Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abbes Laghrour Khenchela
Faculté sciences et technologie
Département d'informatique

ce mémoire intitulé :

L'apprentissage profond pour l'analyse d'opinion des textes arabes

- Réalisé par :

  Dekkiche Abdellah
  Hassad Takieddine

- Dérigé par :

  Dr. Bakhouche Abdelali

- Devant les Jurys :

  Dr. Mahdaoui rafik
  Dr. Abdelhadi Adel

2019/2020

Sommaire

Table des figures

remerciement

Au terme de ce modeste travail, je tiens tout d'abord à remercier Allah, le tout puissant, de m'avoir accordé le courage, la patience, la volonté et surtout LA Santé pour réaliser et mener à bien mon travail.

J'adresse mes remerciements les plus sincères à mon directeur de recherche Mr Bakhouch abdelali pour sa direction, sa patience, son soutien, ses encouragements, ses remarques et ses conseils éclairants.

Je tiens à exprimer mes profonds remerciements à mes parents pour leurs conseils et leurs encouragements.

Je tiens à remercier spécialement tous mes professeurs qui m'ont enseigné durant les cinq ans au département de informatique.

Je présente mes remerciements les plus vifs pour tous ceux qui ont participé de près et de loin pour la réalisation de ce modeste travail.

Enfin, J'adresse mes remerciements aux membres du jury qui ont accepté de lire et d'évaluer ce travail de recherche.

Résumé

L'analyse du sentiment est extrêmement utile en veille des medias sociaux car elle permet d'obtenir une vue d'ensemble sur l'opinion du public au sujet de certains thèmes. Les utilisations de l'analyse du sentiment sont à la fois vastes et puissantes. La possibilité d'extraire des insights à partir des données du web social est une pratique qui est largement adoptée par les entreprises à travers le monde. La tâche d'identification d'opinions à partir des textes nécessite une analyse lexicale et syntaxique profonde, surtout dans notre cas, où la langue traitée est l'arabe, cette dernière, qui se caractérise par une morphologie complexe, présente l'un des grands challenges que nous devons faire face.

Abstract

Sentiment analysis is extremely useful in social media watch as it provides an overview of public opinion on certain topics. The uses of sentiment analysis are both broad and powerful. The ability to extract insights from social web data is a practice that is widely adopted by companies around the world. The task of identifying opinions from texts requires a deep lexical and syntactic analysis, especially in our case, where the language treated is Arabic, the latter, which is characterized by a complex morphology, presents one of the great challenges we face.

Introduction général

Contexte du travail

L'analyse des sentiments est une tâche de traitement automatique des langues et d'extraction d'information. Pour un texte donné, il faut identifier la polarité du texte comme étant soit positif, soit négatif. Indiquent plusieurs méthodes, plusieurs références de performance et ressources pour réaliser cette tâche. La polarité d'un sentiment peut être calculée selon plusieurs seuils et peut être vue comme plusieurs différentes classes. Dans notre projet, nous considérons les textes comme pouvant appartenir à seulement deux classes (classification binaire) : soit le texte est positif ou négatif.

Récemment, l'analyse de sentiments a reçu beaucoup d'attention non seulement de la part de la recherche scientifique mais aussi par les autres domaines . Cela peut être attribué aux récentes avancées dans les réseaux sociaux et à la rapidité du relais de l'information.

Les grandes masses de données réelles issues des réseaux sociaux sont largement utilisées pour l'analyse des sentiments. Analyser les messages récents issus des réseaux sociaux pourrait donner l'opinion générale des utilisateurs envers un sujet spécifique.

La plupart des recherches existantes sur l'analyse des sentiments se concentrent sur le texte Anglais. En dépit de son importance en tant que l'une des langues les plus utilisées dans le monde, seules un nombre limité de recherches sur l'analyse du sentiment du texte Arabe ont été réalisées. Les approches de l'analyse du sentiment Arabe proposées se concentrent principalement sur l'Arabe moderne standard parmi lequel peu d'études ont étudié le cas des dialectes Arabes (Arabe familier).

Problématique et objectifs

Notre travail concerne l'analyse automatique des énoncées d'opinion en arabe. En basant sur les méthodes de l'apprentissage automatique et sur le réseau lexical Wordnet pour déterminer les caractéristiques telles que la Force (intensité), Le Focus (prototypicalité) et la polarité de tels énoncés

Organisation de la thèse

Notre thèse est structurée en quatre chapitres et une conclusion générale.

— Tout d'abord, Le premier chapitre contient deux parties, la première partie donne la définition de l'IA ; la définition de l'apprentissage automatique ; la définition et Un tour d'horizon et quelque définition et les types de l'apprentissage profond. La deuxième partie donne l'historique et les type des réseaux neurones, aussi des avantage et des inconvénients de réseaux neurones.

— Les bas sur l'analyses des sentiments et leurs les déférents types.

— Le chapitre 3 est consacré pour présenter la conception de notre système

— L'annexe de notre mémoire

Chapitre 1

L'apprentissage profond et réseaux de neurones

1.1 L'apprentissage profond

1.1.1 L'intelligence artificielle

Depuis l'émergence de la robotique et de l'informatique, les chercheurs essaient d'injecter des notions d'intelligence humaine dans des machines. Étant conçue et fabriquée par l'homme, on qualifie cette forme d'intelligence comme "L'intelligence artificielle" ou IA.

L'IA est devenue un sujet en vogue dans les médias et magazines scientifiques en raison des nombreuses réalisations, dont beaucoup sont le fruit des progrès accomplis dans le domaine de l'apprentissage automatique. De grandes entreprises dont Google, Facebook, IBM, Microsoft mais aussi des constructeurs automobiles à l'instar de Toyota, Volvo et Renault, sont très actifs dans la recherche en IA et prévoient d'y investir davantage encore dans le futur. Plusieurs scientifiques spécialisés dans l'IA dirigent désormais les laboratoires de recherche de ces grandes entreprises et de nombreuses autres. La recherche en IA a permis de réaliser d'importants progrès dans la dernière décennie, et ce dans différents secteurs. Les avancées les plus connues sont celles réalisées dans l'apprentissage automatique, grâce notamment au développement d'architectures d'apprentissage profond, des réseaux de neurones convolutifs multicouche dont l'apprentissage s'opère à partir de gros volumes de données sur des architectures de calcul intensif. Parmi les réalisations de l'apprentissage automatique, il convient de citer la résolution de jeux Atari (Bricks, Space invaders, etc.) par Google DeepMind, utilisant les pixels images affichés à l'écran comme données d'entrée afin de décider quelle action adopter pour atteindre le plus haut score possible à la fin de la partie.[Bertrand Braunschweig ,2016]

1.1.2 L'apprentissage automatique

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle (IA), il consiste à doter l'ordinateur de capacités pour se programmer lui - même, au lieu de le programmer manuellement, ainsi il pourra résoudre de nouveaux problèmes à partir de données déjà fournies. En général, l'objectif de l'apprentissage automatique est de comprendre la structure des données et de les intégrer dans des modèles qui peuvent être compris et utilisés par tout le monde. Bien que l'apprentissage automatique soit un domaine de l'informatique, il diffère des approches informatiques traditionnelles, en effet dans cette dernière, les algorithmes sont des ensembles d'instructions explicitement programmées utilisées par les ordinateurs pour calculer ou résoudre des problèmes. Les algorithmes d'apprentissage automatique permettent

aux ordinateurs de s'entraîner sur les entrées de données et utilisent l'analyse statistique pour produire des valeurs qui se situent dans une plage spécifique. Pour cette raison, l'apprentissage automatique facilite l'utilisation des ordinateurs dans la construction de modèles à partir de données d'échantillonnage afin d'automatiser les processus de prise de décision en fonction des données saisies.

Pourquoi l'apprentissage automatique ?

Machine learning(ML) utilise des ordinateurs pour simuler l'apprentissage humain et permet aux ordinateurs d'identifier et d'acquérir des connaissances du monde réel, et d'améliorer les performances de certaines tâches en fonction de ces nouvelles connaissances [Portugal, I., Alencar, P., Cowan, D. (2017)]

Pour mieux comprendre l'utilisation de l'apprentissage automatique, nous allons sites quelques domaines de ce dernier :

- Moteurs de recommandation en ligne comme des suggestions d'amis sur Facebook.
- Auto-conduite de Google
- Netflix présentant les films et émissions que vous aimerez et «plus d'éléments à prendre en compte» et «obtenez-vous un petit quelque chose» sur Amazon sont autant d'exemples d'apprentissage automatique appliqué.
- Détection de cyber-fraude

Tous ces exemples font écho au rôle essentiel que l'apprentissage automatique a commencé à prendre dans le monde riche en données d'aujourd'hui. Les machines peuvent aider à filtrer des informations utiles qui aident à des avancées majeures, et nous voyons déjà comment cette technologie est mise en œuvre dans une grande variété d'industries. Le flux de processus représenté ici représente le fonctionnement de l'apprentissage automatique [Priyadharshini. (March 8, 2018)]

1.1.3 L'apprentissage profond

Après avoir la définition de IA et l'apprentissage automatique, nous allons maintenant intéresser plus particulièrement à l'apprentissage profond.

L'apprentissage profond est un sous-ensemble de l'apprentissage automatique, utilise un niveau hiérarchique de réseaux neuronaux artificiels pour réaliser le processus d'apprentissage automatique.

Un tour d'horizon sur l'apprentissage profond

Dans les années 1950, le mathématicien britannique Alan Turing imagine une machine capable d'apprendre, une «Learning Machine». Au cours des décennies suivantes, différentes techniques de Machine Learning ont été développées pour créer des algorithmes capables d'apprendre et de s'améliorer de manière autonome.Parmi ces techniques, on compte les réseaux de neurones artificiels. C'est sur ces algorithmes que reposent l'apprentissage profond, mais aussi des technologies comme la reconnaissance d'images ou la vision robotique[Bastien L,2018]

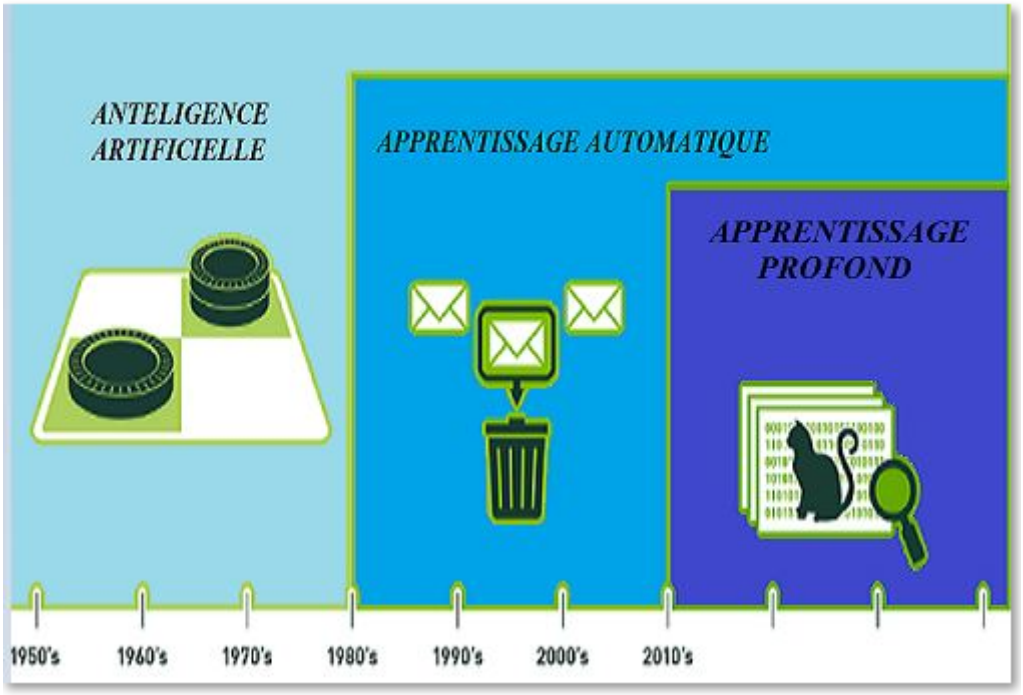


FIGURE 1.1 – Evolution d'apprentissage profond [Bastien L,2018]

Quelque définition del'apprentissage profond

— L'apprentissage profond (deeplearning) est un terme abrégé pour "apprentissage dans les réseaux de neurones profonds". Il s'agit des méthodes d'apprentissage automatique utilisant les réseaux de neurones profonds ; c'est donc un sous-domaine de l'apprentissage automatique (et un sous-sous-domaine de l'IA en général).[Dorianne W ,2017]

— L'apprentissage profond est un ensemble d'algorithmes de machine learning cherchant à modéliser des abstractions de haut niveau au sein des données en utilisant des architectures de modèles composés de multiples transformations non linéaires.[Rémi S, 2014]

L'apprentissage des réseaux de neurones

On appelle apprentissage des réseaux de neurones la procédure qui consiste à estimer les paramètres (poids, biais,...) des neurones du réseau, afin que celui-ci remplisse au mieux la tâche qui lui est affectée. il existe plusieurs paradigmes d'apprentissage.[Guesbaya Tahar,2012].

