donné et sont même disposés à payer plus pour un produit dont l'opinion est plus favorable qu'un autre, ce qui peut augmenter les ventes. Grâce à l'analyse des sentiments, les entreprises peuvent connaître l'opinion des clients sur leurs produits ou leurs services. Dans une perspective d'amélioration de leurs produits et d'augmentation de leurs ventes et revenus.

10.6.Éducation

L'analyse des sentiments peut être utilisée pour extraire des informations utiles sur la méthodologie d'enseignement d'un enseignant et également sur le programme du cours. Il identifie le degré d'apprentissage des étudiants, comprend leurs besoins, prévoit leurs performances et apporte des changements effectifs dans le style. Les résultats de l'analyse des sentiments aident les enseignants et les établissements à prendre des mesures correctives.

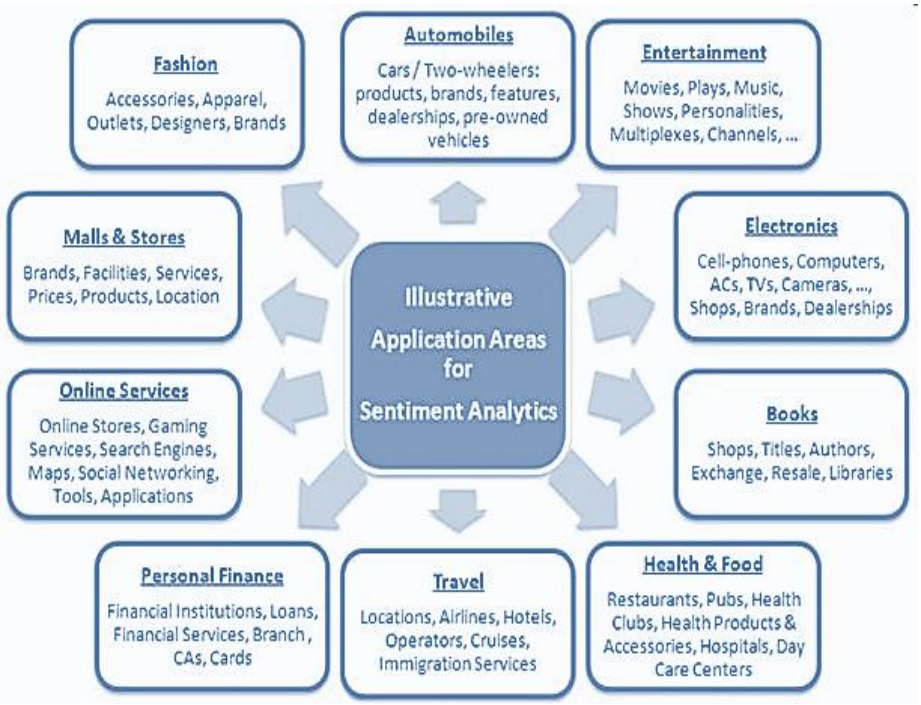


Figure 1. 5:Domaines d'application de l'analyse d'opinions [28]

11.Conclusion

Nous avons étudié dans ce chapitre, les définitions de la fouille de la fouille d'opinion et les types d'opinions, nous avons abordé la classification d'opinion, nous avons expliqué le processus de la fouille d'opinion et ces étapes, et les niveaux d'analyse d'opinion. Puis nous avons défini les deux méthodes d'apprentissage automatique, l'apprentissage supervisé et celui-ci non supervisé. Nous avons présenté les approches d'analyse d'opinion qui sont l'approche basée sur lexique, l'approche basée sur corpus et l'approche hybride. Puis nous avons cité quelques domaines d'application de la fouille d'opinion.

Chapitre 2

L'apprentissage profond & le traitement de Langage naturel

1. Introduction

L'intelligence artificielle se développe chaque jour et l'apprentissage profond est l'un des facteurs contribuant à ce développement. L'apprentissage profond est une branche de l'apprentissage automatique qui traite d'algorithmes inspirés de la structure des réseaux de neurones dans le cerveau humain. En d'autres termes, il imite la façon dont notre cerveau accomplit une tâche spécifique. Les algorithmes d'apprentissage en profondeur sont similaires à la façon dont le système nerveux est organisé, chaque neurone communiquant avec d'autres cellules et passant des informations entre elles. La méthode d'apprentissage profond représente l'essentiel des recherches menées par les professionnels, notamment dans son intervention dans plusieurs domaines comme le traitement du langage naturel.

2. Intelligence artificielle (IA)

John McCarthy a inventé le terme (IA) en 1956 lors d'un rassemblement de recherche d'été au Dartmouth College. Certains des meilleurs chercheurs américains se sont réunis dans différentes disciplines pour discuter de ce qui allait devenir le domaine de l'IA. L'intelligence artificielle. (IA, ou AI en anglais pour Artificial Intelligence) a été définie comme «un sous-domaine de l'informatique, qui implique le développement de machines intelligentes capables d'exécuter des tâches caractéristiques des humains (voir, entendre, traiter le langage)».

Les machines alimentées par l'IA peuvent être classées en deux catégories : générales et étroites. Un système d'IA générale effectuerait toutes les tâches de type humain. Par exemple, un robot qui peut traiter le langage et vous parler, tout en étant capable de voir, d'analyser l'environnement environnant.

D'un autre côté, les systèmes d'IA étroits ne peuvent effectuer qu'une seule tâche, mais ils peuvent le faire extrêmement bien, parfois mieux que les humains. Par exemple, la détection de visage et le marquage de Facebook fonctionnent extrêmement bien, étant capable de détecter des visages similaires dans différentes images (mais il ne peut pas traiter la langue en même temps, il s'agit donc d'un système d'intelligence artificielle étroite) **[29].**

L'apprentissage automatique et L'apprentissage profond sont des sous-ensembles de l'intelligence artificielle.

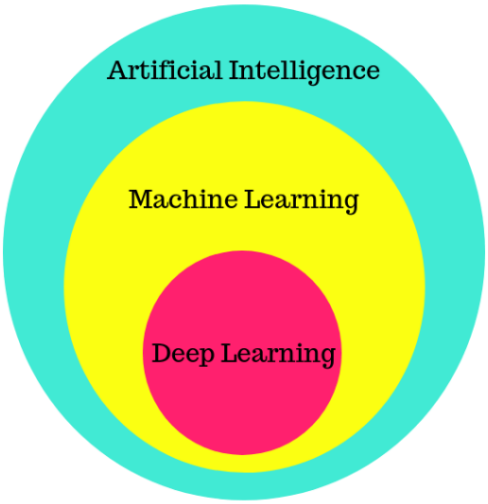


Figure 2. 1:La relation entre l'intelligence artificielle, l'apprentissage automatique et l'apprentissage en profondeur

3. L'apprentissage profond et l'apprentissage automatique

3.1. L'apprentissage profond

Le **Deep Learning**, ou apprentissage profond, est un sous-ensemble du Machine Learning, ou apprentissage automatique, basé sur des réseaux neuronaux artificiels. Le processus d'apprentissage est qualifié de profond parce que la structure des réseaux neuronaux artificiels se compose de plusieurs couches d'entrée, de sortie et masquées. Chaque couche contient des unités qui transforment les données d'entrée en informations que la couche suivante peut utiliser une tâche prédictive spécifique. Grâce à cette structure, une machine est capable d'apprendre au travers de son propre traitement de données **[30].**

3.2. Apprentissage automatique

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle (IA) qui se concentre sur la conception de systèmes qui apprennent – ou améliorent le rendement – en fonction des données qu'ils consomment. Cette technique s'appuie sur le développement de programmes informatiques capables d'acquérir de nouvelles connaissances afin de s'améliorer

et d'évoluer d'eux-mêmes des qu'ils sont à exposer de nouvelles données. Ils fonctionnent en construisant un modèle à partir d'exemples d'entrées afin de faire des prédictions ou des choix basés sur les données plutôt que de suivre des instructions de programme statiques **[31].**

L'apprentissage automatique est généralement divisé en :

- ✓ L'apprentissage automatique supervisé.
- ✓ L'apprentissage automatique non supervisé.

3.2.1. Apprentissage supervisé

La forme la plus commune d'apprentissage automatique est l'apprentissage supervisé. Il est basé sur les données libellées et donc, les étiquettes sont fournies au modèle au cours du processus d'apprentissage. Ces données libellées sont utilisées par l'algorithme d'apprentissage pour donner un modèle qui sera utilisé lors de la prise de décision. Certains modèles d'apprentissage automatique ont été formulés pour classer le texte, c'est-à-dire que l'on soumet au classificateur des exemples pour s'entrainer à classer correctement les documents futurs.

Il existe de nombreuses méthodes d'apprentissage supervisé :

- K plus proches voisins
- Arbres de décisions
- Naïve Bayes (ou encore Simple Bayes)
- Réseaux de neurones
- Machines à support de vecteurs (ou SVM)
- Programmation génétique **[32].**

3.2.2. Apprentissage non supervisé

L'apprentissage non supervisé, encore appelé apprentissage à partir d'observations, partage une propriété commune avec l'apprentissage supervisé: il transforme un jeu de données en un autre. Mais l'ensemble de données dans lequel il se transforme n'est pas connu ou compris auparavant .Contrairement à l'apprentissage supervisé sera quant à lui alimenter uniquement par des exemples, et créera lui-même les classes qui lui semblent les plus judicieuses

(clustering) ou des règles d'associations (algorithmes Apriori). L'algorithme K-moyen (K-means) permet de comprendre facilement le concept de classification non supervisée **([33],[34])**.

3.3. La différence entre l'apprentissage profond et l'apprentissage automatique

Par rapport aux autres méthodes de machine learning, le Deep Learning occupe actuellement une position avantageuse en raison d'une autre avancée technologique considérable : le Big Data.

Auparavant, la collecte de données était un problème et cela mettait le Deep Learning dans une position faible. Les algorithmes nécessitaient un volume élevé de données pour fonctionner efficacement, ce qui n'était pas facilement disponible. Cependant, avec le Big Data envahissant les vannes d'Internet et capturant tous les aspects de notre vie, l'apprentissage en profondeur représente désormais l'outil idéal pour tirer des informations et des actions significatives à partir de volumes élevés de données.

Le volume de données collectées par des systèmes disparates est immense. Il fournit suffisamment de carburant au moteur Deep Learning pour surpasser toutes les autres méthodes d'apprentissage automatique en termes de précision et de vitesse.

Un autre avantage du Deep Learning est qu'il nécessite une moindre connaissance du domaine par rapport aux autres formes de ML. D'autres méthodes d'apprentissage automatique nécessitent qu'un expert du domaine définisse les fonctionnalités à identifier afin que le programme puisse facilement passer au crible les données.

Grâce à une approche progressive de l'apprentissage, les algorithmes DL peuvent essayer d'apprendre des fonctionnalités de haut niveau à partir des données disponibles sans recourir à un expert du domaine pour définir chaque fonctionnalité.

