République Algérienne Démocratique et Populaire Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahleb Blida

Faculté des sciences

Département d'informatique

Mémoire de fin d'études

Pour l'obtention

D'un Diplôme de master en informatique

Option : Traitement automatique de la langue (TAL)

Thème :

Analyse d'opinion multi-cibles de l'actualité en ligne

Réalisé par :

FERHAOUI Ines

BELMADI Akila

Encadreur : Mr. AMRANE Abdesalam (CERIST)

Promotrice : Mme.OUKID Lamia (USDB1)

Présidente du Jury : Mme. GUESSOUM Dalila (USDB1)

Examinatrice : Mme. BEY Fella (USDB1)

Année universitaire 2019/2020

Remerciements

En premier lieu,

Nous tenons à remercier ALLAH qui nous a aidé et nous a donné

La patience et la force à accomplir ce travail.

Nous remercions vivement

Mr. AMRANE Abdesalam notre encadreur et notre promotrice Mme.OUKID Lamia pour leurs aides, leurs conseils et leurs temps consacré pour structurer le travail et améliorer sa qualité.

Nous remercions également toutes les personnes qui nous ont aidés de près ou de loin pour la réalisation de ce travail.

Nous tenons à exprimer nos remerciements aux membres du jury pour nous avoir fait l'honneur d'accepter d'évaluer ce travail.

À tous les enseignants de l'USDB, nous vous remercions pour tous vos enseignements qui nous ont permis de réaliser ce projet et ce, tout au long de ces précédentes années.

FERHAOUI Ines

BELMADI Akila

Résumé

A l'ère d'internet et des réseaux, il y a eu une surabondance d'informations qui a conduit à une multiplicité des sources d'informations en l'occurrence avec la migration de la presse vers le web. En conséquence, les acteurs économiques et politiques sont devenus plus exigeant en matière de collecte et filtrage d'informations et surveille en continu l'évolution de leur image de marque pour améliorer la prise de décision.

Il existe plusieurs travaux dans le cadre de la classification de texte par apprentissage automatique qui ont été réalisés dans diverses langues. Ce projet vise les sources média algériennes et traite le texte exprimé en deux langues : arabe et française. L'objectif est de développer un système d'agrégation de l'actualité permettant d'identifier les entités ciblées et d'analyser la tonalité exprimée.

Des méthodes à base de règles et d'autre à base d'apprentissage automatique pour la classification ont été conçus puis évalués sur un dataset que nous avons collecté. Des améliorations ont été proposés sur le processus de classification et d'analyse d'opinion nous avons atteint des résultats de performances satisfaisons en termes de qualité de prédiction et de minimisation d'erreurs.

Enfin, nous présentons des captures d'écrans qui illustre les fonctionnalités du système proposé. Les entités figurant dans l'information ainsi que la tonalité sont visibles pour chaque article de l'actualité.

Mots clés : Agregation de flux RSS, Extraction d'entités nommées, Analyse d'opinion, Veille média.

Abstract

In the age of the internet and networks, there has been an overabundance of information and a multiplicity of information sources in this case with the migration of the press to the web. As a result, economic and political actors have become more demanding when it comes to collecting and filtering information and continuously monitor changes in their brand image to improve decision-making.

There are several works in the context of machine learning text classification that have been done in various languages. This project targets Algerian media sources and processes the text expressed in two languages, Arabic and French. The goal is to develop a news agregation system to identify targeted entities and analyze the tone expressed.

Rule-based and other machine-learning-based methods for classification were designed and then evaluated on a dataset we collected. Improvements have been proposed on the classification and opinion analysis process and we have achieved satisfactory performance results in terms of prediction quality and error minimization.

Finally, we present screenshots that illustrate the functionality of the proposed system. The entities appearing in the information as well as the tone are visible for each news article.

Keywords : RSS feed aggregation, Extracting named entities, Opinion Mining, média monitoring.

ملخص

في عصر الإنترنت والشبكات، كان هناك وفرة في المعلومات وتعدد مصادر المعلومات في هذه الحالة مع انتقال الصحافة إلى الويب. نتيجة لذلك، أصبحت الجهات الفاعلة الاقتصادية والسياسية أكثر تطلبًا عندما يتعلق الأمر بجمع المعلومات وتصفيتها والمراقبة المستمرة لتطور صورة علامتها التجارية لتحسين عملية صنع القرار.

هناك العديد من الأعمال في سياق تصنيف نصوص التعلم الآلي التي تم إجراؤها بلغات مختلفة. يستهدف هذا المشروع المصادر الإعلامية الجزائرية ويعالج النص المعبر عنه باللغتين العربية والفرنسية. الهدف هو تطوير نظام تجميع الأخبار لتحديد الكيانات المستهدفة وتحليل النبرة المعبر عنها.