Apprentissage supervisé

Un apprentissage est dit supervisé lorsque l'on force le réseau à converger vers un état final précis, en même temps qu'on lui présente un motif. Ce genre d'apprentissage est réalisé à l'aide d'une base d'apprentissage, constituée de plusieurs exemples de type entrées-sorties (les entrées du réseau et les sorties désirées ou encore les solutions souhaitées pour l'ensemble des sorties du réseau). La procédure usuelle dans le cadre de la prévision est l'apprentissage supervisé (ou à partir d'exemples) qui consiste à associer une réponse spécifique désirée à chaque signal d'entrée. La modification des poids s'effectue progressivement jusqu'à ce que l'erreur (ou l'écart) entre les sorties du réseau (ou résultats calculés) et les résultats désirés soient minimisés. Cet apprentissage n'est possible que si un large jeu de données est disponible et si les solutions sont connues pour les exemples de la base d'apprentissage. [Marc.P, 2004].

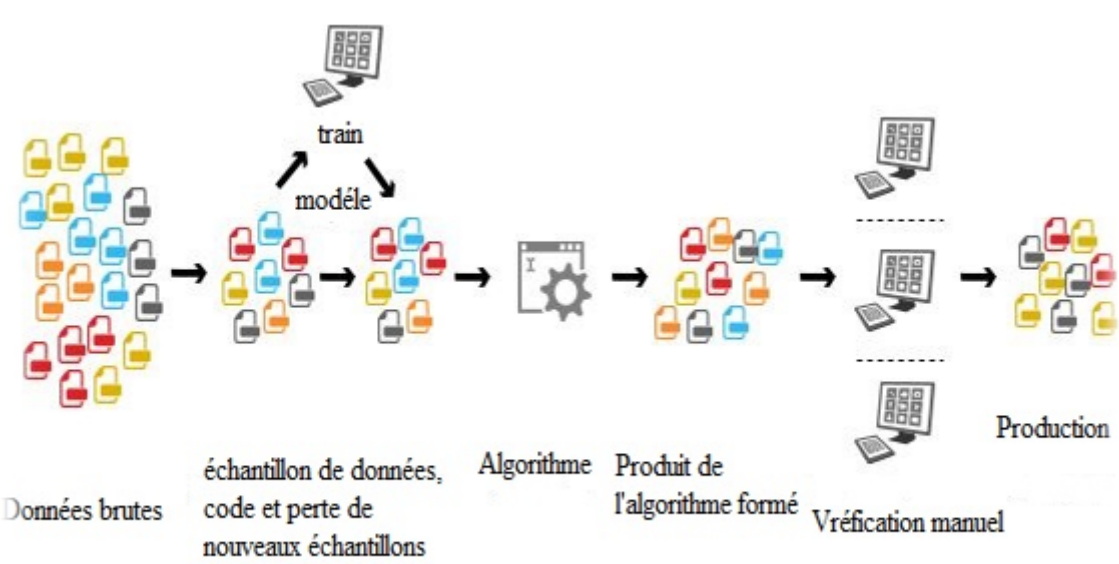


FIGURE 1.2 – Expliquele déroulement d'apprentissage supervisé [Priyadharshini. March 8, 2018]

Apprentissage non supervisé (autoorganisationnel)

L'apprentissage non supervisé consiste à ajuster les poids à partir d'un seul ensemble d'apprentissage formé uniquement de données. Aucun résultat désiré n'est fourni au réseau. Qu'est-ce que le réseau apprend exactement dans ce cas ? L'apprentissage consiste à détecter les similarités et les différences dans l'ensemble d'apprentissage. Les poids et les sorties du réseau convergent, en théorie, vers les représentations qui capturent les régularités statistiques des données. Ce type d'apprentissage est également dit compétitif et (ou) coopératif. L'avantage de ce type d'apprentissage réside dans sa grande capacité d'adaptation reconnue comme une Auto organisation, «self-organizing» [Kohonen, 1987]. L'apprentissage non supervisé est surtout utilisé pour le traitement du signal et l'analyse factorielle. [Jean-François.J , 1994].

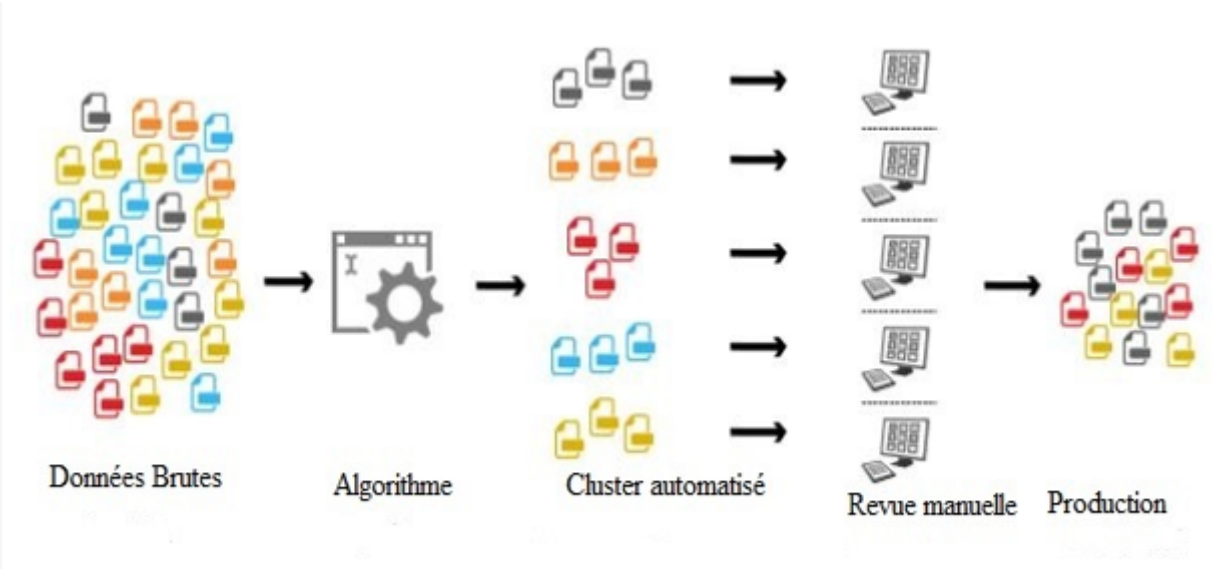


FIGURE 1.3 – Expliquele déroulement d'apprentissage supervisé [Priyadharshini. March 8, 2018]

Apprentissage hybride

Dans ce type on utilise les deux apprentissages supervisé et non-supervisé.

Apprentissage par renforcement

L'apprentissage renforcé est une technique similaire à l'apprentissage supervisé à la différence qu'au lieu de fournir des résultats désirés au réseau, on lui accorde plutôt un grade (ou score) qui est une mesure du degré de performance du réseau après quelques itérations. Les algorithmes utilisant la procédure d'apprentissage renforcé sont surtout utilisés dans le domaine des systèmes de contrôle [Marc.P, 2004].

Apprentissage compétitif

L'apprentissage compétitif, consiste à faire compétitionner les neurones d'un réseau pour déterminer lequel sera actif à un instant donné. Contrairement aux autres types d'apprentissage, où généralement tous les neurones peuvent apprendre simultanément et de la même manière, l'apprentissage compétitif produit un « vainqueur »ainsi que parfois, un ensembles de neurones « voisins » vainqueurs, et seuls se vainqueur et son voisinage bénéficient d'une adaptation de leurs poids, alors que le neurone qui ne gagne pas la compétition ne modifiera aucunement ses poids. Ainsi, les neurones individuels peuvent apprendre à se spécialiser sur des sous-ensembles de données pour devenir des détecteurs de caractéristiques [Dreyfus.G et al, 2002].

1.1.4 Les algorithmes d'apprentissage

L'algorithme d'apprentissage de Hebb

Modifie de façon itérative (petit à petit) les poids pour adapter la réponse obtenue à la réponse désirée. Il s'agit en fait de modifier les poids lorsqu'il y a erreur seulement .

- Initialisation des poids et du seuil S à des valeurs (petites) choisies au hasard
- Présentation d'une entrée $E_1 = (e_1, ...e_n)$ de la base d'apprentissage.
- Calcul de la sortie obtenue x pour cette entrée :
  a=$\Sigma(w_i.e_i$)-S(La valeur de seuil est introduite ici dans le calcul de la somme pondérée)
  x = signe (a) (si a ¿ 0 alors x = +1 sinon a ¡= 0 alors x = -1)
- Si la sortie x est différente de la sortie désirée d_1 pour cet exemple d'entrée E_1 alors modification des poids (µ est une constante positive, qui spécifie le pas de modification des poids) :
  w_{ij}(t+1) = w_{ij}(t) + µ. (x_i. x_j)
- Tant que tous les exemples de la base d'apprentissage ne sont pas traités correctement (Modification des poids), retour à l'étape 2.

L'algorithme d'apprentissage du perceptron

L'algorithme d'apprentissage du perceptron est semblable à celui utilisé pour la loi de Hebb. Les différences se situent au niveau de la modification des poids [Portugal, I., Alencar, P., Cowan, D. (2017)].

- Initialisation des poids et du seuil S à des valeurs (petites) choisies au hasard.
- Présentation d'une entrée $E_1 = (e_1, ...e_n)$ de la base d'apprentissage.

— Calcul de la sortie obtenue x pour cette entrée :

$$\text{a}=\Sigma(w_i.e_i)\text{-S}$$

x = signe (a) (Si a ¿ 0 alors x = +1 sinon a ¡= 0 alors x = -1)

— Si la sortie x du Perceptron est différente de la sortie désirée dl pour cet exemple d'entrée E_1 alors modification des poids (µ le pas de modification) : w_i(t+1) = w_i (t) + µ. ((d_1 - x).e_i)
Rappel : d_1 = +1 si E est de la classe 1, d_1 = -1 si E est de la classe 2 et (d_1 - x) est une estimation de l'erreur.

— Tant que tous les exemples de la base d'apprentissage ne sont pas traités correctement (modification des poids), retour à l'étape 2.

1.2 Réseaux de neurones

1.2.1 Introduction

Les réseaux de neurones sont constitués d'un ensemble de neurones artificiels ou nœuds qui sont analogues aux neurones biologiques. lls sont issus d'une tentative de conception d'un modèle mathématique très simplifié du cerveau humain en se basant sur notre façon d'apprendre et de corriger nos erreurs. Ce chapitre présente une introduction à la théorie de réseaux de neurone.

1.2.2 Historique sur les réseaux de neurones

Les recherches dans le domaine du connexionnisme ont démarré avec la présentation en 1943 par W. MC Culloch et W. Pitts d'un modèle simplifié de neurone biologique communément appelé neurone formel. Ils montrèrent également théoriquement que des réseaux de neurones formels simples peuvent réaliser des fonctions logiques, arithmétiques et symboliques complexes.

En 1949, D. Hebb initie, dans son ouvrage "The Organization of Behavior", la notion d'apprentissage. Deux neurones entrant en activité simultanément vont être associés (c'est-à-dire que leurs contacts synaptiques vont être renforcés). On parle de loi de Hebb et d'associationnisme.

En 1958, F. Rosenblatt développe le modèle du Perceptron. C'est un réseau

de neurones inspiré du système visuel. Il possède deux couches de neurones : une couche de perception (sert à recueillir les entrées) et une couche de décision. C'est le premier modèle pour lequel un processus d'apprentissage a pu être défini. S'inspirant du perceptron, Widrow et Hoff, développent, dans la même période, le modèle de l'Adaline (Adaptive Linear Element). Ce dernier sera, par la suite, le modèle de base des réseaux de neurones multi-couches.

En 1969, Les recherches sur les réseaux de neurones ont été pratiquement abandonnées lorsque M. Minsky et S. Papert ont publié leur livre « Perceptrons » (1969) et démontré les limites théoriques du perceptron, en particulier, l'impossibilité de traiter les problèmes non linéaires par ce modèle.

En 1982, Hopfield développe un modèle qui utilise des réseaux totalement connectés basés sur la règle de Hebb pour définir les notions d'attracteurs et de mémoire associative.