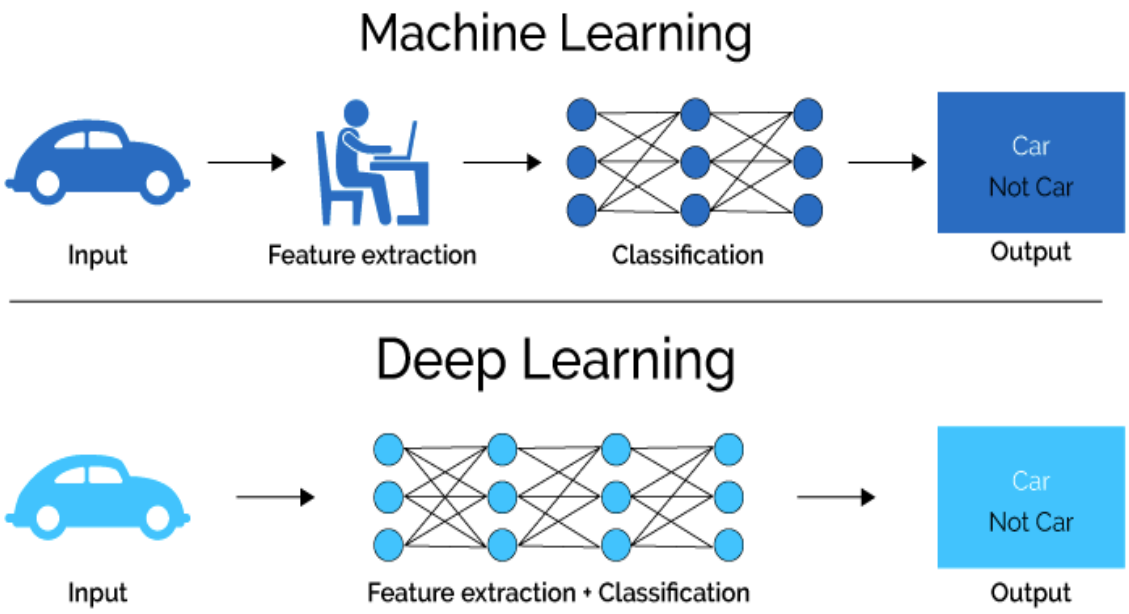


Figure 2. 2:Différence entre l'apprentissage automatique des machines et l'apprentissage profond [35].

4. L'importance de l'apprentissage profond

L'apprentissage automatique n'est pas utile lorsque vous travaillez avec des données de grandes dimensions, c'est-à-dire que nous avons un grand nombre d'entrées et de sorties. Ne peut pas résoudre des problèmes cruciaux d'intelligence artificielle comme le NLP, la reconnaissance d'image etc.

Extraction de caractéristiques est un des grands défis des modèles d'apprentissage machine traditionnels. Cette extraction automatisée des caractéristiques permet aux modèles de Deep Learning d'atteindre un taux de précision particulièrement élevé pour les tâches de vision par ordinateur.

Les modèles d'apprentissage profond sont capables de se concentrer sur les fonctionnalités appropriées par eux-mêmes, nécessitant peu de conseils de la part du programmeur.

4.1. Principe de fonctionnement

La plupart des méthodes de Deep Learning utilisent des architectures de **réseaux de neurones**, ce qui explique pourquoi il est souvent question de **réseaux de neurones profonds** pour désigner des modèles de Deep Learning.

Le terme « profond » se rapporte généralement au nombre de couches cachées du réseau de neurones. Les réseaux de neurones classiques ne comportent que 2 à 3 couches cachées, tandis que les réseaux profonds peuvent en compter jusqu'à 150 **[36]**.

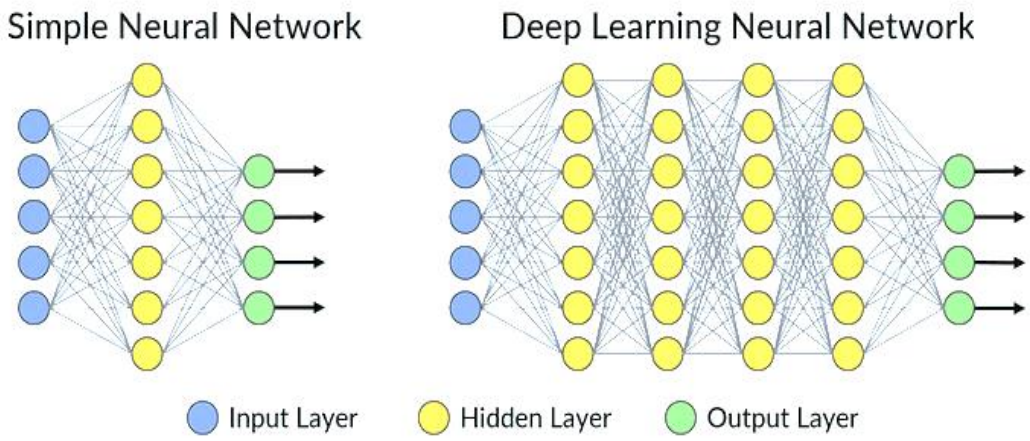


Figure 2. 3:Différence entre un réseau de neurones simple et un réseau de neurones d'apprentissage en profondeur [37].

4.2. Réseau de neurones

Un réseau de neurones est défini comme un ensemble de noeuds (appelés neurones) connectés via des liaisons dirigées (flèche), chaque flèche représente une connexion entre la sortie d'un neurone et l'entrée d'un autre(les flèches entrantes étant les entrées du neurone et les flèches sortantes étant les sorties du neurone), Chaque flèche porte un poids, reflétant son importance, chaque nœud étant une unité de traitement qui exécute une fonction de nœud statique sur son signal entrant pour générer une sortie de nœud unique**[38]**.

Les valeurs d'entrée, ou en d'autres termes, nos données sous-jacentes, sont transmises via ce «réseau» de couches masquées jusqu'à ce qu'elles convergent vers la couche de sortie.

La couche en sortie correspond à notre prédiction: il peut s'agir d'un nœud si le modèle ne génère qu'un nombre ou de quelques noeuds s'il s'agit d'un problème de classification multi-classe. La forme à l'intérieur des neurones dans les couches centrales représente une fonction d'activation (typiquement un 1 = (1 + e − x)) qui est appliquée à la valeur du neurone avant de le transmettre à la sortie **[39]**.

4.3. Fonctions d'activation : algorithmes mathématiques appliqués aux valeurs de sortie

Une fonction d'activation est une fonction mathématique utilisée sur un signal. Elle va reproduire le potentiel d'activation que l'on retrouve dans le domaine de la biologie du cerveau humain. Elle va permettre le passage d'information ou non de l'information si le seuil de stimulation est atteint. Concrètement, elle va avoir pour rôle de décider si on active ou non une réponse du neurone. Un neurone ne va faire qu'appliquer la fonction suivante :

$$\mathbf{X = \sum (entrée * poids) + biais\ [40]}.$$

Ci-dessous sont les types de fonction d'activation (la table2.1).

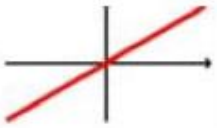
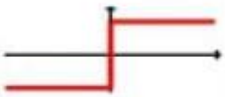
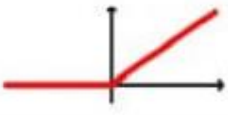
activation function	*equation*	*example*	*1D graph*
Linear	$\phi(z) = z$	Adaline, linear regression	
Unit Step (Heaviside Function)	$\phi(z) = \begin{cases} 0 & z < 0 \\ 0.5 & z = 0 \\ 1 & z > 0 \end{cases}$	Perceptron variant	
Sign (signum)	$\phi(z) = \begin{cases} -1 & z < 0 \\ 0 & z = 0 \\ 1 & z > 0 \end{cases}$	Perceptron variant	
Piece-wise Linear	$\phi(z) = \begin{cases} 0 & z \leq -½ \\ z + ½ & -½ \leq z \leq ½ \\ 1 & z \geq ½ \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multilayer NN	
Hyperbolic Tangent (tanh)	$\phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multilayer NN, RNNs	
ReLU	$\phi(z) = \begin{cases} 0 & z < 0 \\ z & z > 0 \end{cases}$	Multilayer NN, CNNs	

Tableau 2. 1: les fonctions d'activation d'un réseau de neurones [41].

5. Modèles d'apprentissage profond

Dans cette section on va présenter quelques types de modèle ou bien des algorithmes d'apprentissage profond.

Dans les années 1980, la plupart des réseaux de neurones ne formaient qu'une seule couche en raison du coût de calcul et de la disponibilité des données. De nos jours, nous pouvons nous permettre d'avoir plus de couches cachées dans nos réseaux de neurones, d'où le surnom d'apprentissage profond. Les différents types de réseaux de neurones disponibles à l'utilisation ont également proliféré, des modèles tels que les réseaux de neurones convolutionnels (CNN), les réseaux de neurones récurrents (RNN) et (LSTM).

5.1.Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (Convolutional Neural Network CNN) sont un type spécialisé de réseaux de neurones multi-couches généralement utilisés quand l'entrée est structurée selon une grille (ex : une image). Ces réseaux ont été inspirés des travaux de **[42]** sur le cortex visuel des animaux, et plus particulièrement sur ses propriétés : les champs récepteurs locaux et le partage de poids.

Les CNN sont initialement introduits par [**43]** pour une tâche de reconnaissance de formes, et ont été popularisés, dans les années 1990, avec les travaux de **[44]** sur la reconnaissance de caractères.

La figure 4 montre les différentes couches d'un réseau de neurones convolutif. Ce dernier est composé d'un ou plusieurs blocs de convolution et de pooling, une ou plusieurs couches cachées et une couche de sortie. Le CNN prend en entrée une grille multi-dimensionnelle représentant une instance d'apprentissage ou d'inférence, et fournit en sortie la classe correspondante.

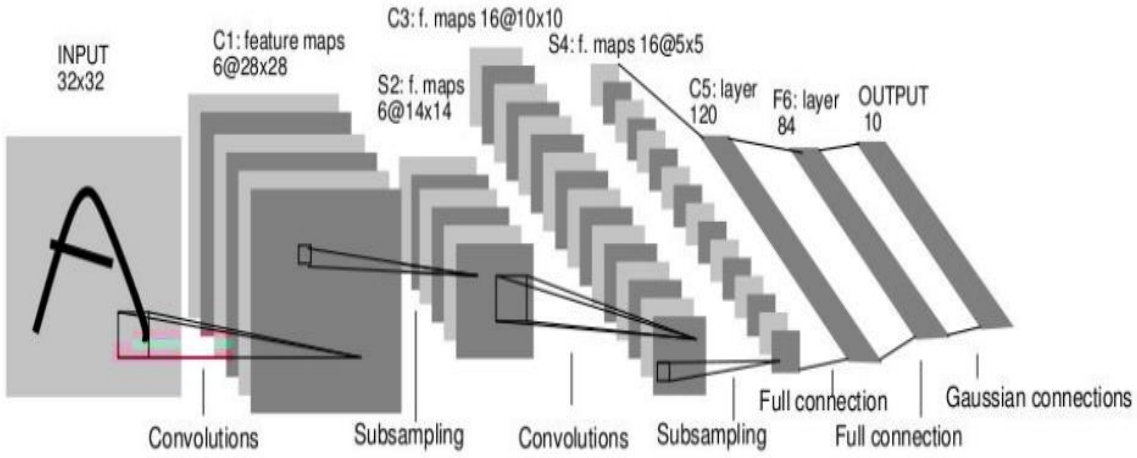


Figure 2. 4:Réseau de neurones convolutif composé de deux couches de convolution et de pooling (subsampling), suivi de deux couches cachées et d'une couche de sortie [45].

Une architecture CNN est formée par un empilement de couches de traitement indépendantes :

- La couche de convolution (CONV) qui traite les données d'un champ récepteur.
- La couche de pooling (POOL), qui permet de compresser l'information.
- La couche de correction (ReLu), souvent appelée par abus 'ReLu' en référence à la fonction d'activation (Unité de rectification linéaire).
- La couche "entièrement connectée" (FC), qui est une couche de type perceptron.
- La couche de perte (LOSS) **[45] .**

5.2.Les réseaux de neurones récurrents (RNN)

Les réseaux neuronaux récurrents (RNN) sont un type de réseau neuronal dans lequel la sortie de l'étape précédente est alimentée en entrée de l'étape en cours. Dans les réseaux de neurones traditionnels, toutes les entrées et sorties sont indépendantes les unes des autres, mais dans des cas comme lorsqu'il est nécessaire de prédire le mot suivant d'une phrase, les mots précédents sont nécessaires et il est donc nécessaire de se souvenir des mots précédents. C'est ainsi que RNN a vu le jour, qui a résolu ce problème à l'aide d'une couche cachée. La caractéristique

principale et la plus importante de RNN est l'état caché, qui se souvient de certaines informations sur une séquence.