تم تصميم طرق التصنيف المستندة إلى القواعد وغيرها من الأساليب القائمة على التعلم الآلي ثم تقييمها على مجموعة بيانات قمنا بجمعها. تم اقتراح تحسينات على عملية التصنيف وتحليل الرأي وحققنا نتائج أداء مرضية من حيث جودة التنبؤ وتقليل الأخطاء.

أخيرًا، نقدم لقطات شاشة توضح وظائف النظام المقترح. تظهر الكيانات التي تظهر في المعلومات بالإضافة إلى النغمة لكل مقالة إخبارية.

الكلمات الرئيسية: تجميع موجز ويب، استخراج الكيانات المحددة، تحليل الرأي، مراقبة الوسائط.

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

CRF	**C**onditional **R**andom **F**eileds
EN	**E**ntité **N**ommée
EN_PERS	**E**ntité **N**ommée de **T**ype **N**om de **P**ersonne
EN_LOC	**E**ntité **N**ommée de **T**ype **N**om de **L**ieu
EN_ORG	**E**ntité **N**ommée de **T**ype **N**om d'**O**rganisation
ML	**M**achine **L**earning
NER	**N**amed **E**ntity **R**ecognition
NERA	**A**rabic **N**amed **E**ntity **R**ecognition
NLP	**N**atural **L**anguage **P**rocessing
SVM	**S**upport **V**ector **M**achines
TAL	**T**raitement **A**utomatique de **L**anguage

Introduction Générale

Les opinions des autres ont une influence significative sur notre processus de prise de décision quotidien. Les opinions sont au cœur de presque toutes les activités humaines et sont des influenceurs clés de nos comportements. Nos croyances et nos perceptions de la réalité et les choix que nous faisons sont, dans une large mesure, conditionnés par la façon dont les autres voient et évaluent le monde. Pour cette raison, lorsque nous devons prendre une décision, il est nécessaire d'analyser et de visualiser les opinions des utilisateurs.

En effet, avec la croissance exponentielle de l'actualité mise en ligne et la quantité de données produite par les systèmes de communication et d'information dont les flux de l'actualité, il serait difficile d'analyser toute l'actualité diffusée. Le texte récupéré qui est de taille importante rend la tâche de traitement manuelle gourmande en termes de ressources et de temps.

Ce travail s'intéresse à l'analyse d'opinion basée sur le traitement du langage naturel (NLP) pour l'exploration d'opinion de l'actualité. En tant que tel, l'objectif de ce travail est d'utiliser une approche d'exploration des données, d'extraction de texte et de classification en utilisant les fragments de texte récupérés dans le but d'analyser les points de vue exprimés envers des entités cibles.

Le système développé est destiné aux agents de veille qui surveillent l'actualité diffusée par les sources média algériennes, leur permettant de spécifier les acteurs économiques à forte influence et d'analyser la tonalité exprimée envers ces derniers.

En prenant les objectifs globaux d'un angle plus près, nous citons les objectifs détaillés ci-dessous :

- Offrir aux acteurs de veille une nouvelle manière d'analyse de données afin de mieux exploiter l'information collectée ceci en exploitant les techniques de Data Mining.
- Elargir le champ d'exploitation des sources médiatiques en termes d'efficacité : la quantité des données analysées par rapport au temps.
- Proposer une solution générique, souple et capable d'évoluer en fonction des nouveaux champs d'exploitation de la cellule de veille.

Notre travail s'appuie sur la réalisation d'un site web offrant une interface à l'utilisateur qui lui permettra de créer son profil et de consulter la page revue de presse de l'actualité qu'elle présente à son tour tous les informations agrégées avec la tonalité exprimée par rapport à ces derniers.

Ce site sera géré par un administrateur qui le mettra à jour et gérer les utilisateurs et les sources.

Le présent mémoire est organisé en quatre chapitres répartis en deux parties :

La première partie est consacrée pour la synthèse bibliographique, qui représente un état de l'art sur les méthodes d'extraction des entités nommées et l'analyse de tonalité en général. Nous recensons les travaux existants dans ce domaine, ceci est réparti sur deux chapitres :

Chapitre1. Méthodes d'extraction des entités nommées : ce chapitre présente les techniques d'identification d'entités nommées ainsi que les différents phénomènes d'ambiguïté rencontrés lors de la détection de ces derniers. Nous finirons par présenter les approches les plus courants pour l'extraction des entités nommées.

Chapitre2. Méthodes d'analyse d'opinion dans l'actualité : dans ce chapitre nous abordons les différents concepts liés à l'analyse de tonalité, les domaines d'application et les défis rencontrés. Pour chaque approche nous présentons une synthèse des travaux les plus pertinents. Et nous finirons par présenter et examiner les outils d'analyse de tonalité existants.

La deuxième partie du rapport comporte notre contribution. Elle s'étale sur deux chapitres.

Chapitre3. Conception : Dans ce chapitre nous exprimons les besoins du système à concevoir et les fonctionnalités qu'il doit assurer. Nous présentons l'architecture du système et une vue globale des algorithmes à implémenter.