En 1984 c'est la découverte des cartes de Kohonen avec un algorithme non supervisé basé sur l'auto-organisation et suivi une année plus tard par la machine de Boltzman (1985). Une révolution survient alors dans le domaine des réseaux de neurones artificiels : une nouvelle génération de réseaux de neurones, capables de traiter avec succès des phénomènes non- linéaires : le perceptron multicouche ne possède pas les défauts mis en évidence par Minsky.

Proposé pour la première fois par Werbos, le Perceptron Multi-Couche apparaît en 1986 introduit par Rumelhart, et, simultanément, sous une appellation voisine, chez Le Cun (1985). Ces systèmes reposent sur la rétropropagation du gradient de l'erreur dans des systèmes à plusieurs couches, chacune de type Adaline de Bernard Widrow, proche du Perceptron de Rumelhart. [Kadous Djamila, 2012][Dreyfus D.et al ,2004].

1.2.3 Les réseaux de neurons

Les réseaux de neurones artificiels sont des combinaisons de fonctions élémentaires. appelées neurones formels, ou simplement neurones associés en couches et fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau.[MERZOUKA.N,2009] Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Ils sont constitués d'un nombre fini de neurones qui sont arrangés sous forme de couches. Les neurones de deux couches adjacentes sont interconnectés par des poids. L'information dans le réseau se propage d'une couche à l'autre, on dit qu'ils

sont de type « feed-forward ». Nous distinguons trois types de couches :

- **Couche d'entrée :** les neurones de cette couche reçoivent les valeurs d'entrée du réseau et les transmettent aux neurones cachés. Chaque neurone reçoit une valeur, il ne fait pas donc de sommation.
- **Couches cachées :** chaque neurone de cette couche reçoit l'information de plusieurs couches précédentes, effectue la sommation pondérée par les poids, puis la transforme selon sa fonction d'activation qui est en général une fonction sigmoïde. Par la suite, il envoie cette réponse aux neurones de la couche suivante
- **Couche de sortie :** elle joue le même rôle que les couches cachées, la seule différence entre ces deux types de couches est que la sortie des neurones de la couche de sortie n'est liée à aucun autre neurone.

Les réseaux de neurones artificiels possèdent une propriété fondamentale qui justifient l'intérêt croissant qui leur est accordé et que sont capable d'intervenir dans des domaines très divers, et qui les distingue des techniques classiques de traitement des données. Les réseaux de neurones sont des approximateurs universels : Cette propriété peut être énoncée comme suit : Toute fonction bornée suffisamment régulière peut être approchée uniformément, avec bonne précision, dans un domaine fini de l'espace de ses variables, par un réseau de neurones qui comporte une couche de neurones cachée en nombre fini, possédant tous la même fonction d'activation et un neurone de sortie linéaire [J.GHOUILI 2005.][B.GOSSELIN 1996.] .
Parcimonie : Lors de la modélisation d'un processus à partir de ses données, on cherche toujours à obtenir les résultats les plus satisfaisants possibles avec un nombre minimum de paramètres. On dit que l'on cherche l'approximation la plus parcimonieuse. Pour obtenir un modèle non linéaire de précision donnée, un réseau de neurone a besoin de moins de paramètres ajustables que les méthodes de régression classiques (par exemple la régression polynomiale). Or le nombre de données nécessaires pour ajuster le modèle est directement lié au nombre de ses paramètres [G. DREYFUS et al,2002],[L.BAGHLI, 1999].

1.2.4 Avantages et Inconvénients de réseaux de neurons :

Avantages

Les principales qualités des réseaux de neurones sont leur capacité d'adaptabilité et d'auto- organisation et la possibilité de résoudre des problèmes non-linéaires avec une bonne approximation [Anand et al, 1992][Watrous.R.L., 1987]. Ils ont une bonne

immunité aux bruits et se prêtent bien à une implantation parallèle. La rapidité d'exécution est une qualité importante et elle justifie souvent à elle seule le choix d'implanter un réseau de neurones. Ces qualités ont permis de réaliser avec succès, plusieurs applications : classification, filtrage, compression de données, contrôleur, etc.[HICHAM.C,2002].

Inconvénients

La difficulté d'interpréter le comportement d'un réseau de neurones est un inconvénient pour la mise au point d'une application. Il est souvent impossible d'utiliser les résultats obtenus pour améliorer ce comportement. Il est également hasardeux de généraliser à partir d'expériences antérieures et de conclure ou de créer des règles sur le fonctionnement et le comportement des réseaux de neurones. Plusieurs paramètres doivent être ajustés et aucune méthode ne permet de choisir des valeurs optimales. Beaucoup d'heuristiques sont utilisées, mais elles se contredisent parfois et elles ne permettent pas toujours de trouver des valeurs optimales.[HICHAM.C,2002].

1.2.5 Conclusion

Dans ce chapitre nous avons représentées deux parties : la première partie on a parlé de l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond ; dans la deuxième partie on a parlé sur les réseaux de neurones.

Chapitre 2

Analyse des Sentiments

2.1 Introduction

La fouille de données d'opinions est un domaine de recherche en plein essor. Elle devient essentielle, par exemple pour le développement de tâches de veille (technologique, marketing, concurrentielle, sociétale) qui peuvent se révéler cruciales pour les entreprises et trouve de très nombreux domaines d'applications. Nous pouvons citer, par exemple, les clients qui souhaitent connaître comment évaluer un produit avant de l'acheter, l'image que les clients peuvent se faire d'une entreprise, la détection de rumeurs sur le web.
Cependant, les approches traditionnelles de fouilles de données ne sont plus adaptées à un contexte dans lequel il faut appréhender non seulement de gros volumes de données mais s'intéresser à la qualité des données : comment déterminer des avis négatifs ou positifs dans des documents aussi divers que des blogs ou des journaux ? Comment valider/évaluer les résultats obtenus ? Quel type de données à utiliser ? Une approche pluridisciplinaire regroupant différentes communautés (fouille de données, aide à la décision, modélisation des connaissances, TAL, Linguistique, etc.) parait aujourd'hui essentielle au développement rigoureux de cette thématique. Notons, que l'un des objectifs de l'opinion-mining est la classification de textes en fonction des jugements favorables ou défavorables qu'ils expriment.
Nous présentons dans ce chapitre la définition de l'analyse des sentiments (opinion-mining), les Types d'opinion, les Niveaux d'analyse des sentiments, les Tâches de l'analyse des sentiments, les domaines d'application, ainsi que les sources de données.

2.2 Définitions de l'analyse des sentiments

Dans la littérature, sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, appraisal ex- traction, sont des termes utilises pour designer des technologies d'analyse automatique des discours, écrits ou parles, a fin de extraire des informations subjectives comme des jugements, des évaluations ou des émotions. L'origine de la discipline l'analyse des sentiments se réfère aux des sciences de la psychologie, la sociologie et l'anthropologie [MEENA RAMBOCAS AND JO,2013]
Le terme Analyse Sentimentale se réfère à l'extraction automatique de texte évaluative, qui aide à produire des résultats prédictifs. Le terme analyse de sentiment est apparu en Nasukawa et Yi en 2003[NASUKAWA,2003] , et le terme extraction de l'opinion terme est apparu dans Dave, Laurent et Pennock en 2003 [KUSHAL DAVE, STEVE LAWRENCE AND DAVID M,2003] Cependant, la recherche sur des sen-

timents et des opinions est apparue plus tôt dans [SANJIV R. DAS ET MIKE Y ,2001] [BO PANG, LILLIAN LEE, SHIVAKUMAR VAITHYANATHAN,2002] [R. M,2001] [PETER D,2002] [JANYCE WIEBE ,2000] .
[$BingLiu, 2015$] a présenté une définition de l'analyse des sentiments comportant les domaines d'application ainsi que sa relation avec le TALN : l'analyse des sentiments est le domaine de l'étude qui analyse les opinions, les sentiments, les évaluations, les attitudes et les émotions des gens vers des entités telles que des produits, des services, des organisations, des particuliers, des problèmes, des événements, des sujets, et leurs attributs. Il représente un grand espace de recherche. l'analyse des sentiments est un domaine de recherche extrêmement actif en traitement automatique des langues. Pour mettre en valeur l'intérêt de l'echange d'opinions dans l'analyse des sentiments [BO PANG AND LILLIAN LEE,2008] considère que l'opinion des autres a toujours été une pièce d'information très précieuse au moment de se faire une opinion ou de prendre une décision. En effet, avant l'apparition du Web et l'Internet, les gens avaient intérêt à connaitre les opinions de leurs amis ou de leur famille. Il leur était demande de faire savoir quel parti politique recevrait leur voix lors des prochaines élections. Grace a l'essor considérable qu'ont connu le Web et l'Internet à partir des années quatre-vingt-dix, il est devenu possible pour tous de consulter l'opinion d'un vaste groupe de personnes a travers le Web.
vDonc l'echange d'opinion est la phase principale qui permet d'éffectuer une analyse de sentiment sur un sujet donne. Selon H Tang et S Tan X la plupart des recherches existantes se sont portées sur la fouille et l'extraction de faits, par exemple, la recherche d'information, la recherche sur le Web et beaucoup d'autres. On assiste, ces dernières années, à une prise de conscience de l'importance de l'opinion sur le web, ce qui explique les nombreux et récents travaux dans ce domaine [HUIFENG TANG, SONGBO TAN AND XUEQI CHENG, 2009]. Ils montrent l'importance de l'analyse de sentiment dans le temps actuel.

2.3 Types d'opinion

On peut distinguer deux types d'opinions la première s'appelle opinion régulière [BING LIU, 2007] L'autre type est appel´ e opinion comparative [NITIN JINDAL AND BING LIU,2006]. En fait, nous pouvons également classer les opinions en fonction de la fa¸ con dont ils sont exprimé´ s dans le texte, l'opinion explicite et l'opinion implicite.

2.3.1 Opinion régulière et opinion comparative

Opinion régulière

Une opinion régulière est souvent simplement considérée comme une opinion dans la littérature et il y a deux sous-types principaux [BING LIU,2007]

— **Opinion directe :** Une opinion directe fait référence à une opinion exprimée directement sur une entité ou un aspect de l'entité, par exemple, "La résolution de cet écran est excellente."

— **Opinion indirecte :** Une opinion indirecte est une opinion exprimée indirectement sur une entité ou aspect d'une entité en fonction de ses effets sur d'autres entités. Ce sous-type se produit souvent dans le domaine médical. Par exemple, la phrase "Après l'injection du médicament, mes articulations senties pire" décrit un effet indésirable du médicament sur" mes articulations ", ce qui donne indirectement une opinion négative ou un sentiment au médicament. Dans le cas, l'entité est le médicament et l'aspect est l'effet sur les articulations. Une grande partie de la recherche actuelle se concentre sur les opinions directes. Ils sont plus simples a manipuler. Les opinions indirects sont souvent plus difficiles a traiter. Par exemple, dans le domaine du médicament, il faut savoir si un état souhaitable et indésirable est avant ou après l'utilisation du médicament. Par exemple, la phrase "Puisque mes articulations étaient douloureuses, mon médecin m'a mis sur ce médicament" n'exprime pas un sentiment ou une opinion sur le médicament parce que "articulations douloureuses" (ce qui est négatif) est arrivé avant d'utiliser le médicament.

Opinion comparative

Un avis comparatif exprime une relation de similitudes ou de différences entre deux ou plusieurs entités et/ou une préférence du détenteur d'opinion sur la base de certains aspects partages des entités [HADY ELSAHAR AND SAMHAA R EL-BELTAGY, 2015] Par exemple, les phrases Dell est meilleur que HP et Dell est le meilleur expriment deux opinions comparatives. Une opinion comparative est habituellement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe, mais pas toujours (par exemple : l'utilisation de verbe préférer).

2.3.2 Opinion explicite et opinion implicite

Opinion explicite

Une opinion explicite est une déclaration subjective qui donne une opinion régulière ou comparative, par exemple : "Le couscous a bon goût" et "Facebook est mieux que Twitter."

Opinion implicite

Une opinion implicite est une déclaration objective qui implique une opinion régulière ou comparative. Une telle déclaration objective exprime habituellement un fait souhaitable ou indésirable, par exemple : "La durée de vie de la batterie de l'ordinateur portable Toshiba est plus longue que celle de l'ordinateur portable HP."

2.4 Niveaux d'analyse des sentiments

En général, l'analyse des sentiments a été étudiée principalement à trois niveaux

2.4.1 Niveau document

La tâche à ce niveau est de classer si un document d'opinion entier exprime un sentiment positif ou négatif [BO PANG, LILLIAN LEE, AND SHIVAKUMAR VAITHYANATHAN,2002] [PETER D TURNEY,2002] Par exemple, à la suite d'une revue de produit, le système détermine si l'avis exprime une opinion globale positive ou négative sur le produit. Cette tâche est communément appelée classification de sentiment au niveau du document. Ce niveau d'analyse suppose que chaque document exprime des opinions sur une seule entité (par exemple, un seul produit). Ainsi, il ne s'applique pas aux documents qui évaluent ou comparent plusieurs entités.