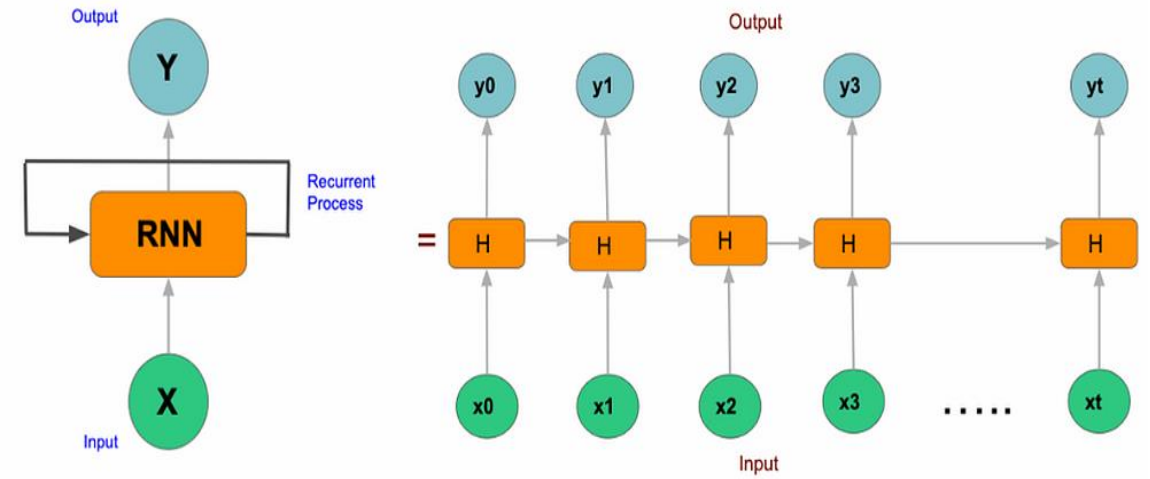


Figure 2. 5Les réseaux de neurones récurrents (RNN) [46]

Les RNN ont une « mémoire» qui mémorise toutes les informations sur ce qui a été calculé. Il utilise les mêmes paramètres pour chaque entrée car il effectue la même tâche sur toutes les entrées ou couches masquées pour produire la sortie. Cela réduit la complexité des paramètres, contrairement aux autres réseaux de neurones **[47].**

5.3. Les réseaux Long Short-Term Memory (LSTM)

Les réseaux de mémoire à long terme à court terme généralement appelés simplement «LSTM» - sont un type spécial de RNN, capable d'apprendre les dépendances à long terme. Ils ont été introduits par Hochreiter & Schmidhuber (1997) et ont été affinés et popularisés par de nombreuses personnes dans le cadre de leurs travaux. Ils travaillent très bien sur une grande variété de problèmes et sont maintenant largement utilisés.

Les LSTM sont explicitement conçus pour éviter le problème de dépendance à long terme. Se souvenir d'une information pendant de longues périodes est pratiquement leur comportement par défaut, pas quelque chose qu'ils ont du mal à apprendre. Tous les réseaux neuronaux récurrents ont la forme d'une chaîne de modules répétitifs de réseau neuronal. Dans les RNN standard, ce module répétitif aura une structure très simple, telle qu'une couche de tanh unique **[48]**.

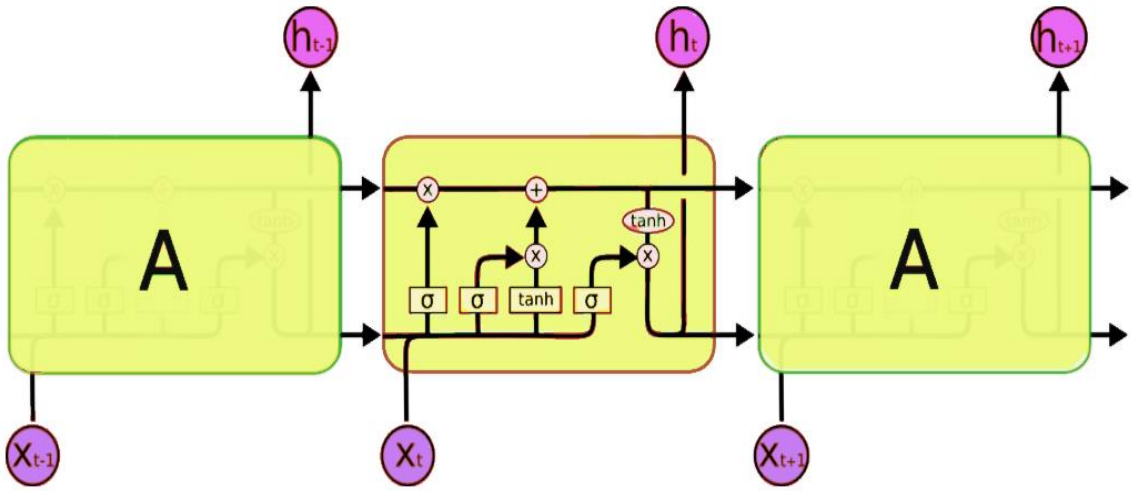


Figure 2. 6:Le module de répétition dans un LSTM [49]

6. Domaines d'application

L'apprentissage profond a de nombreuses applications en informatique, on cite quelques domaines:

- Reconnaissance d'image
- Traduction automatique
- Voiture autonome
- Diagnostic médical
- Recommandations personnalisées
- Robots intelligents
- Reconnaissance de la parole
- Traitement du langage naturel

6.1.Le traitement du langage naturel

Le traitement du langage naturel (TLN, ou NLP en anglais) est la capacité pour un programme informatique de comprendre le langage humain tel qu'il est parlé. Il fait partie des technologies d'intelligence artificielle.

Le développement d'applications TLN est difficile parce que traditionnellement les ordinateurs sont conçus pour que les humains leur « parlent » dans un langage de

programmation précis, sans ambiguïté et extrêmement structuré, ou à l'aide d'un nombre limité de commandes vocales clairement énoncées. Or le discours humain n'est pas toujours précis, il est souvent ambigu et sa structure linguistique peut dépendre d'un grand nombre de variables complexes, notamment l'argot, les dialectes régionaux et le contexte social **[50].**

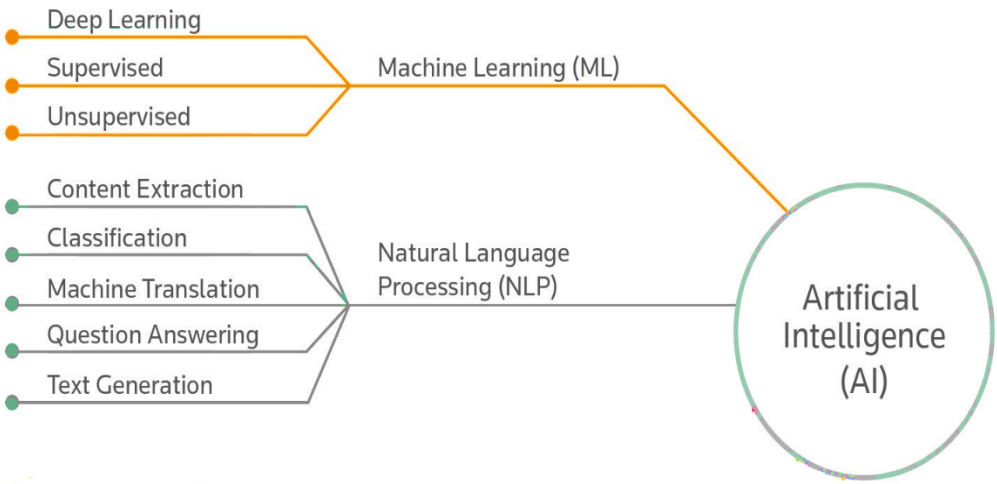


Figure 2. 7:Schéma aux 3 domaines NLP, IA et DL [51]

6.1.1. Taches de traitement du langage naturel

Dans cette section, nous présentons brièvement les quatre tâches standard du NLP.

❖ **Part-Of-Speech Tagging**

C'est une partie essentielle du traitement du langage naturel. C'est le processus de conversion d'une phrase en formes - liste de mots, liste de groupe (où chaque groupe a une forme (mot, étiquette)). Le signe d'état fait partie du signe de la parole et indique si le mot est un nom, un adjectif, un verbe, etc. **[52].**

Exemple

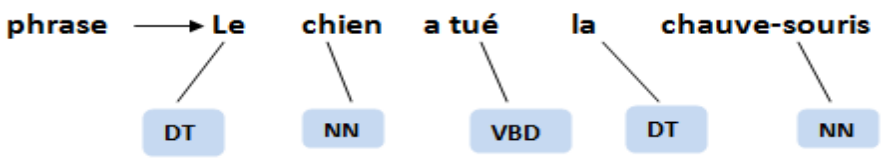


Figure 2. 8:Exemple de Part-Of-Speech Tagging

Liste du Tag & description

La figure 9 ci-dessous représente des quelques Tags et leurs descriptions.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection

Figure 2. 9:Tags et leurs descriptions.

❖ Named Entity Recognition (NER)

La reconnaissance d'entité nommée (NER) - également appelée identification d'entité ou extraction d'entité - est une technique de traitement du langage naturel (NLP) qui identifie automatiquement les entités nommées dans un texte et les classe dans des catégories prédéfinies. Les entités peuvent être des noms de personnes, d'organisations, de lieux, d'heures, de quantités, de valeurs monétaires, de pourcentages, etc.

Avec la reconnaissance d'entités nommées, vous pouvez extraire des informations clés pour comprendre en quoi consiste un texte, ou simplement les utiliser pour collecter des informations importantes à stocker dans une base de données **[53].**

- **Exemple**

Figure 2. 10:Exemple de la tache NER

❖ **Semantic Role Labeling (SRL)**

L'étiquetage des rôles sémantiques (SRL) est une tâche du traitement du langage naturel (NLP) qui vise à attribuer automatiquement des rôles sémantiques à chaque argument pour chaque prédicat dans une phrase d'entrée donnée.

❖ **Parsing ou Chunking**

Est le processus qui consiste à déterminer la structure syntaxique d'un texte en analysant ses mots constitutifs sur la base d'une grammaire sous-jacente (du langage).

- **Exemple**

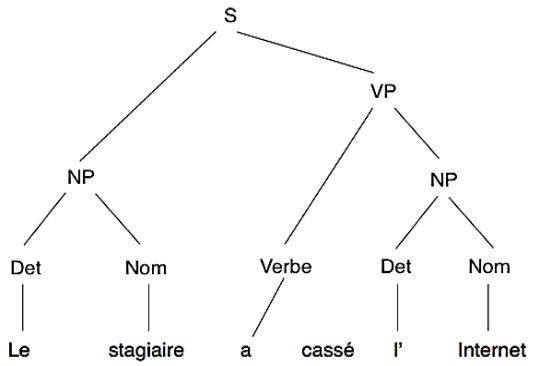


Figure 2. 11:Exemple de la tache Parsing

6.1.2. Objectif

Le domaine du traitement du langage naturel (NLP) vise à convertir le langage humain en une représentation formelle facile à manipuler par les ordinateurs pour étudier des problèmes fondamentaux du traitement de la langue naturelle, ce qui est bien adapté à la modélisation des données textuelles afin d'en extraire des informations et, éventuellement, de représenter les mêmes informations différemment (**[54]** ; **[55]** ; **[56]**).