Chapitre4. Implémentation et évaluation : ce chapitre présente les outils de programmation et l'implémentation de notre application, présentation des interfaces et les résultats d'exécution, ainsi que l'évaluation de notre système.

Finalement, nous clôturons ce mémoire par une conclusion générale.

Partie 1. Synthèse Bibliographique

1 Méthodes d'extraction des entités nommées

1.1 Introduction

Le concept d'entité nommée (EN) est apparu dans les années 90, Les entités nommées constituent un champ de recherche très actif depuis de nombreuses années. La reconnaissance d'entité nommée **NER**, également connue sous le nom de détection des entités nommées, est une technique populaire utilisée dans l'extraction d'informations qui est un élément essentiel pour de nombreuses tâches de traitement automatique des langues **(TAL)**, qu'elles soient monolingues ou multilingues.

Dans ce chapitre nous présenterons les différentes approches d'extraction des entités nommées existantes, nous commençons à définir l'entité nommée ainsi que les catégories et les sous catégories identifiées. Ensuite, nous étudierons les différentes techniques d'identification des entités nommées et par la suite nous discuterons les différents phénomènes d'ambiguïté rencontrés lors de la reconnaissance des entités nommées et enfin nous listerons les travaux effectués sur la détection des entités nommées qui est une partie intégrante du processus de reconnaissance des **EN**.

1.2 Définition

On appelle traditionnellement « entités nommées » (en anglais Named Entity) l'ensemble des noms de personnes, d'entreprises et de lieux présents dans un texte donné. On associe souvent à ces éléments d'autres syntagmes comme les dates, les unités monétaires ou les pourcentages [1]. Comme illustre la figure suivante :

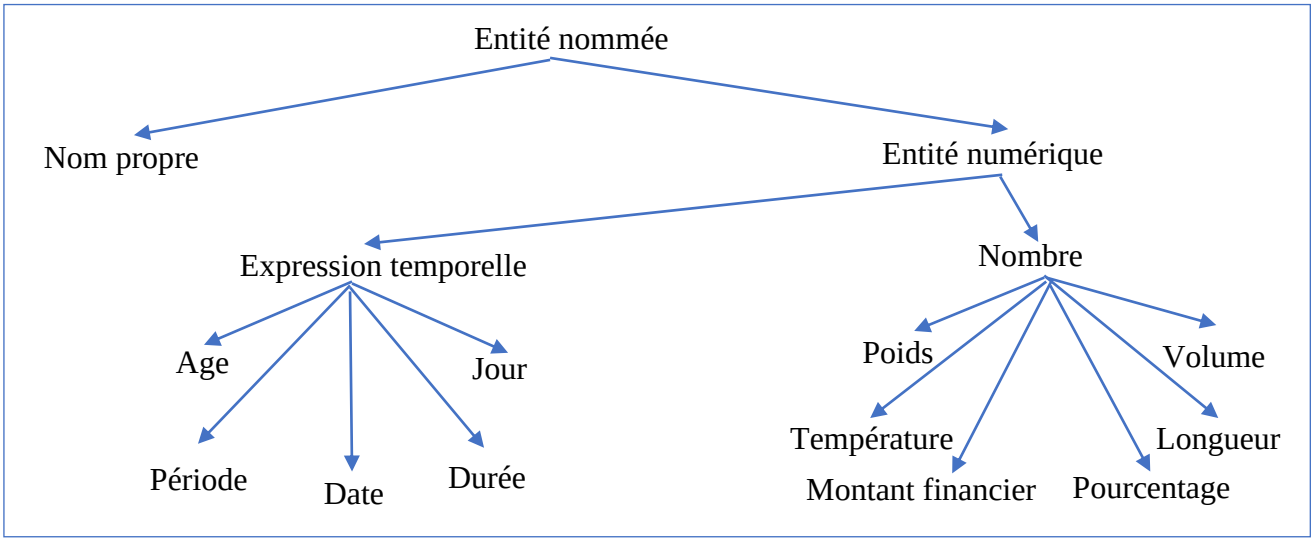

Figure 1:la hiérarchie des types d'entités nommées [2]

1.3 Catégories des entités nommés

On distingue deux formes d'EN : Les EN élémentaire et les EN complexes.