2.4.2 Niveau phrase

a tâche à ce niveau va aux phrases et détermine si chaque phrase exprime une opinion positive, négative ou neutre. Neutre ne signifie généralement aucune opinion. Ce niveau d'analyse est étroitement lié à la classification de la subjectivité [JANYCE M WIEBE, REBECCA F BRUCE, AND THOMAS P O'HARA, 1999] qui distingue les phrases (appelées phrases objectives) qui expriment des informations factuelles, de phrases (appelées phrases subjectives) exprimant des opinions et des opinions subjectives. Cependant, nous devons noter que la subjectivité n'est pas équivalente au sentiment car de nombreuses phrases objectives peuvent impliquer des opinions,

par exemple : Nous avons acheté un nouvel ordinateur portable le mois dernier et l'écran est tombe en panne." Les chercheurs ont également analysé les clauses [THERESA WILSON, JANYCE WIEBE, AND REBECCA HWA,2004] mais le niveau de la clause n'est toujours pas suffisant, par exemple, "L'Algérie se porte très bien dans cette crise économique".

2.4.3 Niveau aspect

Les analyses au niveau du document et de la phrase ne permettent pas de découvrir exactement ce que les gens aimaient et n'aimaient pas. Le niveau d'aspect effectue une analyse plus fine. Le niveau d'aspect était auparavant appelé niveau caractéristique (extraction d'opinion basée sur les caractéristiques et résume)[MINQING HU AND BING LIU, 2004] Au lieu de regarder des constructions de langage (documents, paragraphes, phrases, clauses ou expressions), le niveau d'aspect regarde directement l'opinion elle-même. Il est bas´e sur l'idée qu'une opinion consiste en un sentiment (positif ou négatif) et une cible (d'opinion).

Une opinion, sans que sa cible soit identifiée, est d'une utilité limitée. Réaliser l'importance des cibles d'opinion nous aide également à mieux comprendre le problème de l'analyse des sentiments. Par exemple, bien que la phrase "Bien que le service n'est pas génial, j'aime toujours Cet hôtel. A clairement un ton positif, on ne peut pas dire que cette phrase soit entièrement positive. En fait, la phrase est positive sur l'hôtel (souligné), mais négative sur son service (non souligne). Dans de nombreuses applications, les cibles d'opinion sont décrites par les entités et / ou leurs différents aspects. Ainsi, le but de ce niveau d'analyse est de découvrir les sentiments sur les entités et / ou leurs aspects. Par exemple, la phrase "La performance de Dell est bonne, mais sa durée de vie de la batterie est courte." Evalue deux ´ aspects : performance et autonomie de la batterie, de Dell (entité).

Le sentiment sur la performance de Dell est positif, mais le sentiment sur sa durée de vie est négatif. La performance et la durée de vie de la batterie de Dell sont les cibles d'opinion. Sur la base de ce niveau d'analyse, un résumé structure des opinions sur les entités et leurs aspects peut être produit, ce qui transforme le texte non structuré en données structurées et peut être utilisé pour toutes sortes d'analyses qualitatives et quantitatives. Les classifications au niveau du document et de la phrase sont déjà très difficiles.

Pour rendre les choses encore plus intéressantes et stimulantes, il existe deux types d'opinions, à savoir des opinions régulières et des opinions comparatives [NITIN JINDAL AND BING LIU,2006]

Une opinion régulière exprime un sentiment seulement sur une entité particulière

ou sur un aspect de l'entité, par exemple, "Couscous a un goût très bon", ce qui exprime un sentiment positif sur le goût de Couscous. Un avis comparatif compare plusieurs entités en fonction de certains de leurs aspects communs, par exemple, Couscous a meilleur goût que Chakhchoukha, ce qui compare Couscous et Chakhchoukha en fonction de leurs goûts (un aspect) et exprime une préférence pour Couscous.

2.5 Tâches de l'analyse des sentiments

Nous allons, dans cette section, aborder les différentes tâches qui composent un système d'analyse de sentiments. Ce plan se réfère principalement au modèle de [Liu, 2012] et fournit une définition de chaque tache.

2.5.1 Résumé de l'opinion

Les applications de l'analyse de sentiments requièrent l'étude des opinions de beaucoup de personnes car un seul avis ne suffit pas, de ce fait, une certaine forme de résumé s'impose [BING LIU, 2013] La récapitulation d'opinions consiste finalement à générer un résumé concis et digeste d'un grand nombre d'opinions. [MARTI A HEARST, 1992] sont les premiers à proposer des résumés bas´es sur les aspects à partir de critiques de clients vis-à-vis des produits vendus en ligne. Ils résument leur travail en trois étapes qui sont :

— identification des aspects du produit que les clients ont mentionnés dans leurs opinions
— identification des phrases qui contiennent une opinion positive ou négative pour chaque aspect.
— production d'un résumé en utilisant les informations découvertes.

2.5.2 Détection des spams

Aujourd'hui, à travers les réseaux sociaux, les blogs et les micro-blogs, il est très facile pour les gens d'exprimer leurs opinions d'une façon anonyme. Malgré ses avantages, l'anonymat a produit de nouvelles difficultés pour l'analyse de l'opinion. Il permet aux gens avec des intentions malveillantes fausser les résultats des systèmes en postant de faux avis afin de promouvoir ou de discréditer des produits cibles, des services, des organisations ou des individus sans divulguer leurs véritables intentions. La tâche de la détection des spams vise essentiellement à repérer ces gens (les spammeurs d'opinion) afin d'assurer la fiabilité des sources. Contrairement 'a

l'extraction d'opinions, la détection de spams n'est pas seulement un problème de traitement du langage naturel car elle est considérée aussi comme étant un problème d'extraction de données

2.6 Domaines d'applications de l'analyse des sentiment

L'importance de la détection d'opinion est présente dans plusieurs domaines ainsi plusieurs applications ont vu le jour dans ce contexte. Nous citons brièvement quelques applications ci-dessous :

2.6.1 La politiqueLa politique

Les acteurs politiques ont suivi la tendance de détection d'opinion, tel qu'avant de promulguer une nouvelle loi, les politiciens essayent de récolter l'avis des internautes sur cette loi. Il est intéressant de connaitre aussi l'avis des internautes sur un homme politique pour une élection présidentielle [FAIZA BELBACHIR, 2010]

2.6.2 Les entreprises

À travers l'analyse des sentiments, les entreprises peuvent connaitre l'opinion des clients sur leurs produits ou leur service. Dans une perspective d'améliorer leurs produits et d'augmenter leurs chiffres d'affaires [ALEXANDER PAK AND PATRICK PAROUBEK, 2010]. Dans le domaine du Product review mining, notamment 'a partir des sites de consultation. Les consommateurs viennent y échanger des avis et trouver des conseils pour leurs décisions d'achat (produits technologiques, voitures, voyage et hôtels, ... etc) [ALEXANDER PAK AND PATRICK PAROUBEK , 2010]

Le marketing a rapidement compris l'intérêt de l'analyse de sentiment. Des agences vendent aux entreprises la traque des moindres mots sur leur image, sur leurs produits [DOMINIQUE BOULLIER ET AUDREY LOHARD,2012]

2.6.3 Les clients

L'analyse des sentiments fait partie aussi de vie des internautes. Les sondages dans ce domaine montrent que la majorité des clients avant qu'ils achètent un produit, ils font des recherches d'avis sur se produit ou un service donne et même ils sont prêts 'a payer plus cher un produit dont l'avis est plus favorable qu'un autre [BING LIU, 2012].

2.6.4 Gestion de réputation de la marque (GRM)

La gestion de la réputation de la marque en Anglais Brand Réputation Management (BRM) se préoccupe par la gestion de le réputation de la marque sur le marché. Les opinions des clients ou d'autres parties peuvent endommager ou améliorer une telle réputation. la GRM est s'intéresse au produit et à l'entreprise plutôt qu'au client. Actuellement, un à plusieurs (one-to-many) conversations ont lieu en ligne a un taux élève. Cela crée des opportunités pour les organisations à gérer et à renforcer la réputation de leur marque. Maintenant, la perception de marque est déterminée non seulement par la publicité et les relations publiques. Les marques sont devenues une somme des conversations à leur sujet. L'analyse des sentiments aide à déterminer comment la marque, produit ou service de l'entreprise est perçue par la communauté en ligne [VIVEK KUMAR SINGH AND DEBANJAN MAHATA,2010]

2.7 Sources des Données

Les opinions des utilisateurs présentent le critère principal pour l'amélioration de la qualité des services fournis et la mise en valeur des produits livres. Ces opinions se présentent sous déférentes sources de données, à savoir, sites d'avis, blog et micro-blog.

2.7.1 Sites d'avis

Les opinions ont le rôle de décideur pour tout utilisateur durant la phase d'achat. Les avis générés par les utilisateurs sur les produits et les services sont largement disponibles sur internet. La classification de sentiment utilise les données de l'examinateur collectées à partir des sites Web tels que :

— www.gsmarena.com (revues de téléphone portable).
— www.amazon.com (revues des produits).
— www.CNETdownload.com (revues des produits).

Ces sites accueillent des millions d'avis sur les produits par les consommateurs [ARTI BUCHE, DR. M. B. CHANDAK AND AKSHAY ZADGAONKAR, 2013]. [G.VINODHINI AND RM.CHANDRASEKARAN,2012]

2.7.2 Blogs

Un blog est ou les personnes peuvent écrire les différents sujets dans un but de partage avec d'autres personnes sur le même site. La simplicité de la création des postes blogs ainsi que leur forme libre à rendue le blogging un événement accessible.

La blogosphère nom associe à l'univers de tous les blogs . Sur la blogosphère, nous trouvons un nombre important de messages relatif à une panoplie des sujets d'intérêt. Les blogs sont utilises sources d'opinions dans la plupart des études relatives à l'analyse des sentiments [ARTI BUCHE, DR. M. B. CHANDAK AND AKSHAY ZADGAONKAR , 2013] [VIVEK KUMAR SINGH AND DEBANJAN MAHATA, 2010]

2.7.3 Micro-blogs

Les micro-blogs sont parmi les outils de communication très populaires des utilisateurs d'internet. Chaque jour, des millions de messages apparaissent dans des sites Web populaires pour les microbloging tels que : Twitter, Tumblr, Facebook . Parfois les messages Twitter expriment des opinions qui sont utilisées comme source de données pour classifier le sentiment [] ALEXANDER PAK AND PATRICK PAROUBEK, 2010]

2.7.4 Twitter

En Mars 2006, Twitter a été crée par le développeur Jack Dorsey comme un outil pour rester en contact avec les amis, Twitter est un service sur le Web qui permet aux utilisateurs d'envoyer et de lire un message court [MATTHEW ERIC GLASSMAN, JACOB R. STRAUS AND COLLEEN J. SHOGAN, 2009]

2.8 Twitter et tweet

Twitter est un réseau social et un micro blog qui permet aux utilisateurs de publier des messages en temps réel, appelés tweets. Les tweets sont des messages courts, limites à 140 caractères. En raison de la nature de ce service de microblogging (messages rapides et courts), les gens utilisent des acronymes, commissent des erreurs d'orthographe, utilisent des émoticônes et d'autres caractéristiques qui expriment des significations particulières [APOORV AGARWAL, BOYI XIE, ILIA VOVSHA, 2011] Twitter est actuellement l'un des plates-formes de micro-blogage les plus populaires. Son premier slogan était Que faites-vous ? Néanmoins l'utilisation a pris une autre piste ou les utilisateurs échangent des avis et des informations, le slogan devient " Quoi de neuf ? ". Plusieurs célébrités utilisent Twitter, on y trouve même des chefs d'´Etat.

2.8.1 Selon les derniers chiffres

— Twitter a plus que 645 millions utilisateurs inscrits.

— 58 millions de tweets envoyés chaque jour.

Dans le cadre de l'analyse des sentiments, la petite taille de message formule l'hypothèse que ce message ne renferme pas a priori plus d'une seule idée, ce qui facilite l'identification de la cible d'une opinion. Mais certains tweets apparaissent comme des messages codes à cause de l'usage des hashtags, abréviations en tout genre, argot, et émoticons. Les termes à connaitre pour bien utiliser Twitter, des vocabulaires specifiques sont utilise sur Twitter plus couramment [LAURENT DIJOUX, 2009]

— Followers : les personnes qui vous suivent.

— Followings : les personnes que vous suivez.

— Friends : les personnes que vous suivez et qui vous suivent

— Twittos : les utilisateurs de Twitter.

— Tweet : court message.