6.1.3. Domaines d'application du NLP

Le NLP est terme assez générique qui recouvre un champ d'application très vaste. Voici les applications les plus populaires **[57] :**

- Traduction automatique
- Analyse des sentiments
- Marketing
- Chatbots

Autre domaine d'application

- Classification de texte
- Reconnaissance de caractères
- Correction automatique
- Résumé automatique

7. L'apprentissage profond et le traitement de langage naturel

La plupart de ces technologies PNL sont alimentées par le Deep Learning. La plupart des méthodes d'apprentissage automatique fonctionnent bien en raison des représentations et des fonctionnalités d'entrée conçues par l'homme, ainsi que de l'optimisation du poids pour mieux faire une prédiction finale. D'autre part, dans l'apprentissage en profondeur, l'apprentissage des représentations tente d'apprendre automatiquement de bonnes fonctionnalités ou représentations à partir d'entrées brutes. Les fonctionnalités conçues manuellement dans l'apprentissage automatique sont souvent sur-spécifiées, incomplètes et prennent beaucoup de

temps à concevoir et à valider. En revanche, les fonctionnalités apprises de l'apprentissage en profondeur sont faciles à adapter et rapides à apprendre.

Le Deep Learning fournit un cadre très flexible, universel et apprenable pour représenter le monde, à la fois pour des informations visuelles et linguistiques. Au départ, cela a permis des percées dans des domaines tels que la reconnaissance vocale et la vision par ordinateur. Récemment, les approches d'apprentissage en profondeur ont obtenu des performances très élevées dans de nombreuses tâches PNL différentes. Ces modèles peuvent souvent être formés avec un seul modèle de bout en bout et ne nécessitent pas d'ingénierie de fonctionnalités traditionnelle et spécifique à une tâche.

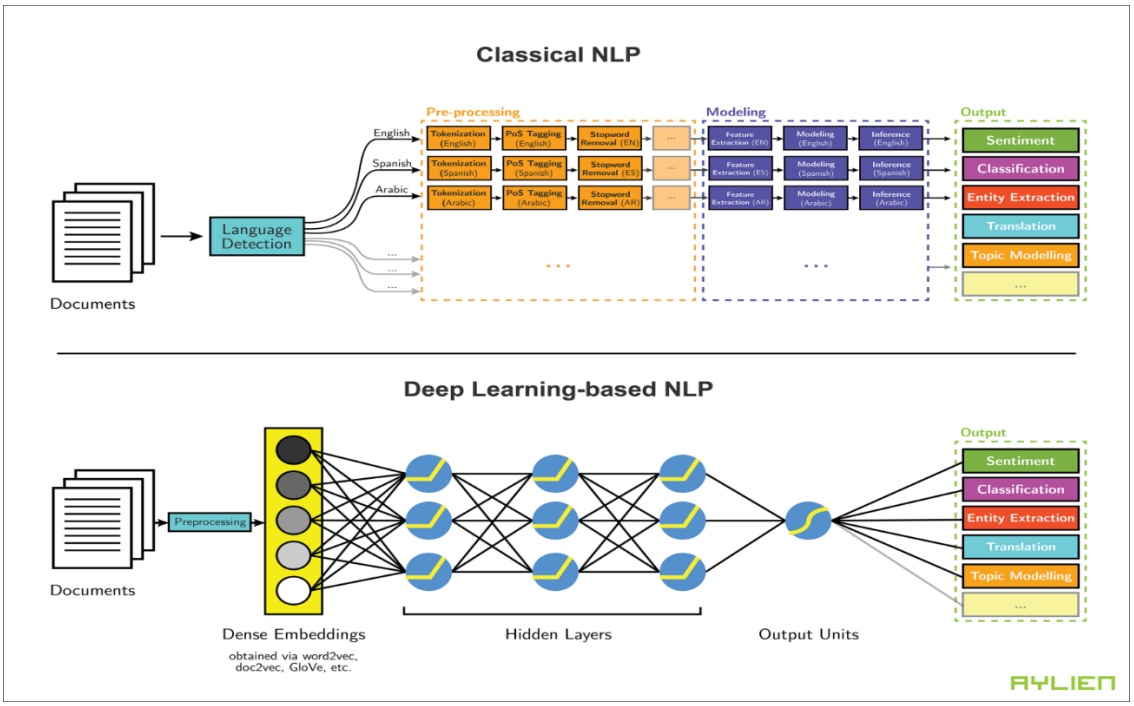


Figure 2. 12:Les applications de NLP Classique vs NLP avec le deep learning [58]

7.1.Les techniques de vectorisation

7.1.1. Bag Of Words

Le sac de mots est une technique de traitement du langage naturel de la modélisation de texte. En termes techniques, on peut dire qu'il s'agit d'une méthode d'extraction de caractéristiques avec des données textuelles. Cette approche est un moyen simple et flexible d'extraire des fonctionnalités à partir de documents.

Un sac de mots est une représentation de texte qui décrit l'occurrence de mots dans un document. Nous suivons simplement le nombre de mots et ignorons les détails grammaticaux et l'ordre des mots. C'est ce qu'on appelle un «sac» de mots parce que toute information sur l'ordre ou la structure des mots dans le document est supprimée. Le modèle se préoccupe uniquement de savoir si les mots connus apparaissent dans le document, et non à l'endroit du document.

7.1.2. TF-IDF

L'acronyme TF-IDF pour Term Frequency & Inverse Document Frequency est une puissante technique d'ingénierie de fonctionnalités utilisée pour identifier les mots importants ou plus précisément les mots rares dans les données textuelles.

La valeur TF-IDF augmente proportionnellement au nombre de fois qu'un mot apparaît dans le document et est compensée par le nombre de documents dans le corpus qui contiennent le mot, ce qui permet d'ajuster le fait que certains mots apparaissent plus fréquemment en général **[59].**

Pour un terme « i » dans un document « j » :

$$\boldsymbol{Wi,j} = \boldsymbol{tfi,j} * \mathbf{log}\,(\boldsymbol{Ndfi})$$

Où :

$\boldsymbol{tfi}$, *est* nombre d'occurrence de i dans j

$\boldsymbol{dfi}$, *est* le nombre de document contenant i

N, est le nombre total des documents.

7.1.3. Word Embedding

L'incorporation de mots est un sujet de base important en PNL. Étant donné que les ordinateurs ne peuvent pas traiter directement le langage naturel, l'incorporation de mots peut transformer le langage naturel en valeurs traitables par ordinateur. La méthode de représentation de mot très basique est appelée représentation instantanée, mais cette méthode n'inclut pas les informations du contexte d'un mot. Pour résoudre ce problème, de nombreuses

méthodes d'inclusion de mots basées sur le contexte ont été mises au point. L'un des outils les plus impressionnants et les plus puissants appelé Word2Vec est proposé **[60].**

Après Word2Vec, Glove **[61]** a été proposé pour améliorer la qualité de la représentation en utilisant des informations statistiques globales sur les mots combinés à des informations contextuelles**.**

- **Word2Vec**

S'appuie sur un simple perceptron multi-couches (réseau de neurone) à une couche cachée (de taille N) où la tâche est de prédire le mot en fonction du contexte ou réciproquement. La vectorisation en dimension N est fournie par l'ensemble des poids des neurones de la couche cachée [62].

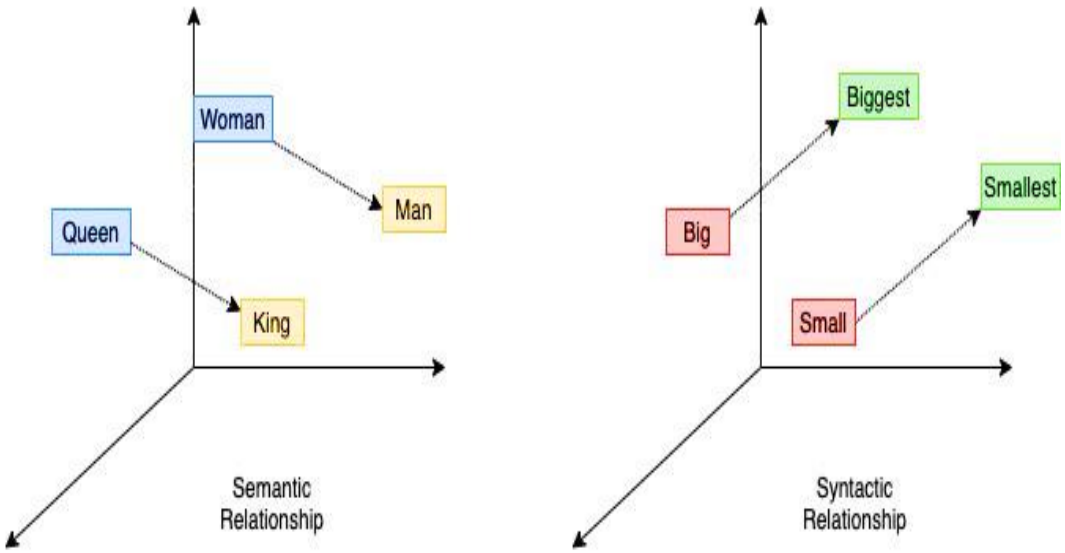


Figure 2. 13:Vecteurs Word2Vec formés avec relation sémantique et syntaxique [63].

- **GloVe**

GLOVE fonctionne de la même manière que Word2Vec. Alors que vous pouvez voir ci-dessus que Word2Vec est un modèle «prédictif» qui prédit un mot donné par le contexte, GLOVE apprend en construisant une matrice de cooccurrence (mots X contexte) qui compte essentiellement la fréquence d'apparition d'un mot dans un contexte. Comme il s'agira d'une matrice gigantesque, nous factorisons cette matrice pour obtenir une représentation de dimension inférieure. Il y a beaucoup de détails dans GLOVE mais c'est l'idée approximative**[64].**

8. Travaux Connexes : état de l'art

Dans cette section, nous présentons l'état de l'art des travaux réalisés en AOA utilisant des réseaux de neurones artificiels. Nous commençons par présenter les méthodes neuronales utilisées et les prétraitements appliqués, et nous finissons par une synthèse. Le tableau 2.2 résume ces travaux en les classant par ordre d'apparition chronologique

Ref	Niveau	Méthodes	Corpus	Résultats
[65]	Phrase (ASM et AD)	DNN, CNN	2026 tweets	CNN= 90%, DNN= 85%
[66]	Phrase (ASM et AD)	DNN	Tweets	DNN= 90.2%
[67]	Phrase (ASM et AD)	CNN, LSTM, CNN+LSTM LSTM-comb	ASTD	CNN=74.1%, LSTM= 80.1%, CNN+LSTM= 73.5%,
			ArTwitter	LSTM-comb=81.6% CNN= 83.2%, LSTM= 83.7%, CNN+LSTM=84.2%, LSTM-comb=87.3%
[68]	Document (ASM et AD)	CNN, LSTM	BRAD	CNN= 89.61%, LSTM=90.05%
[69]	Phrase (ASM et AD)	CNN + lexique	2026 tweets	CNN=92%
[70]	Phrase	CNN, BiLSTM, Vote	ASTD	CNN=64.3%, BiLSTM=64.7%, Vote=65.1%
[71]	Phrase	CNN+LSTM	ASTD, ArTwitter	ASTD= 88.1%, ArTwitter=76.4%
[72]	Phrase	CNN, LSTM, CNN+LSTM	40k tweets	CNN=75.7%, LSTM=81.3%, CNN+LSTM= 78.5%
[73]	Phrase	DE-CNN	ASTD, ArTwitter	ASTD=81.1%, ArTwitter=91.8%

Tableau 2. 2:Travaux réalisés en AOA.