1.3.1 Les EN élémentaires

Une EN élémentaire est une EN simple qui est composée d'un seul mot, comme les noms de lieu **« Canada »** et **« Algérie »** ou le nom de personne **« Ali »**. [3]

1.3.2 Les EN complexes

Une EN complexe est une EN représentée par une liste de propriétés. Par exemple, une EN personne peut être élémentaire si on la représente uniquement par son nom, ou complexe si en plus de son nom, on a une date et un lieu de naissance. [4]

1.4 Techniques d'identification d'entités nommées

Les systèmes de reconnaissance des entités nommées reposent sur les indices qui permettent d'aider à la reconnaissance et la catégorisation des entités nommées. **McDonald** (1996) [5] distingue deux types d'indices : **internes** et **externes**. Les premiers se rapportent à ce qui permet de discerner une EN en se basant seulement sur les formes composant cette dernière. Les secondes en revanche s'intéressent à ce qui apparaît dans le contexte immédiat (formes situées à gauche et à droite de l'entité nommée) ou à partir d'un contexte plus large tel que le document ou le corpus. [6]

1.4.1 Indices internes :

Les principaux indices internes utilisés pour la reconnaissance des entités nommées sont :

- **Informations graphiques :** la majuscule est une marque typographique qui sert à débuter chacune des formes composant une entité nommée. S'appuient seulement sur cet indice pour l'identification et la délimitation des entités nommées. En revanche, la majuscule ne permet pas généralement d'aider à la catégorisation des entités nommées sauf pour les acronymes (une seule forme contenant plusieurs majuscules) qui font référence, dans la plupart des cas mais pas toujours à des organisations ou des personnes.
- **Informations concernant la ponctuation et les caractères spéciaux et numériques :** les signes graphiques peuvent être utilisés dans les acronymes et les noms des organisations et des produits. **Par exemple « I.B.M », « C&A »**, etc. En revanche, les caractères numériques permettent d'aider à l'identification et à la

catégorisation de certaines entités telles que les dates, les pourcentages et les noms des organisations.

- **Informations morphosyntaxiques :** consiste à exploiter les résultats d'un étiquetage morpho syntaxique, décrivant pour chaque mot sa catégorie grammaticale (nom propre, verbe, conjonction, etc.), afin d'élaborer des règles plus généralisatrices pour la délimitation et la catégorisation des entités nommées.
- **Informations issues de lexiques :** consiste à une simple interrogation des listes de noms propres et de mots clefs les plus courants. Ces listes sont préparées a priori manuellement ou automatiquement en utilisant des techniques d'apprentissage. Ces informations permettent dans la plupart des cas de catégoriser la forme à annoter.

1.4.2 Indices externes :

Les indices externes se rapportent au contexte d'apparition de l'entité nommée. Ils sont nécessaires lorsque les indices internes sont ambigus pour ne pas aboutir à des erreurs de classification. Par exemple **« Washington »** est à la fois une personne, une ville et un état américain. Des listes de mots déclencheurs peuvent être utilisées pour aider à catégoriser les entités nommées. Ces listes contiennent les mots qui sont susceptibles d'apparaître dans le contexte. Par exemple, un nom de personne est souvent accompagné d'un titre, d'un grade permettant d'indiquer des propriétés spécifiques. Par exemple **« Monsieur Washington »**, **« Mme Denise »**, **« l'entraîneur Aimé Jacquet »**, etc. D'autres indices externes peuvent aider à la délimitation et à la catégorisation tels que :

- La position du mot dans la phrase, par exemple, s'il est capitalisé et s'il ne se produit pas au début de la phrase alors il a une forte chance d'être une entité nommée.
- Les informations concernant les autres occurrences de l'entité nommée potentielle dans le document ou dans le corpus. Par exemple, si le mot **« Paris »** se produit dans un contexte ambigu alors qu'une autre occurrence de ce mot dans le même texte est étiquetée comme un nom de lieu, alors la première occurrence peut être aussi considérée comme un nom de lieu.
- Les méta-informations, par exemple, les balises XML et HTML pour la reconnaissance des entités nommées dans les documents structurés.

Malgré la présence d'indices internes et externes, la reconnaissance des entités nommées n'est pas une tâche facile. Cela est dû notamment à la présence de plusieurs phénomènes d'ambiguïté.

1.5 Phénomènes d'ambigüité rencontrés

1.5.1 Définition

Une entité nommée est dite ambiguë quand elle est susceptible d'avoir plusieurs interprétations concernant sa délimitation ou sa typologie.

Dans cette partie nous présenterons les différents phénomènes d'ambiguïté qui complexifient la tâche d'annotation des entités nommées. [6]

1.5.1.1 Ambiguïtés graphiques

La majuscule comme un indicateur pour le repérage et la délimitation des entités nommées n'est pas simple à manipuler pour plusieurs raisons :

- Une entité nommée peut contenir des formes commençant par une minuscule par exemple **« le château de Versailles »**, **« la faculté des Lettres »**. Ou peut comporter une ou plusieurs majuscules dans d'autres positions **par exemple « eBay »**, **« WikiLeaks »**.
- La première forme d'une phrase comporte aussi une majuscule, que ce soit une entité nommée ou non.
- L'emploi de la majuscule pour les noms propres n'est pas de règle dans toutes les langues : en allemand, tous les noms, communs ou propres, prennent une majuscule. En outre, il n'y a pas de notion de **majuscules/minuscules** pour les langues n'utilisant pas l'alphabet latin par exemple : **l'arabe, le chinois, etc.**

1.5.1.2 Ambiguïtés sémantiques

À l'instar des noms communs, les entités nommées n'échappent pas à la polysémie, à l'homonymie et à la métonymie. Prenons les exemples suivants :

- **Orange** a invité M. Dupont.
- **La France** a signé le traité de Kyoto.