— Tweeter : envoyer/poster un message.

2.8.2 Caractéristique d'un Tweet

On peut se sentir un peu perdu du vocabulaire de la langue dans les tweets, notamment,'a cause du vocabulaire et symboles spécifiques à l'utilisation de Twitter. A quoi sert le et ? C'est quoi RT ? Toutes ces abréviations peuvent paraitre un peu floues. Dans une perspective de classification, un petit lexique des principaux mots et signes Twitter est présenté [FRED COLANTONIO, 2012]

— Mention @ : se présente sous la forme @NomUtilistauer Il cible un utilisateur de Twitter dans le tweet poste. Exemple : salut à vous de la part de @AthmaneGh et @Souheib . Dans le cadre d'une réponse à un tweet, l'auteur du tweet d'origine est mentionné automatiquement dans la réponse.

— Hashtag ♯ : se présente sous la forme ♯ mot-clé. Il identifie le mot-clé en question comme important et peut en faire un sujet populaire. Exemple : ♯ coronavirus, ♯ graphisme ou encore ♯ facebook.

— RT (ReTweet) : se présente sous la forme RT NomUtilisateur. Il permet de partager le tweet d'un utilisateur. Exemple : RT AthmaneGh Excellent.

— URL (Lien) : se présente sous la forme https :// ou http ://www. Twitter permet à l'utilisateur de rejoindre les liens dans son tweet. Exemple : https ://web.stanford.edu ou http ://www-nlp.stanford.edu/IR-book/.

— VIA : s'utilise pour mentionner votre source d'information, dans votre tweet. Exemple : Via YouTube, Via Facebook.

2.9 Conclusion

L'analyse des sentiments est un domaine intéressant, o'u ce domaine est largement utilisé par les grandes entreprises et les grandes firmes pour avoir une idée de la façon dont les clients sont heureux avec les produits 'a partir du rapport entre les tweet positifs et négatifs à leur sujet. Il peut également être utilisé pour trouver des personnes qui sont satisfaites des produits ou services et leurs expériences peuvent être utilisés pour promouvoir ces produits.

L'analyse des sentiments est une tache un peu complexe et a besoin de grands efforts surtout avec les langages vernaculaires écrits dans les réseaux sociaux.

Chapitre 3

Conception de système

3.1 Introduction

Dans ce chapitre on va expliquer les étapes et les modules composant notre système, ou nous présentons la conception de notre système en commençant par sa conception générale puis sa conception détaillée en expliquant les différents éléments du système et précisant leur fonctionnement.

3.2 Méthodologie suivie

Pour réaliser à notre système, nous avons appliqué une méthode de classification c'est similarité cosinus.
Cette méthode, à besoin d'un grand corpus marqué (chaque donnée est attribuée à sa classe), et besoin d'une technique pour rendre ce corpus compréhensible pour la machine. Et pour cela, nous avons téléchargé une dataset qui contient des tweets qui sont écrit en Arabe standard telle que chaque tweet et marqué sentiments positifs ou négatifs utilisé pour l'entrainement et le test de notre model.

3.3 Conception globale du système

Globalement, on peut représente l'architecture de notre système de catégorisation des sentiments d'un texte comme suit :

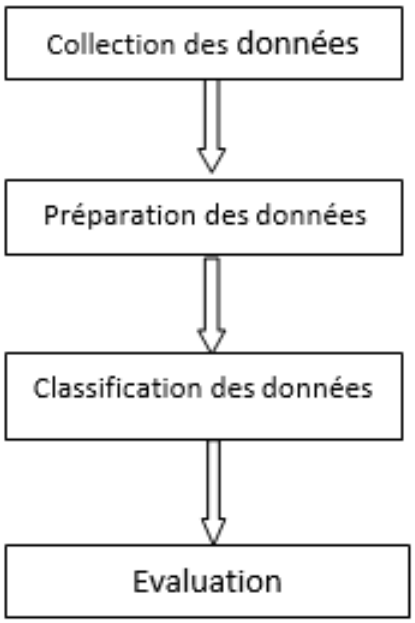


FIGURE 3.1 – L'architecture générale du système

3.3.1 Collection des données

Nous avons utilisé dataset qui contient des tweets étiqueté écrit en Arabe standard (l'entrée d'entrainement de modèle de catégorisation des sentiments) telle que chaque tweet et marqué sentiments positifs ou négatifs et utilisé pour l'entrainement et le test de notre models.
On a téléchargé La dataset qui nous avons utilisé de site suivant :
https ://github.com/bakrianoo/aravec

3.3.2 Préparation des données

Dans se chapitre on va utiliser le prétraitement des textes.
Les corpus textuels constituent la matière première des applications de traitement automatiques de la langue. Mais souvent les textes contiennent des mots mal orthographiés ou collés, des incohérences typographiques, des phrases grammaticales, des caractères bizarres ou dans un encodage différents de celui attendu par le programme, etc. Les textes en quelque sorte « bruités ». Les textes doivent subir un « nettoyage » et de normalisation appropriée. Avant de pouvoir être traité par les programmes de traitement automatiques des langages naturels. Cette phase appelée prétraitement du texte est donc essentielle. Si elle est négligée ou réalisée de façon trop simpliste. Les systèmes risquent de fausser leur résultats (J.-M. Torres-Moreno, 2011). Au cours de cette phase, nous allons mener trois phases la segmentation, la suppression des mots inutiles et la stemmatisation

— **Segmentation :** La segmentation est une étape fondamentale dans le traitement automatique d'un texte, son rôle est de découper un texte en unités d'un certain type. La segmentation d'un texte informatisé est l'opération de délimitation des segments de ses éléments de base qui sont les caractères, en éléments constituants de différents niveaux structurels : paragraphe, phrase, syntagme, mot.

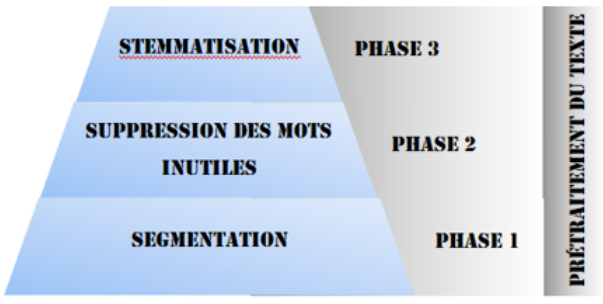


FIGURE 3.2 – Les différentes phases du processus de prétréitémént du texte

Exemples :

	Avant la segmentation	Après la segmentation
Exemple 1	/walilwakalat/وللوكالات « Et pour les agences »	/wa+li+al+wakalat+at/و+ل+ال+وكالة+ات Et+pour+les+agences+pluriel
Exemple 2	/wasanaf`aluhaa/وسنفعلها « et on va la faire »	/wa+sa+na+f`alu+ha/و +س +ن +فعل +ها « et+on+nous+faire+elle »

FIGURE 3.3 – Exemples de segmentation de mots dans la langue arabe

— **Suppression des mots inutiles :** Les mots les plus fréquents n'apportent pas, généralement une grande quantité d'informations. On les appelle mots ou termes vides de sens. En système d'information, les termes vides comme les articles, les conjonctions, les chiffres, la ponctuation, et les symboles spéciaux peuvent être supprimé lors d'un filtrage. Les mots sont appelés en anglais stop-words. Typiquement environs de 26,037 des mots vides dans la langue arabe. Ils sont réunis dans le site web Sourceforge[sourceforge.net/projects/arabicstopwords]

— **Stemmatisation :** La stemmatisation est une technique morphologique largement utilisée pour la préparation des textes dans une recherche documentaire. Elleconsiste à rechercher la racine lexicale ou stem[La forme du mot après l'enlèvement de toutes les affixes] pour des mots en langue naturelle, et ceci, par l'élimination des affixes qui leur sont rattachés, en d'autre terme regrouper sous un même identifiant des mots dont la racine est communes.

Exemple :

Avant la stemmatisation	Après la stemmatisation
صادق، الصدق الصديق، صدق، يصدق	les mots sont des flexions du mot "صدق"

FIGURE 3.4 – Exemple de stemmatisation de mots dans la langue arabe

La majorité des données textuelles utiles de Twitter sont habituellement représentées sous une forme non structurée. Par conséquent, l'application directe de la classification des sentiments à ces données textuelles peut donner de mauvais résultats.

C'est pourquoi les techniques de prétraitement sont importantes pour améliorer la valeur des données, contribuant ainsi aux processus d'analyse des sentiments. Voici quelques-unes des techniques de prétraitement qui doivent être appliquées :

- Nettoyage des tweets : c'est-à-dire la suppression de renseignements non pertinents, par exemple, les URL, les caractères et les noms d'utilisateur spéciaux de Twitter, par exemple #, @, FF et tous les mots non critiques. De plus, nous éliminons les retweets.
- Normalisation : cela transformera le texte arabe en une forme cohérente. La normalisation consiste en plusieurs sous-tâches : éliminer les signes diacritiques; éliminer la lettre Hamza (ء); unifier la lettre alif, donc أ et إ sont remplacés par ا; remplacer ة par ه; (e) remplacer ى par ي; et supprimer la lettre kashida (allongement des mots), par exemple, الحمـــد devient الحمد.
- Suppression de mots-clés : ce sont des mots qui n'ont pas de sens ou qui ne contiennent pas d'information, comme les conjonctions, les articles et les prépositions. L'élimination des stop-words du texte aide à identifier les mots les plus importants. Ici, nous allons éliminer des mots tels que (في dans), (من de) et (على sur). En outre, il est sage de supprimer les mots fréquemment présents car ils ont peu de contenu d'information, par exemple, (مثل comme), (يريد besoin) et (يقول dire).
- Élimination de l'effet de la parole : c'est une pratique courante dans les gazouillis et les autres médias sociaux, où l'une des lettres est répétée plusieurs fois, par exemple COOOOOOOOL.

Exemple :

Dans set exemple , nous appliquons les étapes précédentes a un Tweet .

Pre-processing step	Processed tweet
The original tweet	#مرفوووض قيادة_المرأة_للسيارات مستحيييل أنا أرفض قيادتها تماما ... كافي شوار عنا زحمممة!! هذا الشي
Tweet cleaning	مستحيييل أنا أرفض قيادتها تماما كافي شوار عنا زحمممة هذا الشي مرفوووض
Normalization	مستحيييل انا ارفض قيادتها تماما كافي شوار عنا زحمممة هذا الشي مرفوووض
Stop-word filtering	مستحيييل انا ارفض قيادتها تماما كافي شوار عنا زحمممة مرفوووض
Elimination of speech effect	مستحيل انا ارفض قيادتها تماما كافي شوار عنا زحمة مرفوض
Stemming	مستحيل انا رفض قياده تماما كافي شوارع زحمه رفض

3.3.3 La classification des données :

L'analyse des sentiments renvoie parfois à la classification des sentiments, car celle-ci implique les opérations importantes de l'analyse des sentiments. Nous pouvons classer les approches de classification des sentiments qui ont été utilisées dans la recherche en deux catégories : l'approche de l'apprentissage automatique et l'approche d'orientation sémantique. Dans la première approche, chaque document est étiqueté avec sa caractéristique, par exemple, la polarité d'opinion positive ou négative, et les données étiquetées sont utilisées comme une entrée à un classificateur, qui utilise un algorithme d'apprentissage automatique pour le former. Ainsi, le classificateur construit automatiquement un modèle. Dans cette dernière approche, nous créons un lexique de sentiment d'une langue. Chaque mot du lexique a un degré de positivité qui indique sa classe (positive, négative ou neutre). La polarité du document est calculée par un lexique qui extrait tous les mots du sentiment d'un texte, puis résume son degré de positivité pour déterminer si le sentiment global du document est positif ou négatif. L'avantage de l'approche d'orientation sémantique est qu'aucune donnée de formation n'est requise et qu'elle correspond à tous les domaines. La collecte de l'ensemble de données est habituellement un travail ardu ; par conséquent, une classification manuelle d'une grande quantité de données est nécessaire pour construire l'ensemble de formation. En revanche, l'approche de l'apprentissage automatique est considérée comme plus précise puisque la formation sur les données provient d'un domaine spécifique.