9. Synthèse

Le but des systèmes d'analyse d'opinion est de déterminer la polarité, où le degré de cette dernière varie de 1 à 5 niveaux :

- Positif et négatif
- Positif, Neutre et Négatif
- Très positif, positif, neutre, négatif, extrêmement négatif
- Positif, négatif, neutre et mixte

Les travaux les plus connus dans ce domaine sont les travaux basés sur les réseaux de neurones de classification binaire et trinaire, ainsi que la classification quinaire (5 classes) , qui est également considérée comme une classification binaire en raison de la collecte d'opinions allant de 3 à 5 à des opinions positives , et entre 1 et 2 aux avis négatifs. Peu de travaux ont été effectués sur la classification quaternaire (4 classes) , compte tenu des catégories additives positives, négatives, neutre et mixtes.

Des études ont montré que les réseaux de neurones sont plus efficaces que les classificateurs de base traditionnels tels que (SVM, NB, KNN**) [68].** Les réseaux de neurones récursifs se sont avérés efficaces dans l'analyse d'opinion au niveau de la phrase en ASM **[68]** et les réseaux convolutifs et récursifs se sont également avérés efficaces pour l'analyse d'opinion au niveau de la phrase et document.

La première couche de chaque réseau est basée sur les Word embeddings quel que soit prétraité ou non.

Semblable aux langues étrangères, la langue arabe n'a pas été très populaire dans la recherche pour analyser et catégoriser les opinions.

Où il est classé parmi les recherches récentes et nouvelles dans le domaine du traitement du langage naturel.

La classification des textes arabes est l'un des problèmes graves de la classification à l'aide d'algorithmes d'apprentissage automatique. Atteindre une grande précision dans la classification du texte arabe dépend principalement des techniques de prétraitement utilisées pour préparer l'ensemble de données.

De plus, l'ensemble de données n'est pas aussi volumineux que les données disponibles en anglais. Dans certains cas, il peut être petit et insuffisant pour construire un modèle intégré.

Une autre difficulté qui imprègne l'ensemble de données arabe est que la majorité des groupes sont déséquilibrés. Le déséquilibre de groupe peut dégrader les performances du modèle par rapport aux groupes minoritaires avec des classificateurs de base ou des réseaux de neurones profonds.

En plus de la taille du corps et de la répartition de la polarité, la structure du document est complexe et peut consister en un seul mot, un groupe de mots, une phrase ou plusieurs phrases.

Notre travail dans ce mémoire vise à essayer de surmonter ces lacunes et à valoriser la langue arabe, car c'est notre langue maternelle et la langue utilisée pour évaluer les hôtels, les produits et autres sur les sites Internet.

Nous recherchons principalement un ensemble de données d'opinions en arabe, puis essayons d'utiliser des modèles d'apprentissage automatique pour les analyser et obtenir des résultats pouvant être prêts à être utilisés dans les applications requises pour analyser les opinions.

10. Conclusion

L'apprentissage profond est le domaine le plus émergeant de l'apprentissage automatique et a apporté une contribution importante dans divers domaines de recherche. Cela a permis de surmonter les inconvénients des méthodes traditionnelles en rendant les systèmes moins complexes et plus rapides. L'apprentissage profond a été utilisé avec le traitement automatique du langage dans plusieurs domaines de recherche, ce qui est très prometteur et constitue un succès. Dans ce chapitre nous avons exposé la technique de l'apprentissage profond, ainsi que ses avantages, et ses limites, le traitement du langage naturel, et les différentes méthodes de construction d'embeddings. À la fin, nous avons établi l'état de l'art des méthodes neuronales pour l'analyse d'opinions en arabe. Nous allons présenter, dans le reste de ce thème, notre approche neuronale pour l'analyse d'opinions en arabe.

Chapitre 3

Conception & Réalisation

1. Introduction

Dans ce chapitre, nous présentons le système d'analyse d'opinion Arabe basé sur des algorithmes d'apprentissage profond. Tout d'abord, nous commençons par une introduction au langage Arabe, ensuite nous présentons une brève définition de l'architecture de classification automatisé populaire dans tout système d'analyse de données, qui est la collecte, le traitement, la classification et l'évaluation des données, puis nous présenterons notre contribution à la classification proposée en détaillant les différentes unités qui le composent.

2. Le langage Arabe

La langue Arabe est considérée comme l'une des langues utilisées, car elle se classe cinquième en termes de mots les plus utilisés et également quatrième langue les plus utilisées sur Internet (statistiques pour l'année 2020).

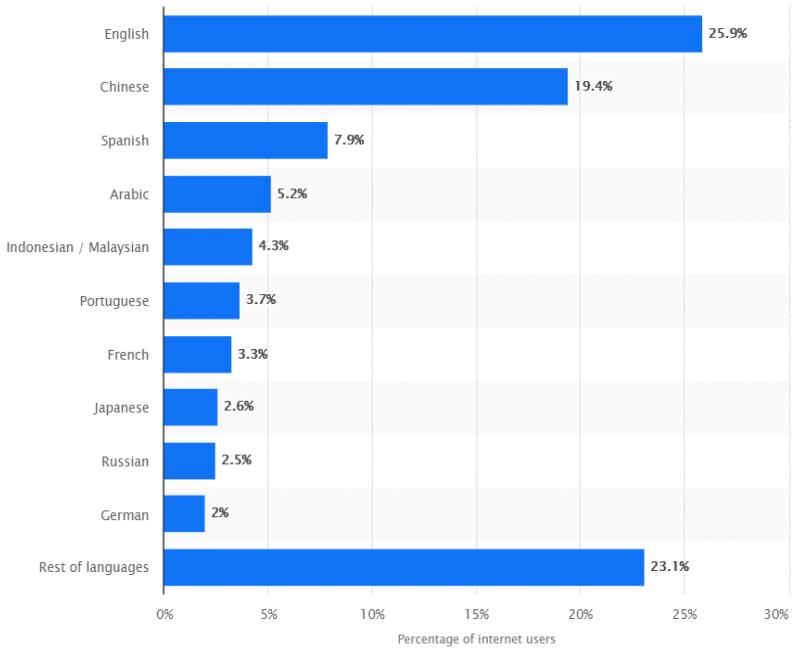


Figure 3. 1 : Langues les plus couramment utilisées sur Internet 2020[1]

[1] *Source : www.statista.com*

De plus, l'Arabe est utilisé par près de 500 millions de personnes :

- ✓ 300 millions comme première langue
- ✓ Le reste comme deuxième langue.

La langue Arabe a une écriture particulière. Contrairement à l'anglais, au français et à toutes les langues occidentales en général :

- ✓ la langue Arabe s'écrit de droite à gauche
- ✓ Il n'y a ni minuscule ni majuscule dans les lettres Arabes.
- ✓ La forme de certaines lettres peut changer selon leur position dans le mot :

  Par exemple la lettre « ع » peut s'écrire de trois manières différentes selon sa place dans le mot :

  - au début du mot « , عـ : علبة » ,
  - au milieu du mot « jeu, ـعـ : لعبة »
  - et à la fin du mot «ع :مربع : ـع » .

C'est l'une des langues les plus difficiles au monde avec sa morphologie riche, sa syntaxe complexe et sa sémantique difficile. Cela rend son analyse et son traitement automatique très difficiles et complexes.

La langue Arabe est composée de 28 lettres (25 consonnes et trois longues voyelles).Contrairement aux autre langages, qui a des lettres dédiées pour représenter les voyelles courtes, l'Arabe a des diacritiques qui jouent le même rôle que les voyelles courtes en anglais et déterminent la prononciation .

Lettre arabe	Symbole	Lettre arabe	symbole	Lettre arabe	Symbole	Lettre arabe	symbole
ء	'	د	d	ض	ḍ	ك	k
ب	b	ذ	ḏ	ط	ṭ	ل	l
ت	t	ر	r	ظ	ẓ	م	m
ث	ṯ	ز	z	ع	ʿ	ن	n
ج	j	س	s	غ	ġ	ه	h
ح	ḥ	ش	š	ف	f	و	w
خ	ḫ	ص	ṣ	ق	q	ي	y

Voyelles brèves		Voyelles longues	
ـَ	a	ا	ā
ـُ	u	و	ū
ـِ	i	ي	ī

Figure 3. 2: Les lettres de langage Arabe

L'Arabe a ses propres caractéristiques qui présentent des défis pour les applications de PNL La classification des textes ne fait pas exception Contrairement à l'anglais, l'Arabe est une langue

d'agglutination où plusieurs éléments (racines de mots, préfixes de mots ou suffixes de mots) peuvent être liés pour former un seul mot écrit. Cela rend le processus de radicalisation/encodage plus difficile en Arabe, contrairement aux langues latines qui sont plus faciles à séparer avec des espaces blancs.

3. Méthodologie

Notre sujet de recherche est lié à l'analyse des sentiments pour la langue Arabe, et pour améliorer les performances de l'analyse, nous avons proposé une architecture de type CNN pour l'analyse d'opinion Arabes.

L'approche qu'on propose est une classification des opinions exprimées dans des phrases. Cette approche s'intègre dans la classe des méthodes d'analyse au niveau texte.

3.1.Architecture générale

Le script général représente les étapes nécessaires dans tout modèle de traitement de texte basé sur l'apprentissage profond, nous collectons d'abord les données textuelles en fonction du type de problème, puis nous pré-traitons les données textuelles et les divisons en deux parties (données d'entraînement et test données), comme le montre la figure suivante. Ensuite, nous construisons un modèle de trainement qui prend les entrées de données de formation traitées à l'étape précédente, puis les vérifie sur l'ensemble de données de test afin de calculer l'efficacité du modèle et produire un système final qui fait la classification du nouveau texte.

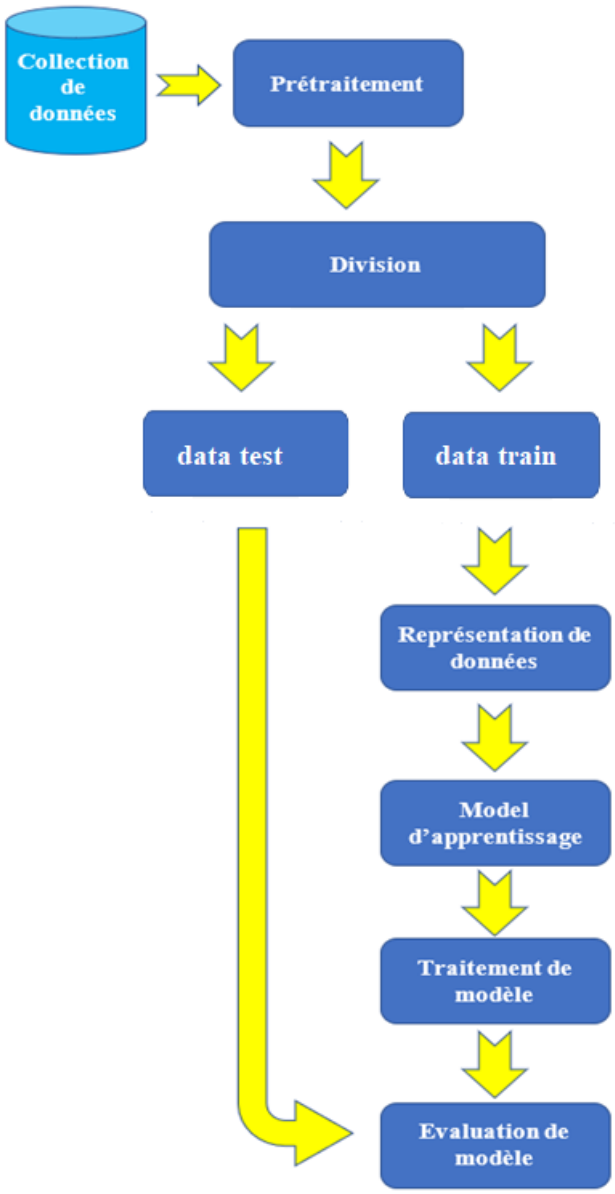


Figure 3. 3: Le processus générale des application NLP avec le deep learning

3.2.Architecture détaillée

L'architecture détaillée de notre système est illustrée dans la figure suivante, et nous l'expliquerons dans la section suivante.