Ces phénomènes complexifient la tâche de catégorisation, par exemple, est-il question de la ville d'Orange ou bien de la société de téléphonie ?

Faut-il préférer une annotation de France en tant qu'**« organisation »** ou **« gouvernement »** ou en tant que **« lieu »** ou **« pays »** ?

1.5.1.3 Ambiguïtés liées à la délimitation

L'identification des limites d'une entité nommée se heurte à plusieurs problèmes :

- **Limite droite :** le début d'une entité nommée est plus facile à identifier (présence d'une majuscule ou d'un mot déclencheur) alors que la limite droite est difficile à

trouver car les mots qui suivent ne sont pas capitalisés, **par exemple « La Fédération nationale de la Mutualité française »**.

- **Coordination** : consiste à unir deux entités nommées en effaçant l'un des constituants communs, **par exemple : « Bill et Hillary Clinton se sont envolés pour Chicago ensemble le mois dernier… ».** Ici, le problème revient à séparer **« Bill Clinton »** de **« Hillary Clinton »**. Une analyse sémantique est nécessaire.
- **Imbrication** : imbriquer une entité nommée dans une autre, **par exemple « le Comité exécutif de « l'Union des associations européennes de football »**.

1.5.1.4 Ambiguïtés liées à la langue arabe

La reconnaissance des entités nommées **(NER)** est une tâche complexe mais son niveau de complexité ne se limite pas seulement aux problèmes théoriques de **NER** tels que l'imbrication ou la coordination des entités nommées si la langue étudiée est l'arabe, mais aussi à certaines spécificités de cette langue à savoir :

- **Aucune capitalisation :** la capitalisation n'est pas une caractéristique de script arabe, à la différence des langues latines où une **EN** commence habituellement par une majuscule. Par conséquent, l'utilisation de la caractéristique de capitalisation n'est pas une option dans **NER** arabe. [7]
- **La nature agglutinative** : l'arabe a une nature agglutinative élevée en laquelle un mot peut se composer des préfixes, du lemme et des suffixes dans différentes combinaisons, qui a conduit à une morphologie très compliquée comme illustre la figure suivante. [8]