Exemple :

FIGURE 3.5 – Classification de polarité des tweets arabes

L'objectif de la classification de la subjectivité est de déterminer si la phrase est subjective ou objective. Une phrase subjective contient certaines orientations ou sentiments personnels, alors qu'une phrase objective est davantage une information factuelle. En outre, la classification des sentiments peut être divisée sur la base du texte donné. Le sentiment ou l'opinion peut être classé au niveau du document, de la phrase ou de l'aspect. La figure 2 présente un exemple d'analyse des sentiments à des niveaux différents. Dans le cas du niveau de document, nous nous préoccupons du sentiment de l'ensemble du document, par exemple, l'analyse du sentiment d'un film ou une revue de produit. Dans l'analyse au niveau de la phrase, nous suivons le sentiment de chaque phrase. Un exemple courant de classification au niveau de la phrase est la classification du sentiment des tweets où chaque tweet ne dépasse pas 140 caractères. Pour la classification au niveau de l'aspect, nous sommes concernés par l'extraction du sentiment ou de l'opinion sur différents aspects des objets, par exemple, un appareil photo numérique ou mobile. C'est plus difficile car il faut identifier les objets du texte avant de s'attaquer à leurs aspects (caractéristiques) et déterminer le sentiment qui s'exprime pour chaque caractéristique, qu'elle soit positive ou négative.

Analyse des sentiments au niveau de la phrase		
Texte	Sentiment	
L'écran tactile est merveilleux	Positive	
Analyse des sentiments au niveau de document		
Texte	Sentiment	
'ai acheté un iPhone il y a quelques jours. C'est un tel téléphone, bien qu'un petit grand. L'écran tactile est cool. La qualité vocale est aussi claire. Je l'aime simplement.	Positive	
Analyse des sentiments au niveau de L'aspect		
Texte	Aspect	Sentiment
La qualité des appels de l'iphone est bonne, mais la durée de vie de la batterie est courte.	qualité des appels	Positive
	durée de vie de la batterie	Négative

FIGURE 3.6 – Exemple d'analyse des sentiments à différents niveaux

Nous avons l'intention de développer un système d'analyse des sentiments qui vise à identifier la polarité des tweets, qu'ils soient positifs ou négatifs, qui répondent à l'un des problèmes sociaux en Algérie.

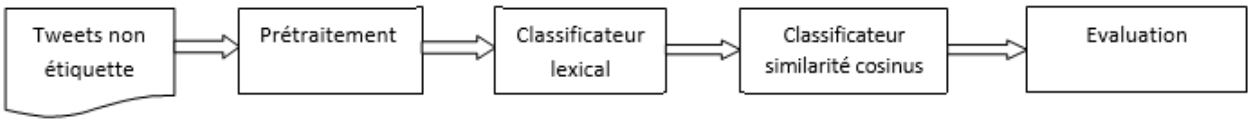


FIGURE 3.7 – Architecture générale de la conception recommandée.

Dans ce module, on va classifier nos données (les mots arabes), on utilise similarité cosinus, pour utilisé cette méthode ces données doivent être converties en vecteurs, la tâche de convertir un texte donné en un vecteur de caractéristique est une tâche importante dans le traitement de texte en termes d'extraction des caractéristiques les plus importantes, cette tache déjà traité dans dataset qui on a téléchargé.

Le vecteur de mot " قطة " :

```
[ 0.6214019 ,   2.664876  ,  -2.4490244 ,  -0.13141291,   1.0106287 ,
  1.4277642 ,  -0.6019407 ,  -0.37155798,   2.2610269 ,  -0.51503485,
 -1.7400011 ,   1.4599515 ,   1.3110927 ,   0.4506139 ,   1.1511235 ,
 -2.3989084 ,   0.0108205 ,  -0.93597263,   0.20742278,   2.7626824 ,
 -0.21789424,  -2.6269352 ,  -0.033042  ,  -2.0458148 ,   1.4766251 ,
 -2.589866  ,  -1.7341375 ,  -1.5589778 ,   0.57571614,   4.2727513 ,
  0.02701492,   1.77316   ,  -1.1816478 ,   0.24516247,  -0.04227808,
  0.57215565,   3.2628767 ,  -1.2422727 ,   1.2351261 ,   1.7213373 ,
 -2.2107098 ,   3.334359  ,   0.835815  ,   0.27691752,  -0.61994714,
  2.0607152 ,  -0.33151346,   2.132865  ,  -1.1516991 ,  -0.39679298,
 -2.1682317 ,   1.5982645 ,  -1.1571178 ,   1.3672193 ,  -0.81996626,
  0.5634883 ,   0.8571397 ,   1.2602032 ,   1.5811064 ,  -2.6346667 ,
 -0.21950944,  -1.7665412 ,   1.3162723 ,  -0.9176698 ,  -0.5075662 ,
 -0.6396452 ,  -0.57308793,   2.6602883 ,   1.466169  ,  -0.54523975,
  1.0440696 ,  -0.8016639 ,   0.95874494,  -0.8008114 ,   1.4913949 ,
  0.9796351 ,  -1.3504812 ,  -0.16031194,   0.779816  ,  -2.036837  ,
  0.5117812 ,  -1.174033  ,   3.356553  ,   1.5459414 ,  -1.1024675 ,
  1.4124179 ,   1.0076581 ,  -0.23065878,   3.9290988 ,   0.24867593,
 -0.8912038 ,   0.7108348 ,   0.40351257,   3.1929119 ,  -0.7811022 ,
 -2.341077  ,   0.38009226,  -0.7102923 ,  -0.6132934 ,  -0.88354295 ]
```

La similarité cosinus mesure la similitude entre deux vecteurs d'un espace intérieur du produit. Il est mesuré par le cosinus de l'angle entre deux vecteurs et détermine si deux vecteurs pointent dans à peu près la même direction. Il est souvent utilisé pour mesurer la similitude des documents dans l'analyse des textes.
La similarité cosinus est fréquemment utilisée en tant que mesure de ressemblance

entre deux documents. Il pourra s'agir de comparer les textes issus d'un corpus dans une optique de classification (regrouper tous les documents relatifs à une thématique particulière), ou de recherche d'information (dans ce cas, un document vectorisé est constitué par les mots de la requête et est comparé par mesure de cosinus de l'angle avec des vecteurs correspondant à tous les documents présents dans le corpus. On évalue ainsi lesquels sont les plus proches)[wikipédia].
La mesure d'angle entre deux vecteurs ne pouvant être réalisée qu'avec des valeurs numériques, il faut imaginer un moyen de convertir les mots d'un document en nombres. On partira d'un index correspondant aux mots présents dans les documents puis on attribuera à ces mots des valeurs. La forme la plus simple pourrait être de compter le nombre d'occurrences des mots dans les documents [wikipédia].

$$Similarity = \cos(\beta) = \frac{A.B}{|A|.|B|}$$

Ou A et B sont des vecteurs représentés les phrases. A c'est le statut et B le commentaire

Exemple :

Le tweet : تشهد الطرقات الجزائرية كثره النساء التي تقّود السيارات

Commentaire 1 : مستحيل أن أسمح لأي امرأه أن تقّود سيارتي

Commentaire 2 : يعود ذلك للوعي في عقّلية المواطن الجزائري

En utilisant les étapes précédemment de prétraitement de donnes, puis on utilise dataset pour extraire le vecteur de chaque mot, puis on faire la somme de tous les vecteurs de chaque phrase (statut ou commentaire) alors
La somme de tous les vecteurs des mots de notre statut est :

```
[ 0.2313719 ,   2.694856  ,   1.4490586 ,   1.69011923, -2.01112834 ,
  0.6947642 ,  -0.5459469 ,   0.37354398,  -2.2610269 ,  -0.51503485,
 -1.7400011 ,   1.4599515 ,   1.3110927 ,   0.4506139 ,   1.4475235 ,
 -2.3989084 ,   0.0108205 ,  -0.93597263,   0.20742278,   2.7626824 ,
 -0.21789424,  -2.6269352 ,  -0.033042  ,  -2.0458148 ,   1.4766251 ,
 -2.589866  ,  -1.7341375 ,  -1.5589778 ,   0.57571614,   4.2727513 ,
  0.02701492,   1.77316   ,  -1.181634  ,   0.24516247,  -0.04227808,
  0.57215565,  -3.2628767 ,  -1.2422727 ,   1.2351261 ,   1.7213373 ,
 -2.2107098 ,   3.334359  ,   0.835815  ,   0.27691752,  -0.61994714,
  2.0607152 ,  -0.33151346,   2.132865  ,  -1.1516991 ,  -0.39679298,
 -2.1682317 ,   1.5982645 ,  -1.1571178 ,   1.3672193 ,  -0.81996626,
  0.5634883 ,   0.8571397 ,   1.2602032 ,   1.5811064 ,  -2.6346667 ,
 -0.21950944,  -1.7665412 ,   1.3162723 ,  -0.9176698 ,  -0.5075662 ,
 -0.4752452 ,  -0.57308793,   2.6602883 ,   1.466169  ,  -0.54523975,
  1.0440696 ,  -0.8016639 ,   0.95874494,  -0.8008114 ,   1.4567949 ,
  0.9796351 ,  -1.3504812 ,  -0.16031194,   0.779816  ,  -2.036837  ,
  0.5117812 ,  -1.174033  ,   3.356553  ,   1.5459414 ,  -1.1024675 ,
  1.4124179 ,   1.0076581 ,  -0.23065878,   3.9290988 ,   0.24867593,
 -0.8912038 ,   0.7108348 ,   0.40351257,   3.1929119 ,  -0.7811022 ,
 -2.341077  ,   0.66449226,  -0.7858523 ,  -0.6132934 ,   3.85353829 ]
```

La somme de tous les vecteurs des mots de commentaire 1 est :

```
[-3.6213245 ,  -2.664876  ,  -0.447544 ,  -4.1333745 ,   1.0106287  ,
  1.4277642 ,  -0.6019407 ,  -0.37155798,   2.2610269 ,  -0.51503485,
  1.7000011 ,   1.4599515 ,   1.3110927 ,   0.4506139 ,   1.1511235 ,
 -2.3989084 ,   0.0108205 ,  -0.93597263,   0.20742278,   2.7626824 ,
 -0.21789424,  -2.6269352 ,  -0.033042  ,  -2.0458148 ,   1.4766251 ,
 -2.589866  ,  -1.7341375 ,  -1.5589778 ,   0.57571614,   4.2727513 ,
  0.02701492,   1.7731633 ,  -1.1816478 ,   0.24516247,  -0.04227808,
  0.57215565,   3.2628767 ,  -1.2422727 ,   1.2351261 ,   1.7213373 ,
 -2.2107098 ,   3.334359  ,   0.835815  ,   0.27691752,  -0.61994714,
  2.0607152 ,  -0.33001346,   2.132865  ,  -1.1516991 ,  -0.39679298,
 -2.1682317 ,   1.5982645 ,  -1.1571178 ,   1.3672193 ,  -0.81996626,
  0.5634883 ,   0.8571397 ,   1.2602032 ,   1.5811064 ,  -2.6346667 ,
 -0.21950944,  -1.7665412 ,   1.3162723 ,  -0.9176698 ,  -0.5075662 ,
 -0.6396452 ,  -0.57308793,   2.6607673 ,   1.466169  ,  -0.54523975,
  1.0440696 ,  -0.8016639 ,   0.95874494,  -0.8008114 ,   1.4913949 ,
  0.9796351 ,  -1.3504812 ,  -0.16031194,   0.779816  ,  -2.036837  ,
  0.5117812 ,  -1.174033  ,   3.3565539 ,   1.7649414 ,  -1.1024675 ,
  1.4124179 ,   1.0076581 ,  -0.23065878,   3.9290988 ,   0.24867593,
 -0.8912038 ,   0.7108348 ,   0.40351257,   3.1929119 ,  -0.7811022 ,
 -2.341077  ,   0.38009226,  -0.7102923 ,  -3.6132934 ,  -0.38856285 ]
```

La somme de tous les vecteurs des mots de commentaire2 est :

```
[ 3.477382  ,   1.8877542 ,  -1.4477246 ,   1.69011923, -2.01112834 ,
  0.6947642 ,  -0.5459469 ,   0.373548  ,  -2.2610269 ,  -0.51503485,
 -1.7400011 ,   1.4599515 ,   1.3110927 ,  -0.4506139 ,   1.4475235 ,
 -2.3989084 ,   0.0108205 ,  -0.93597263,  -0.20742278,   2.7626824 ,
 -0.21789424,  -2.6269352 ,  -0.033042  ,  -2.0458148 ,   1.4766251 ,
 -2.589866  ,  -1.7341375 ,  -1.5589778 ,  -0.57571614,   4.2727513 ,
  0.02701492,   1.77316   ,  -1.181634  ,   0.24516247,  -0.04288808,
  0.57215565,  -3.2628767 ,  -1.2422727 ,   1.2351261 ,   1.7213373 ,
 -2.2107098 ,   3.334359  ,   0.835815  ,   0.27691752,  -0.61994714,
  2.0607152 ,  -0.33151346,   2.384865  ,  -1.1516991 ,  -0.39679298,
 -2.1682317 ,   1.5982645 ,  -1.1571178 ,   1.3672193 ,  -0.81996626,
  0.5634883 ,   0.8571397 ,   1.2602032 ,   1.5811064 ,  -2.6346667 ,
 -0.21950944,  -1.7665412 ,   1.3162723 ,  -0.9176698 ,  -0.5075662 ,
 -0.4752452 ,  -0.57308793,   2.62883   ,   1.466169  ,  -0.54523975,
  1.0440696 ,  -0.8016639 ,   0.9587449 ,  -0.8008114 ,   1.4567949 ,
  0.9796351 ,  -1.3504812 ,  -0.16031194,   0.7678816 ,  -2.03683776,
  0.5117812 ,  -1.174033  ,   3.356553  ,   1.5459414 ,  -1.1024675 ,
  1.4124179 ,   1.0076581 ,  -0.23065878,   3.9290988 ,   0.24867593,
 -0.8912038 ,   0.7108348 ,   0.40351257,   3.1929119 ,  -0.7811022 ,
 -2.341077  ,   0.66449226,  -0.7858523 ,  -0.6132934 ,   3.85353829 ]
```

Maintenant nous appliquons le loi de cosinus similarité pour le statut et le commentaire 1 alors :

cos (β_1) = 0,32402

Maintenant nous appliquons le loi de cosinus similarité pour le statut et le commentaire 2 alors :

cos (β_2) = 0,80214

3.3.4 Evaluation

Dans ce module on utilise le résultat de similarité cosinus pour classifier les commentaires de chaque tweet, on suppose que le seul de classification est 0,5.