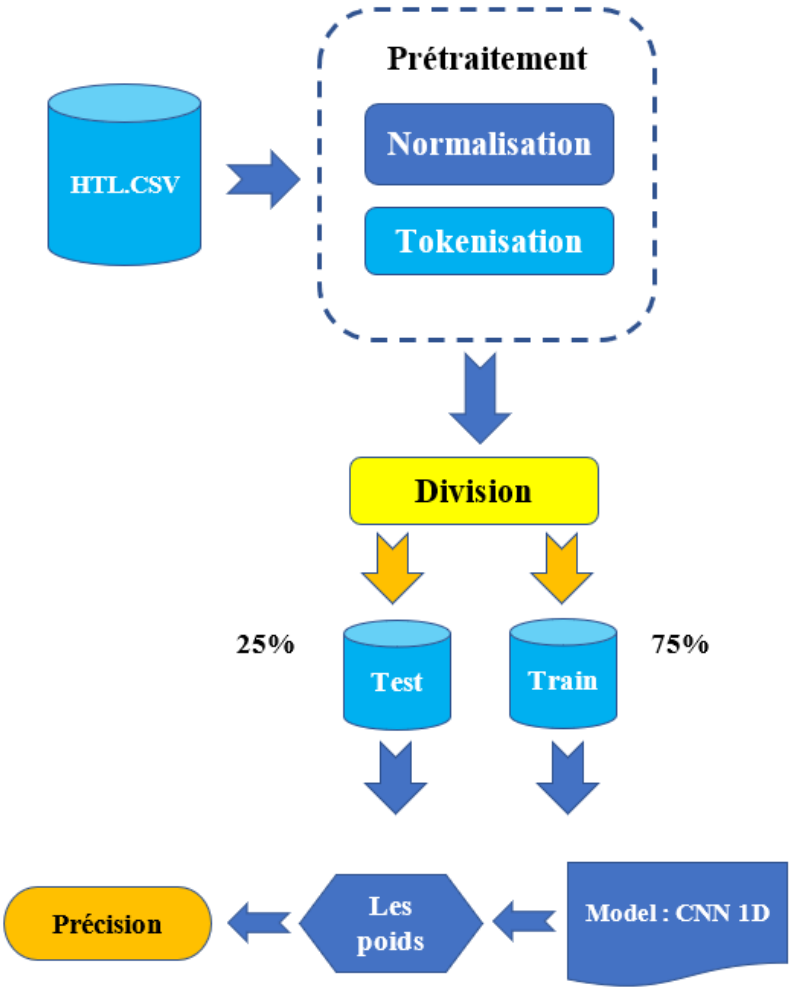


Figure 3. 4: Architecture générale de l'approche proposée

3.2.1. Description de dataset

L'ensemble de données décrit dans l'article [74], qui a remporté le troisième prix du meilleur article lors de la conférence internationale CICLing2015 sur la linguistique informatique et le traitement de texte intelligent, a été utilisé.

Le référentiel comprend les éléments suivants : 33K avis annotés automatiquement dans les domaines des films, des hôtels, des restaurants et des produits

Étant donné que la création de model se fait sur notre ordinateur personnel et en utilisant également la plate-forme Google Colab pendant 12 heures par jour seulement, cela ne suffit pas pour former le modèle et obtenir de bons résultats, nous avons donc choisi uniquement l'ensemble de données pour les opinions sur les hôtels qui est collecté depuis le site de

TripAdvisor.com , qui contient 15572 reviews :

- Positive review : 10775
- Négative review :2647
- Neutre review : 2150

3.2.2. Prétraitement

L'étape de prétraitement englobe diffèrent technique de traitement de texte tel que la normalisation, la tokenisation, la lemmatisation... Dans notre cas en s'intéresse seulement par l'application de normalisation et de tokenisation.

- **Normalisation**

Dans la phase de normalisation nous avons applique diffèrent étape pour que le jeu de données final soit bien nettoyer. Le code suivant résume les différentes étapes appliquées qui sont tout trouvé dans la bibliographie re de python :

- ✓ **remove_diacritics(string)**: supprime tous les signes diacritiques d'une chaîne et renvoie la version nettoyée
- ✓ **remove_numbers(string)**: supprime tous les nombres d'une chaîne et renvoie la version propre
- ✓ **remove*non*arabic_words(string)**: supprime tous les mots non Arabes (ont un symbole non arabe) d'une chaîne et renvoie la version nettoyée
- ✓ **remove*extra*whitespace(string)**: supprime les espaces supplémentaires d'une chaîne et renvoie la version propre
- ✓ **remove*non*arabic_symbols(string)**: supprime tous les symboles non Arabes d'une chaîne et renvoie la version propre
- ✓ **remove_ponctuations(string)**: supprime toutes les ponctuations d'une chaîne et renvoie la version propre
- ✓ **remove*dubplicated*letters(string)**: supprime les lettres en double et renvoie la chaîne de résultat

```
def clean_text(text):
    search = ["أ","إ","آ","ة","_","-","/",".","،"," و "," يا ",'"',"-","'","ى",
              "\\",'\n', '\t','&quot;','?','؟','!']
    replace = ["ا","ا","ا","ه"," "," ","","","","و","يا",
               "","","","ي","",' ', ' ',' ',' ? ',' ؟ ', ' ! ']
        tashkeel = re.compile(r'[\u0617-\u061A\u064B-\u0652]')
    text = re.sub(tashkeel,"", text)
        longation = re.compile(r'(.)\1+')
    subst = r"\1\1"
    text = re.sub(longation, subst, text)
    text = re.sub(r"[^\w\s]", '', text)
    text = re.sub(r"[a-zA-Z]", '', text)
    text = re.sub(r"\d+", ' ', text)
    text = re.sub(r"\n+", ' ', text)
    text = re.sub(r"\t+", ' ', text)
    text = re.sub(r"\r+", ' ', text)
    text = re.sub(r"\s+", ' ', text)
    text = text.replace('وو', 'و')
    text = text.replace('يي', 'ي')
    text = text.replace('اا', 'ا')
        for i in range(0, len(search)):
        text = text.replace(search[i], replace[i])
        text = text.strip()
        return text
```

Figure 3. 5: Code nettoyage de jeux de données.

- **Tokenisation**

La tokenisation consiste essentiellement à diviser une phrase, un paragraphe ou un document texte entier en unités significatifs plus petites, telles que des mots ou des termes individuels. Chacune de ces petites unités est appelée **Token**. Apprès cette phase en applique directement **le Padding** pour que tous les review soit de meme taille (la taille de le review maximal).

3.2.3. Protocole de Train et Test

Dans la phase de division des données nous avons choisi de souffler les donnes et les diviser en 2 parties :

1. **Train :** Partie de jeux de données (échantillon) pour la phase l'apprentissage dans les réseaux de nouerons.
2. **Test :** Partie de jeux de données (échantillon) utilisées uniquement pour évaluer les performances d'un modèle final.

3.2.4. Création de model

Dans notre proposition nous avons créé un modèle avec la couche embedding et un algorithme CNN-1D.

Dans ce qui suit on présente l'architecture de notre modèle :

```
Build model...
Model: "sequential_12"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_12 (Embedding)     (None, 300, 300)          3000000
_________________________________________________________________
dropout_23 (Dropout)         (None, 300, 300)          0
_________________________________________________________________
conv1d_20 (Conv1D)           (None, 298, 250)          225250
_________________________________________________________________
conv1d_21 (Conv1D)           (None, 296, 250)          187750
_________________________________________________________________
max_pooling1d_7 (MaxPooling1 (None, 148, 250)          0
_________________________________________________________________
flatten_11 (Flatten)         (None, 37000)             0
_________________________________________________________________
dense_22 (Dense)             (None, 250)               9250250
_________________________________________________________________
dropout_24 (Dropout)         (None, 250)               0
_________________________________________________________________
activation_22 (Activation)   (None, 250)               0
_________________________________________________________________
dense_23 (Dense)             (None, 1)                 251
_________________________________________________________________
activation_23 (Activation)   (None, 1)                 0
=================================================================
Total params: 12,663,501
Trainable params: 12,663,501
Non-trainable params: 0
```

Figure 3. 6: Architecture de model CNN-1D proposé.

Le modèle que nous présentons est composé d'une couche de embedding, deux couches de convolution et une couche de maxpooling , flatten et et deux couches de sortie. Les informations en entrée sont de nombre maximum des mots 300 , et la dimension de la représentation vectoriel 300.

L'information passe d'abord à les 2 premières couches de convolution. Chaque couche est composée de 250 filtres de taille 3.

Ensuite on applique Maxpooling pour réduire la taille de l'information ainsi la quantité de paramètres et de calcul. À la sortie de cette couche, nous aurons 148 features maps de taille 250.

En applique la couche de flatten , le résultat est vecteur de caractéristiques à une dimension de 37000.

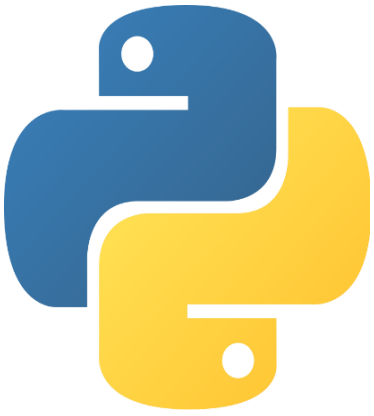
Finalement on a créé une couche cache de 250 neurones , et une couche de sortie contient un seule nouerons qui utilisée la fonction d'activation un sigmoïde qui permet de calculer la distribution de probabilité des 2 classes (positive et négative).

4. Implémentation

Dans cette partie nous présenter dans une première partie le langage de programmation python utilisé dans l'implémentation de model et les outils software tel que les bibliographies impoerter : **Keras , pickle , Matplotlib , Numpy , Sklearn** …

Et le hardware utilise pour implémenter notre modèle, ensuite nous expriment les résultats obtenus.

4.1. Le langage de programmation



Python : Version 3.7.10

Est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions.

4.2.Le software

Kears : Version 2.5.0

Est considéré comme une puissante de la bibliothèque Python, facile à utiliser pour développer et évaluer des modèles d'apprentissage en profondeur. Il a un design minimaliste qui permet de construire un réseau couche par couche ; l'entraîner et l'exécuter. Il englobe les bibliothèques de calcul numérique efficaces Theano et TensorFlow et permet de définir et de former des modèles de réseaux neuronaux en quelques courtes lignes de code. Il s'agit d'une API (Application programming interface) de réseau neuronal de haut niveau, aidant à utiliser largement l'apprentissage profond et l'intelligence artificielle. Il s'exécute au-dessus d'un certain nombre de bibliothèques de niveau inférieur, notamment TensorFlow, Theano, etc.

Pickle : Version 0.0.11

Le module pickle implémente des protocoles binaires pour sérialiser et désérialiser une structure d'objet Python. "Pickling" est le processus par lequel une hiérarchie d'objets Python est convertie en un flux d'octets, et "unpickling" est l'opération inverse, par laquelle un flux d'octets (à partir d'un fichier binaire ou d'un objet de type octets) est reconverti en une hiérarchie d'objets.