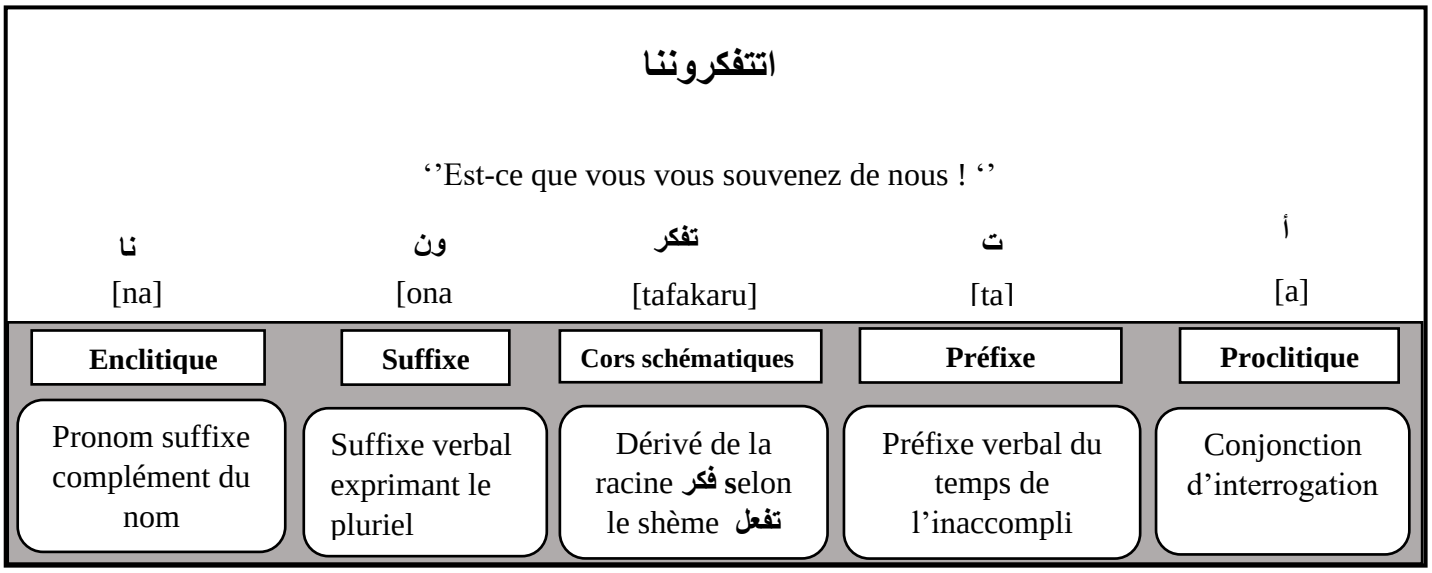

Figure 2: exemple illustratif de la nature agglutinative de la langue arabe

- **Voyelles courtes facultatives :** dans la théorie, les voyelles courtes, ou les signes diacritiques, sont nécessaires pour la prononciation et la désambigüisation. L'absence des voyelles dans les textes arabes engendre une certaine ambiguïté en ce qui concerne le sens du mot d'une part, et augmente la difficulté à identifier sa fonction dans la phrase d'autre part. Cependant, les textes arabes modernes n'incluent pas des signes diacritiques, donc, un mot arabe peut se rapporter deux ou plus différentes significations selon le contexte qu'il apparaît dedans.[9]
- **Variantes d'orthographe :** en script arabe, le mot peut être orthographié différemment et se réfère toujours au même mot avec le même sens. **Par exemple** : le mot **« جرام», jrAm**1, « **gramme** », peut également être écrit comme **« غرام »,** gramme, avec la même signification.[9]

1.6 Approches d'extraction d'entités nommées

La reconnaissance des entités nommées est une tâche **NLP** essentielle qui nous permet de détecter les références à des personnes, lieux, entreprises, dates, etc. qui sont contenues dans un texte.

Les approches les plus courantes d'extraction des entités du texte sont les suivants :

1.6.1 Approche symbolique

Appelée aussi approche linguistique ou approche à base de règles, elle est utilisée par la majorité des systèmes de reconnaissance d'entités nommées. Son principe de base est d'utiliser des connaissances linguistiques pour établir une liste de règles d'annotation. Ces règles sont écrites manuellement par des experts du domaine elles portent soit sur les constituants de l'entité nommée, soit sur leur contexte. Elles peuvent être de natures différentes : syntaxiques, informations lexicales, morphologiques ou encore sémantiques.[10]

1.6.2 Approche statistique

Contrairement aux approches symboliques qui reposent sur l'intuition humaine, l'approche statistique appelée aussi approche par apprentissage, utilise des processus automatiques pour l'extraction d'information. Son principe est de mettre au point, d'une manière automatique, des modèles d'analyse à partir de masses importantes de données. [10]
Cependant il y a trois types d'apprentissage automatique : supervisé, semi supervise et non supervisé. Voir figure 3 :

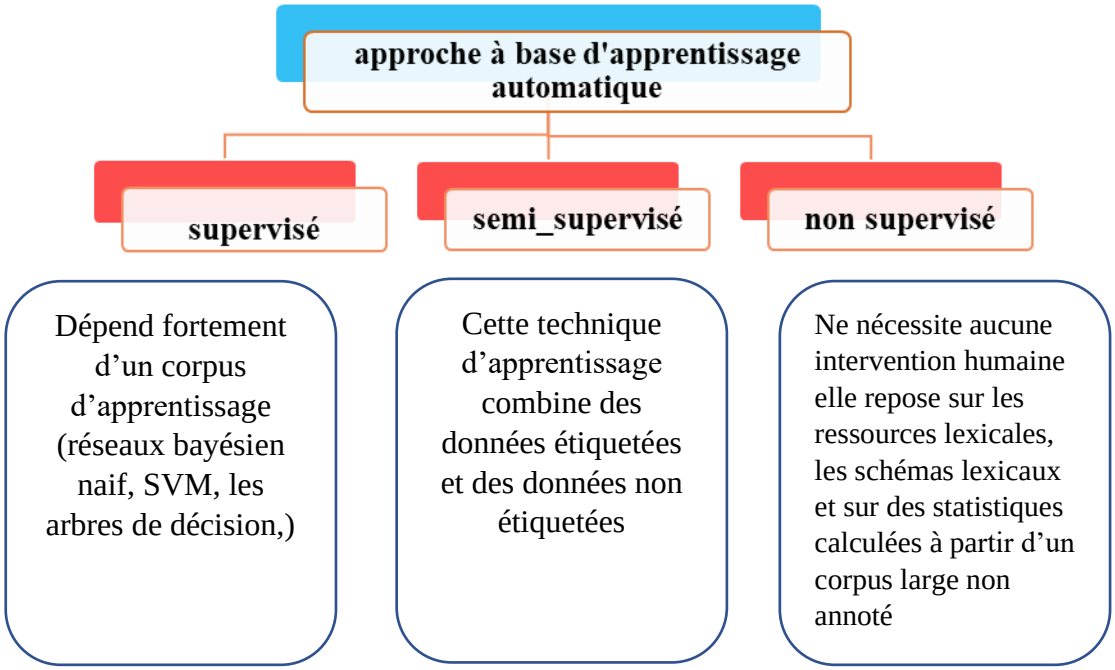

Figure 3: les types des approches statistiques pour l'EN [8]