Si

0 ¡ cos (β) ¡ 0,5 le commentaire est négative

0,5 ¡ cos ($beta$) ¡ 1 le commentaire est positive

cos ($beta$) = 0,5 le commentaire est neutre

3.4 Conclusion

Dans ce chapitre, on a présenté la conception de notre système et on a bien expliqué les modules utilisés et les trois phases essentielles (Collection des données, Préparation des données et la classification des donnes) ainsi que nous avons vue les déférentes méthodes utilisé de prétraitement des textes.

Chapitre 4

Réalisation

4.1 Introduction

Nous allons présenter l'environnement de travail, le langage de programmation, et les outils que nous avons utilisé pour construire l'application.

4.2 Environnement et outils de développement

4.2.1 Environnement de développement

FIGURE 4.1 – Python logo

Python est un langage de programmation de haut niveau utiliser pour la programmation générale. Crier par Guido van Rossum et sorti en 1991, Python a une philosophie de conception qui met l'accent sur la lisibilité du code, notamment en utilisant des espaces importants. Il fournit des constructions qui permettent une programmation claire à petite et à grande échelle. Python dispose d'un système de type dynamique et d'une gestion automatique de la mémoire. Il prend en charge de multiples paradigmes de programmation, y compris orientée objet, impératif, fonctionnel et procédural, et dispose d'une bibliothèque standard vaste et complète. Les interpréteurs Python sont disponibles pour de nombreux systèmes d'exploitation. CPython, l'implémentation de référence de Python, est un logiciel open source et possède un modèle de d´enveloppement bas´e sur la communauté, comme presque toutes ses implémentations de variantes. CPython est g´er´e par la fondation Python Software à but non lucratif [www.python.org.]

4.3 Les outils utilisés

— **Twitter API**

L'API Twitter est une plate-forme pour créer des applications qui sont disponibles pour les membres du réseau social de Twitter. L'API permet d'extraire des tweets.

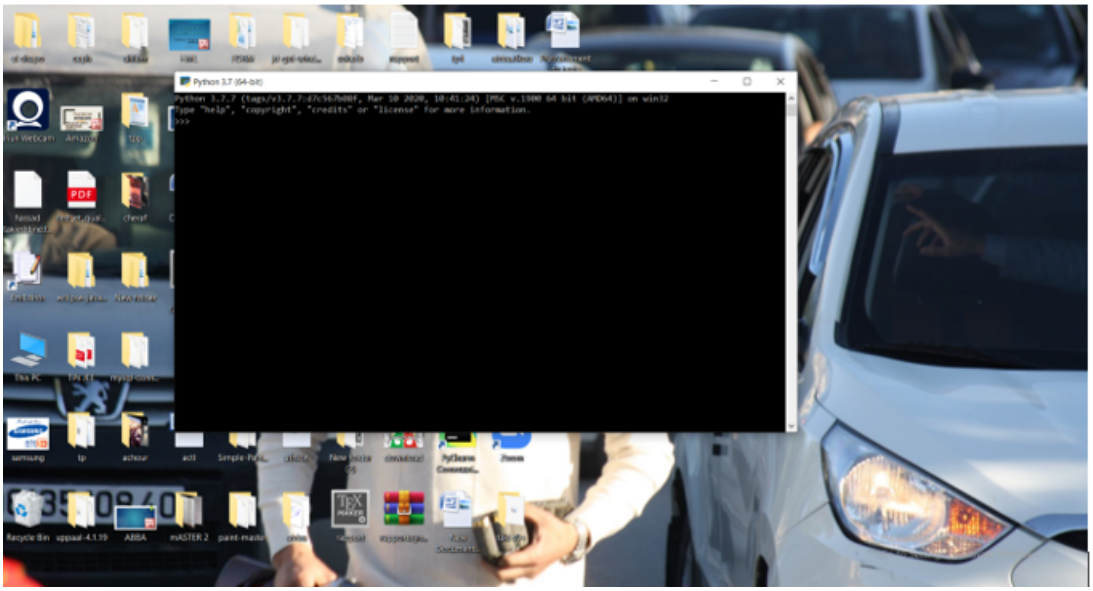
— **Sklearn**

Cette bibliothèque contient de nombreux outils efficaces pour l'apprentissage automatique et la modélisation statistique, notamment la classification, la régression, le regroupement et la réduction des réductions. Veuillez noter que scikit-learn est utilisé pour construire des modèles.

— **Numpy**

NumPy est le paquet fondamental du calcul scientifique en Python. C'est une bibliothèque Python qui fournit un objet tableau multidimensionnel, divers objets dérivés (tels que des tableaux et matrices masqués) et un assortiment de routines permettant d'effectuer des opérations rapides sur des tableaux

4.4 Quelque capture



4.5 Conclusion

Dans cette annexe Nous avons présenté l'environnement de travail, le langage de programmation, et quelque capture de notre travail.

```python
import gensim
import re
import spacy

# Clean/Normalize Arabic Text
def clean_str(text):
    search = ["أ","إ","آ","ة","_","-","/",".","،"," و "," يا ",'"',"ـ","'","ى","\\",'\n', '\t','&quot;','?','؟','!']
    replace = ["ا","ا","ا","ه"," "," ","","",""," و"," يا","","","","ي","",' ', ' ',' ',' ? ',' ؟ ',' ! ']

    #remove tashkeel
    p_tashkeel = re.compile(r'[\u0617-\u061A\u064B-\u0652]')
    text = re.sub(p_tashkeel,"", text)

    #remove longation
    p_longation = re.compile(r'(.)\1+')
    subst = r"\1\1"
    text = re.sub(p_longation, subst, 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])

    #trim
    text = text.strip()

    return text
```

Figure 4.2 – Ce code la faire le prétraitement des donnes (tweet et commentaire)

```
Python 3.7.7 (tags/v3.7.7:d7c567b08f, Mar 10 2020, 10:41:24) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import re
>>> import math
>>> from collections import Counter
>>>
>>>
>>> def get_cosine(vec1, vec2):
...     intersection = set(vec1.keys()) & set(vec2.keys())
...     numerator = sum([vec1[x] * vec2[x] for x in intersection])
...
>>>     sum1 = sum([vec1[x]**2 for x in vec1.keys()])
  File "<stdin>", line 1
    sum1 = sum([vec1[x]**2 for x in vec1.keys()])
    ^
IndentationError: unexpected indent
>>>     sum2 = sum([vec2[x]**2 for x in vec2.keys()])
  File "<stdin>", line 1
    sum2 = sum([vec2[x]**2 for x in vec2.keys()])
    ^
IndentationError: unexpected indent
>>>     denominator = math.sqrt(sum1) * math.sqrt(sum2)
  File "<stdin>", line 1
    denominator = math.sqrt(sum1) * math.sqrt(sum2)
    ^
IndentationError: unexpected indent
>>>
>>>     if not denominator:
  File "<stdin>", line 1
    if not denominator:
    ^
IndentationError: unexpected indent
>>>         return 0.0
  File "<stdin>", line 1
    return 0.0
    ^
IndentationError: unexpected indent
>>>     else:
  File "<stdin>", line 1
    else:
    ^
IndentationError: unexpected indent
>>>         return float(numerator) / denominator
  File "<stdin>", line 1
    return float(numerator) / denominator
    ^
IndentationError: unexpected indent
>>>
>>>
>>> def text_to_vector(text):
```

FIGURE 4.3 – code de similarité cosinus

```
Python 3.7.7 (tags/v3.7.7:d7c567b08f, Mar 10 2020, 10:41:24) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> from numpy import matrix
>>>
>>> # espaces vectoriel de matrices
>>>
>>> #matrices réelles
>>> A=matrix([[1,-1,1,2],[4,4,3,5],[7,8,5,-6]]) # 3 lignes, 4 colonnes
>>> B=matrix([[-1,0,1,-2],[4,2,3,6],[0,1,5,6]]) # idem
>>>
>>> #matrices complexes
>>> C=matrix([[1,1+2.j],[1.j, -1.j]]) # carrée (2,2)
>>>
>>> print (0.5*A) # produit par un réel
[[ 0.5 -0.5  0.5  1. ]
 [ 2.   2.   1.5  2.5]
 [ 3.5  4.   2.5 -3. ]]
>>> print (A+B) # somme de deux matrice
```

FIGURE 4.4 – la somme de deux matrices

AraVec 1.0	V 2.0
AraVec 2.0	AraVec 3.0
assets	AraVec 3.0
queries	New Download links
README.md	fix queries link
aravec-with-spacy.ipynb	instructions on how to load aravec in spaCy (#11)
utilities.py	New Download links

FIGURE 4.5 – dataset utilisé dans cette conception

Model	Docs No.	Vocabularies No.	Vec-Size
Twitter-CBOW	66,900,000	1,476,715	300
Twitter-CBOW	66,900,000	1,476,715	100
Twitter-SkipGram	66,900,000	1,476,715	300
Twitter-SkipGram	66,900,000	1,476,715	100

FIGURE 4.6 – la taille de dataset

Conclusion générale

Le web est devenu une plateforme de lecture et écriture où les utilisateurs ne sont plus strictement des consommateurs d'informations mais aussi des producteurs. Le contenu généré par l'utilisateur, sous forme de texte libre non structuré, devient partie intégrante du web principalement en raison de l'augmentation spectaculaire des sites de réseaux sociaux, des sites de partage de vidéos, des nouvelles en ligne, des sites de critiques en ligne, des forums en ligne et des blogs. En raison de cette prolifération de contenu généré par les utilisateurs, l'exploration de contenu web suscite une attention considérable en raison de son importance pour de nombreuses entreprises, agences gouvernementales et institutions, où l'analyse des sentiments est un sous-domaine important de l'exploration de contenu web. Pour effectuer l'analyse des sentiments sur un texte Arabe, nous avons utilisé des méthodes de l'apprentissage de la machine (Machine Learning) cosinus similarité et le prétraitement du texte et faire une comparaison entre elles selon la précision des résultats.

Dans ces méthodes nous avons entrainé les modèles de catégorisation des sentiments sur dataset et on a atteint des précisions acceptables est qui nous encourage à améliorer nos recherches de plus en plus.

Bibliographies

— Russell, S. J., Norvig, P. (2016). Artificial intelligence : a modern approach. Malaysia ; Pearson Education Limited, « Meghyn. Bienvenu. Introduction à l'IntelligenceArtificielle. Université Aix-Marseille 1. 2007-2008 et 2008-2009. ».

— Touzet, C. (1992). Les réseaux de neurones artificiels, introduction au connexionnisme. EC2.

— Thibodeau-Laufer, E. (2014). Algorithmes d'apprentissage profonds supervisés et non-supervisés : applications et résultats théoriques, (maître ès sciences, Université de Montréal).

— Portugal, I., Alencar, P., Cowan, D. (2017). The use of machine learning algorithms in recommender systems : a systematic review. Expert Systems with Applications.

— Priyadharshini. (March 8, 2018). Machine Learning : What it is and Why it Matters, ''simplilearn,"https ://www.simplilearn.com/.

— Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. "O'Reilly Media, Inc.".

— Chapelle, O., Scholkopf, B., Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds. ; 2006) [book reviews]. IEEE Transactions on Neural Networks, 20(3), 542-542.

— Hamel, P. (2012). Apprentissage de représentations musicales à l'aide d'architectures profondes et multiéchelles, (Doctoral dissertation, Université de Montréal).