Sklearn : Version 0.22.2. post1

Scikit-learn est une bibliothèque Python, libre et dédiée à l'apprentissage automatique. Elle comprend des fonctions pour estimer des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support. Elle est conçue pour s'harmoniser avec les autres bibliothèques libre Python, notamment NumPy et SciPy

Matplotlib : Vesrion 3.2.2

Est une bibliothèque de traçage pour le langage de programmation Python et son extension mathématique numérique NumPy C'est un logiciel de premier ordre qui fait de Python un concurrent averti à des outils scientifiques tels que MatLab ou Mathematica. elle conçue pour la génération de visualisations simples et puissantes . La bibliothèque est prise en charge par différentes plates-formes et utilise différents kits d'interface graphique pour la représentation des visualisations résultantes. Les différents IDE (comme IPython) prennent en charge la fonctionnalité de Matplotlib.

Numpy : Version 1.19.5

Numpy est une bibliothèque pour effectuer des opérations arithmétiques numériques en Python. Le package de base autour duquel se construit la pile de calcul scientifique est appelé numpy (Numerical PYthon).

La bibliothèque Numpy fournit une gestion plus facile des tables de nombres et des fonctions complexes (propagation).

4.3.Le hardware

Le Deep Learning est un domaine avec des exigences en calculs intenses et la disponibilité des ressources (surtout en GPU) dédiés à cette tache vont fondamentalement influencer sur l'expérience de l'utilisateurcar sans ses ressources, il faudra trop de temps pour apprendre de ses erreurs ce qui peut être décourageant.alors nous avons choisi la platformes Google Colaboratory pour pour effectueé les expérimentations .

Google Colab

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique.

Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le cloud.

5. Réalisation

Pour notre model on a utilisé le model CNN-1d à l'aide de la couche d'embedding de dimension 300 et le nombre d'époque de 10.

Dans les figures suivants, la précision de l'apprentissage augmente avec le nombre d'époque, ceci reflète qu'à chaque époque le modèle apprenne plus d'informations. Aussi la validation augmente de la même façon.

De même, l'erreur d'apprentissage et diminue, d'autre part la validation diminue aussi avec le nombre d'époque.

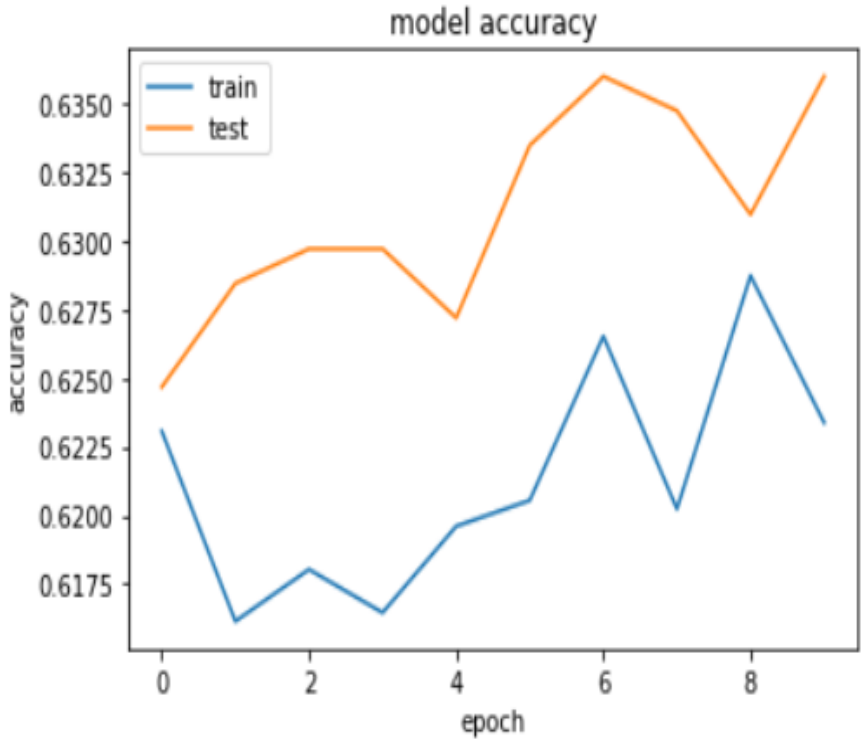


Figure 3. 7: precison de model

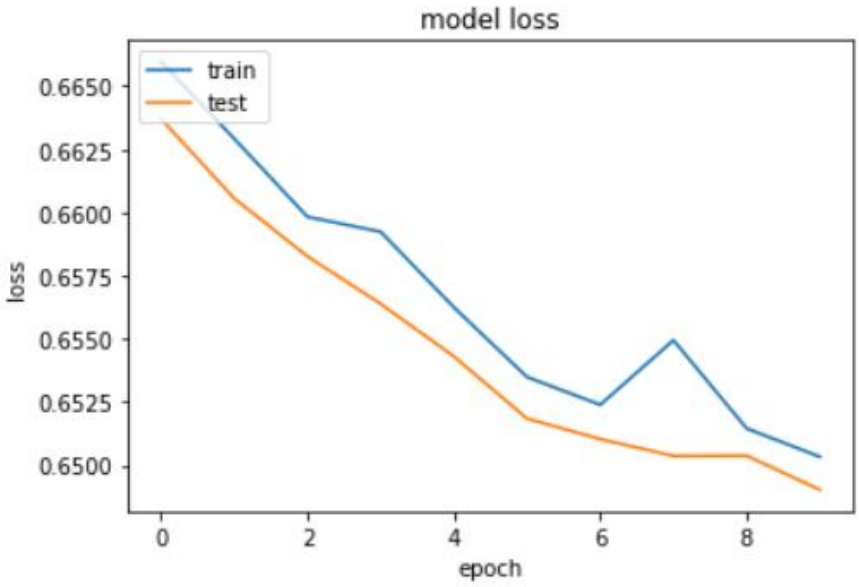


Figure 3. 8: erreur de model

La figure suivantes montre la matrice de confusion de la prédiction du notre model ainsi que le rapport de classification , le taux de précision est de 64 %. Le modèle a mal classé 480 reviews (177 + 303).

Le rapport de classification montre la Précision, le Rappel et le F-score de la prédiction du notre modèle, le rappel pour les données étiqueté en 0 est de 67% et pour les donnes étiquetées en 1 est de 61 %. Ce que signifie que le percentage de bais est tré faible.

```
[[363 177]
 [303 480]]
              precision    recall  f1-score   support

           0       0.55      0.67      0.60       540
           1       0.73      0.61      0.67       783

    accuracy                           0.64      1323
   macro avg       0.64      0.64      0.63      1323
weighted avg       0.65      0.64      0.64      1323
```

Figure 3. 9: Matrice de confuison + Rapport de classification

Dans ce modèle en va prendre le modèle précèdent et en va faire une augmentation de 100 dans le nombre d'itération.

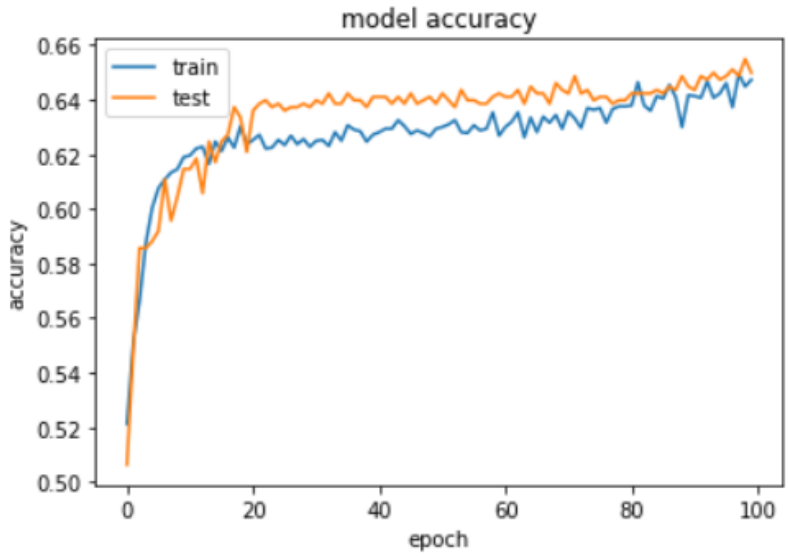


Figure 3. 10:precison de model

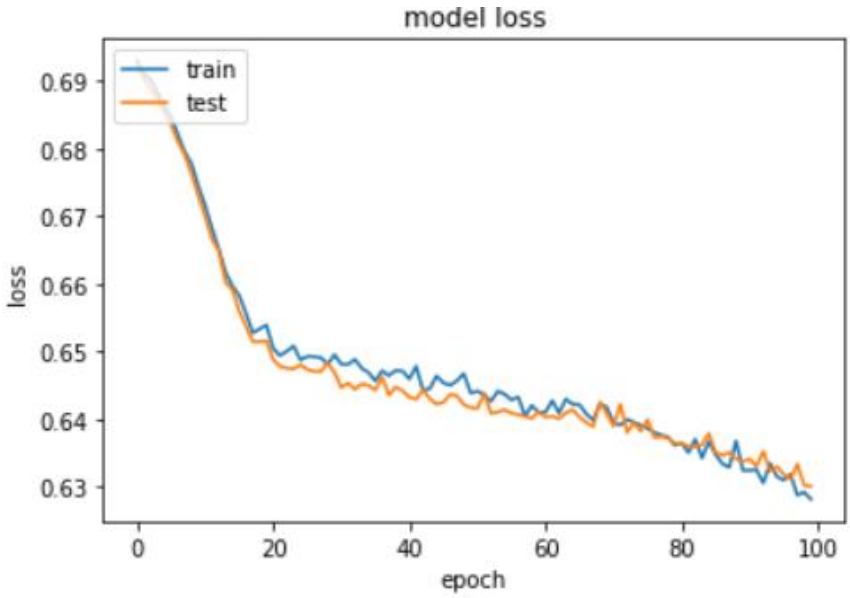


Figure 3. 11erreur de model

```
[[433 220]
 [233 437]]
              precision    recall  f1-score   support

           0       0.65      0.66      0.66       653
           1       0.67      0.65      0.66       670

    accuracy                           0.66      1323
   macro avg       0.66      0.66      0.66      1323
weighted avg       0.66      0.66      0.66      1323
```

Figure 3. 12:Matrice de confuison + Rapport de classification

On remarque qu'avec l'augmentation de nombre des époques, le modèle ne tombe pas dans le problème de overffing.

Ainsi, nous pouvons constater comme résultat final le meilleur modèle qui donne des bons résultats c'est le model CNN-1D avec 100 époques.

6. Conclusion

Dans ce dernier chapitre nous avons présenté notre proposition et illustrer son implémentation pour la classification des opinions Arabes. L'idée est basé sur le modèle de type CNN-1d avec la couche d'embedding comme un support de représentation de connaissances. Nous avons expliqué notre modèle ainsi que la validation et les divers résultats obtenus. Nous avons conclut que le model « CNN-1D » avec 100 époques de traitement nous donne le meilleur résultat

Conclusion générale

1. Conclusion

Dans ce travail, nous avons exploré le domaine de l'analyse des opinions qui, comme tous les autres domaines du traitement du langage naturel, a connu une évolution majeure depuis les années 2000 et a réalisé une évolution majeure et un grand intérêt depuis la naissance de l'apprentissage profond.

Bien que le sujet de l'analyse d'opinion soit un vaste domaine de recherche, il n'a pas été beaucoup étudie par rapport à la langue Arabe, même si c'est l'une des langues les plus difficiles au monde.