1.6.3 Approche hybride

L'approche hybride consiste à combiner 1' approche à base de règles et 1' approche d'apprentissage pour l'extraction des **EN**. Cette combinaison permet de produire un système idéal qui profite des avantages de l'utilisation des deux approches : symbolique et statistique **(Zribi et al, 2010)** [11] Dans cette approche, les règles sont généralement apprises automatiquement mais elles doivent être révisées par un expert **(Poibeau, 2001)** [12] **; (Mansouri et al, 2008).** [13]

1.7 Travaux connexes

1.7.1 Systèmes NER basés sur des règles

- **Shaalan et Raza (2009) [14]** ont développé un système de reconnaissance des **EN** arabes **(NERA)** en utilisant une approche fondée sur des règles. Ce système repose sur l'utilisation d'un ensemble de dictionnaires d'**EN** et sur une grammaire sous forme d'expressions régulières pour la reconnaissance des **EN**. Le meilleur taux F-mesure acquis par ce système est de 98.6%.
- **Zaghouani et al (2010) [15]** Suivant le même principe ont présenté un module de repérage des EN à base de règles pour la langue arabe, la seule différence est qu'ils ont

procédé à une première étape de prétraitement lexical qui prépare le texte pour son analyse linguistique, ce module a été évalué sur un corpus de presse. La valeur de F-mesure apportée par ce module est de 47.35% pour le type organisation et 95.10% pour le type date.

- **Al-Ahmari et Al-Johar (2016) [16]** ont introduit une approche basée sur des règles pour la reconnaissance des **EN**, qui inclut les noms de personnes, de lieux et d'organisations en texte arabe. Le système fonctionne sur la base d'un ensemble de règles grammaticales indépendantes du domaine ainsi que d'un tagueur de partie arabe du langage en plus des nomenclatures et des listes de mots déclencheurs. Ce système comprend trois étapes principales : prétraitement, (catégorisation morphosyntaxique) POS tagging et algorithme de reconnaissance. La méthode a été appliquée à deux corpus arabes de domaines différents. La précision du système a été mesurée en termes de précision comme suit pour la personne 80,7%, pour l'emplacement 93,2% et pour l'organisation 75,4%.
- **(Friburger, 2002) [17]** Le système d'extraction linguistique des noms propres en français baptisé ExtractNP.Ce système est fondé sur les cascades des transducteurs. Grâce à celles-ci de multiples transformations sont réalisés sur les textes soit au niveau de l'analyse syntaxique, soit au niveau de l'extraction de l'information. ExtractNP utilise le système CasSys pour la génération de deux cascades de transducteurs d'extraction des noms propres : la première cascade a pour objectif de catégoriser et extraire une partie des noms propres d'un texte c'est-à-dire ceux qui possèdent des indices permettant de les repérer. La seconde cascade exploite les résultats de la première cascade afin d'en trouver de nouveau. Cette dernière apporte une bonne amélioration pour les résultats d'extraction. Dans le système ExtractNP de nombreuses ressources linguistiques ont été utilisées pour l'extraction. Pour le repérage des organisations, l'auteur a utilisé des preuves externes et internes. Pour l'extraction des noms des lieux, elle tend à extraire les preuves internes d'un côté et d'utiliser des dictionnaires sans preuves de l'autre côté. Pour la localisation des noms de personne, elle a opté pour l'utilisation des preuves internes et externes outre l'exploit de la morphologie des prénoms et des patronymes. Le système ExtractNP fournit actuellement de bons résultats sur la langue française. En effet, il obtient un rappel de 93% et une précision de 94%.
- **Abuleil (2004) [18]** a développé un système à base de règles pour extraire les noms de personne à partir de systèmes de Question/Réponse. ll a déterminé un ensemble de

règles en se basant sur 1' étude des relations entre les mots dans la phrase. Il a utilisé un ensemble de mots-clés et de verbes spéciaux qui ont été collectés dans un autre projet présenté dans **(Abuleil et Evens, 2002) [19]**, La précision de la reconnaissance **EN** a été estimée en termes de précision par l'auteur ; Personnes (90%), emplacement (93%) et organisation (92%).

- **(Fourour, 2002) [20]** propose le système Nemesis est qui permet la délimitation et la catégorisation des entités nommées développé pour le français et pour du texte bien formé (c'est-à-dire qui respecte les règles du français écrit). Il se base essentiellement sur les indices internes et externes définis par McDonald (1996) [5]. L'architecture de Nemesis se compose principalement de trois modules qui s'exécutent séquentiellement : prétraitement lexical, projection des lexiques et application des règles.
  Le taux de rappel de ce système 84,5 % et le taux de précision 89 %.