— Glorot, X. (2015). Apprentissage des réseaux de neurones profonds et applications en traitement automatique de la langue naturelle, (Doctoral dissertation, Université de Montréal).

— Bastien L, Deep Learning ou apprentissage profond : définition,le magazine Cloud and big datahttps ://www.lebigdata.fr/, 27 février 2018

— Dorianne W, Intelligence artificielle, machine learning, deep learning : kézako ?, ledigitalab https ://www.ledigitalab.com/fr/,2 October 2017,page2/10.

— Rémi S, Le Deep Learning pour tous, internetactu http ://www.internetactu.net/,2 October 2014.

— Philippe B : une première introduction au Deep Learning, Microsoft—Developer, Machine Learning France https ://msdn.microsoft.com/fr-fr/,April 28, 2016.

— Baghli.L « Contribution à la commande de la machine asynchrone, l'utilisation de la logique floue, des réseaux de neurones et des algorithmes génétiques

» Thèse de Doctorat, université Nancy I, 1999.
— Benoît.V ,Etude prospective des applications possibles des réseaux de neurones formels dans le traitement des données psychométriques,2001.
— Dreyfus D., Martinez J.-M., Samuelides M., Gordon M. B., Badran F., Thiria S. et Hérault L., Réseaux de neurones, méthodologie et applications, Eyrolles, 2ème édition, (2004).
— DREYFUS.G, MARTINEZ.J-M. SAMUELIDES.M, GORDON.M.B, . BADRAN.F, THIRIA.S. ET HERAULT.L « Réseaux de neurone, Méthodologies et Application » , Paris, Edition Eyrolles.2002.
— Hicham.C, CONCEPTION ET COMPARAISON DE LOIS DE COMMANDE ADAPTATIVE À BASE DE RÉSEAUX DE NEURONES POUR UNE ARTICULATION FLEXIBLE AVEC NON-LINÉARITÉ DURE, 2002
— Ghouili.J « Commande sans capteur d'une machine asynchrone avec estimation de la vitesse par les réseaux de neurones » Thèse de Doctorat, Université de Québec. Avril 2005.
— Jean-François.J , Les Réseaux neuromimétiques,1994.
— Kadous.D, Utilisation des réseaux de neurones comme outil du datamining :Génération de modèle comportemental d'un processusphysique à partir de données ,2012.
— Merzouka.N, Etude des performances des réseaux de neurones dynamiques à représenter des systèmes réels : une approche dans l'espace d'état, 2009.
— Zemouri.R,Contribution à la surveillance des systèmes de production à l'aide des réseaux de neurones dynamiques : Application à la e-maintenance,2003.
— Marc.P 'Réseaux de neurones', université Laval, 2004.
— MEENA RAMBOCAS AND JO ?O GAMA MARKETING RESEARCH : THE ROLE OF SENTIMENT ANALYSIS, FEP ECONOMICS AND MANAGMENT, 2013 .
— NASUKAWA, TETSUYA ET JEONGHEE YI SENTIMENT ANALYSIS : CAPTURING FAVORABILITY USING NATURAL LANGUAGE PROCESSING, KNOWLEDGE CAPTURE, 2003
— KUSHAL DAVE, STEVE LAWRENCE AND DAVID M. PENNOCK MINING THE PEANUT GALLERY : OPINION EXTRACTION AND SEMANTIC CLASSIfiCATION OF PRODUCT REVIEWS, 2003
— SANJIV R. DAS ET MIKE Y. CHEN YAHOO ! FOR AMAZON : EXTRACTING MARKET SENTIMENT FROM STOCK MESSAGE BOARDS,2001.
— SATOSHI MORINAGA,KENJI YAMANISH,KENJI TATEISHI,AND TOSHIKAZU FUKUSHIMA MINING PRODUCT REPUTATIONS ON THE WEB,PROCEEDINGS OF THE EIGHTH ACM SIGKDD INTERNATIO-

NAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, NEW YORK, NY, USA, 2002

— BO PANG, LILLIAN LEE, SHIVAKUMAR VAITHYANATHAN, THUMBS UP ? : SENTIMENT CLASSI- fiCATION USING MACHINE LEARNING TECHNIQUES, PROCEEDINGS OF THE ACL-02 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING - VOLUME 10,STROUDS- BURG, PA, USA 2002

— R. M. TONG AN OPERATIONAL SYSTEM FOR DETECTING AND TRACKING OPINIONS IN ONLINE DIS- CUSSION, IN WORKING NOTES OF THE ACM SIGIR 2001 WORKSHOP ON OPERATIONAL TEXT CLASSIfiCATION 2001

— R. M. TONG AN OPERATIONAL SYSTEM FOR DETECTING AND TRACKING OPINIONS IN ONLINE DIS- CUSSION, IN WORKING NOTES OF THE ACM SIGIR 2001 WORKSHOP ON OPERATIONAL TEXT CLASSIfiCATION 2001

— PETER D. TURNEY, THUMBS UP OR THUMBS DOWN ? : SEMANTIC ORIENTATION APPLIED TO UNSUPERVISED CLASSIfiCATION OF REVIEWS, PROCEEDINGS OF THE 40TH ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, STROUDSBURG, PA, USA, 2002.

— JANYCE WIEBE, LEARNING SUBJECTIVE ADJECTIVES FROM CORPORA,PROCEEDINGS OF THE SEVENTEENTH NATIONAL CONFERENCE ON ARTIfiCIAL INTELLIGENCE AND TWELFTH CON- FERENCE ON INNOVATIVE APPLICATIONS OF ARTIfiCIAL INTELLIGENCE, 2000

— BING LIU, OPINIONS, SENTIMENT, AND EMOTION IN TEXT,CAMBRIDGE UNIVERSITY PRESS, 2015

— BO PANG AND LILLIAN LEE, OPINION MINING AND SENTIMENT ANALYSIS,NOW PUBLISHERS INC, 2008.

— HUIFENG TANG, SONGBO TAN AND XUEQI CHENG, A SURVEY ON SENTIMENT DETECTION OF REVIEWS,INFORMATION SECURITY CENTER, INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES, BEIJING 100080, PR CHINA, 2009.

— BING LIU. WEB DATA MINING : EXPLORING HYPERLINKS, CONTENTS, AND USAGE DATA. SPRINGER SCIENCE BUSINESS MEDIA, 2007

— NITIN JINDAL AND BING LIU. MINING COMPARATIVE SENTENCES AND RELATIONS. IN AAAI, VOLUME 22, PAGES 1331–1336, 2006

— BING LIU. WEB DATA MINING : EXPLORING HYPERLINKS, CONTENTS, AND USAGE DATA. SPRINGER SCIENCE BUSINESS MEDIA, 2007.
— HADY ELSAHAR AND SAMHAA R EL-BELTAGY. BUILDING LARGE ARABIC MULTIDOMAIN RESOURCES FOR SENTIMENT ANALYSIS. IN INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, PAGES 23–34. SPRINGER, 2015
— BO PANG, LILLIAN LEE, AND SHIVAKUMAR VAITHYANATHAN. THUMBS UP ? : SENTIMENT CLASSIFICATION USING MACHINE LEARNING TECHNIQUES. IN PROCEEDINGS OF THE ACL-02 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSINGVOLUME 10, PAGES 79–86. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2002.
— PETER D TURNEY. THUMBS UP OR THUMBS DOWN ? : SEMANTIC ORIENTATION APPLIED TO UNSUPERVISED CLASSIFICATION OF REVIEWS. IN PROCEEDINGS OF THE 40TH ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, PAGES 417–424. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2002.)
— JANYCE M WIEBE, REBECCA F BRUCE, AND THOMAS P O'HARA. DEVELOPMENT AND USE OF A GOLD-STANDARD DATA SET FOR SUBJECTIVITY CLASSIFICATIONS. IN PROCEEDINGS OF THE 37TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, PAGES 246–253. ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 1999.),
— THERESA WILSON, JANYCE WIEBE, AND REBECCA HWA. JUST HOW MAD ARE YOU ? FINDING STRONG AND WEAK OPINION CLAUSES. IN AAAI, VOLUME 4, PAGES 761–769, 2004.)
— MINQING HU AND BING LIU. MINING AND SUMMARIZING CUSTOMER REVIEWS. IN PROCEEDINGS OF THE TENTH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, PAGES 168–177. ACM, 2004.).
— NITIN JINDAL AND BING LIU. MINING COMPARATIVE SENTENCES AND RELATIONS. IN AAAI, VOLUME 22, PAGES 1331–1336, 2006.).
— BING LIU. SENTIMENT ANALYSIS AND OPINION MINING. SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES, 5(1) :1–167, 2012.)
— MARTI A HEARST. DIRECTION-BASED TEXT INTERPRETATION AS

AN INFORMATION ACCESS REFINEMENT. TEXT-BASED INTELLIGENT SYSTEMS : CURRENT RESEARCH AND PRACTICE IN INFORMATION EXTRACTION AND RETRIEVAL, PAGES 257–274, 1992.)

— FAIZA BELBACHIR, EXP´ERIMENTATION DE FONCTIONS POUR LA D´ETECTION D'OPINIONS DANS LES BLOGS, UNIVERSIT´E DE PAUL SABATIER, INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE 2010..

— ALEXANDER PAK AND PATRICK PAROUBEK, TWITTER AS A CORPUS FOR SENTIMENT ANALYSIS AND OPINION MINING, UNIVERSIT´E DE PARIS-SUD, LABORATOIRE LIMSI- CNRS,FRANCE 2010..

— DOMINIQUE BOULLIER ET AUDREY LOHARD, OPINION MINING ET SENTIMENT ANALYSIS : M´ETHODES ET OUTILS, 2012.

— VIVEK KUMAR SINGH AND DEBANJAN MAHATA, A CLUSTERING AND OPINION MINING APPROACH TO SOCIO-POLITICAL ANALYSIS OF THE BLOGOSPHERE, COMPUTATIONAL INTELLIGENCE AND COMPUTING RESEARCH (ICCIC), 2010 IEEE INTERNATIONAL CONFERENCE ON 2010.

— ARTI BUCHE, DR. M. B. CHANDAK AND AKSHAY ZADGAONKAR, OPINION MINING AND ANALYSIS :A SURVEY, INTERNATIONAL JOURNAL ON NATURAL LANGUAGE COMPUTING (IJNLC), INDIA 2013.

— G.VINODHINI AND RM.CHANDRASEKARAN, SENTIMENT ANALYSIS AND OPINION MINING : A SURVEY,INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN COMPUTER SCIENCE AND SOFTWARE ENGINEERING ,INDIA 2012.

— VIVEK KUMAR SINGH AND DEBANJAN MAHATA, A CLUSTERING AND OPINION MINING APPROACH TO SOCIO-POLITICAL ANALYSIS OF THE BLOGOSPHERE, COMPUTATIONAL INTELLIGENCE AND COMPUTING RESEARCH (ICCIC), 2010 IEEE INTERNATIONAL CONFERENCE ON 2010.

— ALEXANDER PAK AND PATRICK PAROUBEK, TWITTER AS A CORPUS FOR SENTIMENT ANALYSIS AND OPINION MINING, UNIVERSIT´E DE PARIS-SUD, LABORATOIRE LIMSI- CNRS,FRANCE 2010.

— MATTHEW ERIC GLASSMAN, JACOB R. STRAUS AND COLLEEN J. SHOGAN, SOCIAL NET- WORKING AND CONSTITUENT COMMUNICATIONS : MEMBERS USE OF TWITTER AND FACEBOOK DUR- ING A TWO-MONTH PERIOD IN THE 112TH CONGRESS,CONGRESSIONAL RESEARCH SERVICE, 2009.

— APOORV AGARWAL, BOYI XIE, ILIA VOVSHA, OWEN RAMBOW AND REBECCA PASSON- NEAU, SENTIMENT ANALYSIS OF TWITTER DATA, LSM 11 PROCEEDINGS OF THE WORKSHOP ON LANGUAGES IN SOCIAL MEDIA,2011.
— LAURENT DIJOUX, BOOSTEZ VOTRE BUSINESS AVEC TWITTER,ALMABIC, 2009.
— FRED COLANTONIO, COMMUNICATION PROFESSIONNELLE EN LIGNE : COMPRENDRE ET EXPLOITER LES M´EDIAS ET R´ESEAUX SOCIAUX,EDIPRO, 2011.
— TIM O'REILLY AND SARAH MILSTEIN, THE TWITTER BOOK, 2012.
— Sigrid Maurel, Paolo Curtoni et Luca Dini, (). L'analyse des sentiments dans les forums. URL : http ://www2.lirmm.fr/ mroche/FODOP08/ArticlesFODOP08 /Article2.pdf.