Nous avons discuté quelques concepts de base liés à l'analyse d'opinion, puis nous nous sommes déplacés vers les réseaux de neurones. Méthodes d'apprentissage automatique et profond, en particulier les réseaux de neurones convolutifs et les réseaux de neurones récurrents.

Comme contribution modeste de notre part, nous avons proposé un modèle d'analyse des opinions pour la langue arabe à l'aide de réseaux de neurones convolutifs (CNN) sur un ensemble de données HTL.csv

2. Perspectives

Se basant sur les résultats présentés par le modèle, ils ne sont pas suffisants pour être prêts à l'emploi, et cela est principalement dû au déséquilibre des catégories de données en plus de la différence de longueur de chaque opinion et de l'utilisation de mots d'argot difficiles pour représenter à l'aide de la couche d'embedding :

1. **L'augmentation de dataset**
2. **Appliques d'autres algorithmes de deep Learning tel que le LSTM et le Bi-Lstm**
3. **Analyser les opinions avec leur diacritiques**

Références bibliographies

[1] : https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1 23/2/2021

[2] :Dictionnaire d'informatique, M. GINGUAY, A. LAURET, Masson, 4° édition, 1990

[3] :Dictionnaire d'informatique, M. GINGUAY, A. LAURET, Masson, 4° édition, 1990

[4] : https://www.larousse.fr/dictionnaires/francais/%C3%A9motion/28829 .24/ 2/ 2021

[5] :Gabriel Dabi-Schwebel, Microblogage " microblogging". 14 Avril 2014. AGENCE 1MIN30. https://www.1min30.com/dictionnaire-du-web/microblogage 24/2/2021

[6] :https://www.gingersoftware.com/english-online/spelling-book/confusing-words/objective-subjective 25/2/2021

[7] :Bilal Saberi, Saidah Saad. 2017 Sentiment Analysis or Opinion Mining: A Review

[8] :https://datafranca.org/wiki/Analyse_des_sentiments 1/3/2021

[9] :LIU B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1–167

[10] :[ZIANI Amel2017/2018] : these La recommandation via l'analyse d'opinions Université de Badji Mokhtar Annaba

[11] :Reyes, A., Rosso, P. and Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. Language Resources and Evaluation, 47 (1), pp. 239–268.

[12] :https://artificial-intelligence-deep-learning.blogspot.com/2020/09/lanalyse-des-sentiments-basee-sur-l.html 1/3/2021

[13] :https://www.sciencedirect.com/topics/computer-science/neutral-opinion 1/3/2021

[14] :PONTIKI M., GALANIS D., PAVLOPOULOS J., PAPAGEORGIOU H., ANDROUTSOPOULOS I. & MANANDHAR S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), p. 27–35, Dublin, Ireland: Association for Computational Linguistics.

[15] :Fouille d'opinions méthodes et outils Étude des méthodes existantes de classification de textes d'opinion Université de Larbi Tébessa –Tébessa-

[16] :[Cynthia Van Hee ,2013] : L'analyse des sentiments appliquée sur des tweets politiques : une étude de corpus, Faculté associée de linguistique appliquée Université Bruxelles Belgique, 2013

[17] :Liu, B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167, 2012.

[18] :[Damien Poirier et Françoise Fessant…,2010] : Damien Poirier et Françoise Fessant et Cécile Bothorel et Emilie Guimier de Neef et Marc Boullé, Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films, Revue des Nouvelles Technologies de l'Information RNTI-E-17, 2010, 147-169.

[19] :https://www.irit.fr/publis/SIG/2010_M2R_B.pdf 3/3/2021

[20] :ALTRABSHEH N., GABER M. M. & COCEA M. (2013). Sa-e : sentiment analysis for education. In International Conference on Intelligent Decision Technologies, volume 255, p. 353–362.

[21] :MUNEZERO M., MONTERO C. S., MOZGOVOY M. & SUTINEN E. (2013). Exploiting sentiment analysis to track emotions in students' learning diaries. In Proceedings of the 13th Koli Calling International Conference on Computing Education Research, Koli Calling '13, p. 145–152, New York, NY, USA : ACM.

[22] :SOLÍS-AVILÉS E., ESPINOZA A. H., ORTIZ-ZAMBRANO J. & VARELA-TAPIA E. (2018). Sentiment analysis in education domain: A systematic literature review. In Technologies and Innovation : 4th International Conference, CITI 2018, Guayaquil, Ecuador, November 6-9, 2018, Proceedings, volume 883, p. 285 : Springer.

[23] :SAUNDERS C. H., PETERSEN C. L., DURAND M.-A., BAGLEY P. J. & ELWYN G. (2018). Bring on the machines: Could machine learning improve the quality of patient education materials ? a systematic search and rapid review. JCO clinical cancer informatics, 2, 1–16

[24] :CLARK E. M. (2019). Applications in sentiment analysis and machine learning for identifying public health variables across social media.

[25] :SONG M. (2019). Health social network analytics: Analysis of chronic diseases with extracted entities and their relations. J Med Internet Res, 21(6), e12876.

[26] :GABARRON E., DORRONZORO E., RIVERA-ROMERO O. & WYNN R. (2019). Diabetes on twitter: a sentiment analysis. Journal of diabetes science and technology, 13(3), 439– 444.

[27] :MODAVE F., ZHAO Y., KRIEGER J., HE Z., GUO Y., HUO J., PROSPERI M. & BIAN J. (2019). Understanding perceptions and attitudes in breast cancer discussions on twitter.

[28] :http://www.thebeaconservices.com/sentiment_analysis.php

[29] :Clearing the Confusion: Artificial Intelligence vs Machine Learning vs Deep Learning by Csongor Barabasi | Jan 2, 2019 | Artificial Intelligence

[30] :Josh Patterson & Adam Gibson, 2017. *Deep Learning A Practitioner's Approach,* 1ère (Ed), O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, Mike Loukides & Tim McGovern, 532 p, (pp. 28).

[31] :Annina S., Mahima S, S. Venkatesan3, D.R. Ramesh Babu, An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE).

[32] :FAREK Lazhar, 2009, MEMOIRE Présentation en vue de l'obtention du diplôme de magister Identification d'opinions dans les journaux arabes, Faculté de Sciences de l'ingénieur Annaba

[33] :Andrew W. Trask, 2019. grokking Deep Learning.

[34] :Vincent Boucher, 2017. Mémoire de Master Machine learning en fnance.

[35] :Alex, D., 2017: Difference between Machine Learning and Deep Learning. https://artificialintelligencehow.com/2017/10/18/difference-machine-learning-deep-learning/ 20/4/2021

[36] :https://fr.mathworks.com/discovery/deep-learning.html 10/5/2021

[37] :https://www.securityinfowatch.com/videosurveillance/videoanalytics/article/21069937/deep-learning-to-the-rescue 13/5/2021

[38] : https://link.springer.com/article/. 8/5/2021

[39] :Yoav Goldberg, 2015. A Primer on Neural Network Models for Natural Language Processing.

[40] : https://deeplylearning.fr/cours-theoriques-deep-learning/fonction-dactivation/ 10/5/2021

[41] : http://staff.univ-batna2.dz/sites/default/files/melkami_kameleddine/files/chap2_2.pdf 9/5/2021

[42] :HUBEL D. H. & WIESEL T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1), 106–154.

[43] : FUKUSHIMA K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), 193–202.

[44] : LECUN Y., BOSER B. E., DENKER J. S., HENDERSON D., HOWARD R. E., HUBBARD W. E. & JACKEL L. D. (1990). Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems, p. 396–404.

[45] : MUHAMMAD ADNAN MUSHTAQ MULTILINGUAL, SENTIMENT ANALYSIS AS PRODUCT REPUTATION INSIGHT. University Lecturer Timo Aaltonen, Master of Science Thesis, 2017.

[46] : https://www.mlground.com/introduction-to-recurrent-neural-network-part-1/ 13 /5 /2021

[47] : https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/ 14 /15 /2021

[48] : BEGHDAD Abdelkrim - OUSERIR Amina Une approche Deep Learning pour l'analyse des Sentiments Sur Twitter 2017 2018

[49] : https://arabicprogrammer.com/article/4117317328/ 20 5 2021

[50] : https://www.lemagit.fr/definition/Traitement-du-langage-naturel-TLN 15 /5 /2021

[51] : https://medium.com/@remybonnaffe/new-technologies-and-the-law-the-impact-of-artificial-intelligence-on-the-practice-of-law-c456904688d1 15/ 5/ 2021

[52] : https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging / 14 / 5 /2021

[53] : https://monkeylearn.com/blog/named-entity-recognition/ 16/ 5 /2021

[54] : Silvia F., Eric S., Juan M., Torres M., 2007. *Énergie textuelle de mémoires associatives*.

[55] : Ronan C., Jason W., *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning.*

[56] : Alexis C., Holger S., Yann Le Cun, 2016. Very Deep Convolutional Networks for Natural Language Processing.

[57] : https://datascientest.com/introduction-au-nlp-natural-language-processing 15 5 2021

[58] : https://analyticks.wordpress.com/2016/08/14/leveraging-deep-learning-for-multilingual-sentiment-analysis/

[59] : https://morioh.com/p/57f7f3b0cb33 16 5 2021

[60] : T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, https://arxiv.org/abs/1301.3781.View at: Google Scholar

[61] : J. Pennington, R. Socher, and C. D. Manning, "Glove: global vectors for word representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, Doha, Qatar, October 2014.View at: Google Scholar

[62] : https://gist.github.com/phileas-condemine/2db55ae38e78c61728716dc3a7b91979 20 5 2021

[63] : https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30 25/5/2021

[64] : https://ichi.pro/fr/word2vec-glove-fasttext-et-word-embeddings-de-base-etape-par-etape-229010274898187 20 5 2021

[65] : ALAYBA A. M., PALADE V., ENGLAND M. & IQBAL R. (2017). Arabic language sentiment analysis on health services. In 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), p. 114–118: IEEE.

[66] : ABDELHADE N., SOLIMAN T. H. A. & IBRAHIM H. M. (2017). Detecting twitter users' opinions of Arabic comments during various time episodes via deep neural network. In International Conference on Advanced Intelligent Systems and Informatics, p. 232–246: Springer.

[67] : AL-AZANI S. & EL-ALFY E.-S. M. (2017). Hybrid deep learning for sentiment polarity determination of Arabic micro blogs. In International Conference on Neural Information Processing, p. 491–500: Springer.

[68] : ELNAGAR A., LULU L. & EINEA O. (2018b). An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis. Procedia computer science, 142, 182–189.

[69] : ALAYBA A. M., PALADE V., ENGLAND M. & IQBAL R. (2018b). Improving sentiment analysis in Arabic using word representation. In 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), p. 13–18: IEEE.

[70] : HEIKAL M., TORKI M. & EL-MAKKY N. (2018). Sentiment analysis of Arabic tweets using deep learning. Procedia Computer Science, 142, 114–122.

[71] : ALAYBA A. M., PALADE V., ENGLAND M. & IQBAL R. (2018a). A combined cnn and lstm model for Arabic sentiment analysis. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction, p. 179–191: Springer.

[72] : MOHAMMED A. & KORA R. (2019). Deep learning approaches for Arabic sentiment analysis. Social Network Analysis and Mining, 9(1), 52.

[73] : DAHOU A., ELAZIZ M. A., ZHOU J. & XIONG S. (2019). Arabic sentiment classification using convolutional neural network and differential evolution algorithm. Computational intelligence and neuroscience, 2019.

[74] : ElSahar, H., & El-Beltagy, S. R. (2015, April). Building large arabic multi-domain resources for sentiment analysis. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 23-34). Springer, Cham.