1.7.2 Systèmes NER basés sur approche apprentissage machine

- **Benajiba et al (2007) [21]** ont basée sur une approche d'apprentissage machine premièrement ont construit le corpus ANERcorp et les gazetteers [1]ANERgazet pour développer le système ANERsys. ANERsys est un système d'extraction des EN pour la langue arabe qui est basé sur l'algorithme d'apprentissage statistique d'entropie maximale. Le corpus d'apprentissage automatique du système ANERsys est de 125 000 mots et le corpus de test est de 25 000 mots. Les résultats obtenus par ce système ont donné un rappel de 37.51%, une précision de 51.39% et une f_mesure de 43.36%.
- Ensuite, le système ANERsys a été amélioré à ANERsys 2.0 **(Benajiba et Rosso, 2007) [22]** L'amélioration a été faite pour reconnaître les noms propres longs en combinant l'approche du maximum d'entropie avec l'étiquetage morphosyntaxique. ANERsys 2.0 a donné des résultats améliorés par rapport à 1' ancienne version ANERsys. Le rappel a été amélioré à 49,04 %, la précision a été améliorée à 63,21% et la F-mesure a été améliorée à 55,23 %. Pour améliorer encore la précision de ANERsys, (**Benajiba et Rosso 2008) [23]** ont utilisé un autre modèle probabiliste qui est les champs markoviens conditionnels (CMC) **(Lafferty et al., 2001) [24]**. Ils ont ajouté aussi la segmentation (en anglais, tokenisation) des données, ce qui amène à de meilleurs résultats.

[1] Un gazetteer est une liste d'EN de différentes types (Mikheev et al., 1999).

- **Mohammed et Nazlia (2012) [25]** qui ont développé un système d'extraction des **EN** en arabe avec l'utilisation des réseaux de neurones. Premièrement, ils ont prétraité le texte en entrée qui est en langue arabe. Ensuite, ils ont converti les phrases de ce texte en caractères romains, puis ils ont classifié les types de mots en utilisant les réseaux de neurones. Les réseaux de neurones consistent à apprendre la reconnaissance automatique des types d'EN et à prendre des décisions intelligentes basées sur les données disponibles. Le système de **Mohammed et Nazlia** a obtenu une F-mesure de 69.90% pour les EN-PERS, 43.30% pour les EN-LOC, et 59.20% pour les EN-ORG.
- **Gahbiche-Braham et al (2012) [26]** et **(2014) [27]** ont développé un système d'extraction des **EN** basé sur l'apprentissage supervisé en utilisant les algorithmes d'apprentissage statistiques CMC avec l'outil Wapiti **(Lavergne et al, 2010).[28]** Ensuite, ils ont adapté leur système à un apprentissage non supervisé (autoapprentissage), et les résultats ont été améliorés après l'adaptation.

1.7.3 Systèmes NER basés sur approche hybride

- **Mansouri et al (2008) [13]** ont présenté une étude comparative entre les trois approches : à base de règles, à base d'apprentissage machine et hybride. Cette étude a montré que l'approche hybride donne de bons résultats par rapport aux deux autres approches.

- **Zribi et al (2010) [11]** présente un système qui est composé de deux phases : la phase d'analyse morphologique du texte en arabe et la phase d'extraction automatique de règles pour détecter les **EN** selon leur type. Ce système a été combiné avec 1' algorithme d'apprentissage des règles RIPPER, qui utilise un ensemble d'attributs représentant les éléments les plus influents sur le résultat d'apprentissage. **Zribi et al**. Ont choisi d'utiliser deux types d'attributs pour l'extraction des règles : attributs morphologiques et attributs à base de lexique de noms propres.
- **Oudah et Shaalan (2013) [29]** présente un système basé sur l'approche hybride. Ce système est formé de deux composants. Le premier composant est à base de règles avec l'utilisation de la plateforme **GATE**[2]. Le deuxième composant est à base d'apprentissage machine en utilisant trois techniques

[2] GATE est disponible dans le lien http://gate.ac.uk/

d'apprentissage qui sont : l'arbre de décision, la MVS et la régression logistique. Le système développé a été testé sur le corpus ANERcorp de **(Benajiba et al 2007) [21]** et a donné un rappel de 94,9%, une précision de 94.2% et une F-mesure de 94.5%.

Le tableau ci-dessous présente la synthèse des travaux dédiés à l'extraction d'EN dans le texte. Pour chacun d'entre eux, le type d'EN ciblé et l'approche d'extraction mise en œuvre est spécifiée à propos les différents approches (à base des règles, à base d'apprentissage automatique et l'approche hybride).

Méthodes	**Références**	**Types-EN**	**Résultat**
Approche Basés sur Des Règles	Shaalan et razza (2009)	Pers Org, loc	F_mesure :98.6%
	Zaghouani et al (2010)	Org	F_mesure :47.35%
		Date	F_mesure :95.10%
	Al-Ahmari et Al-Johar (2016)	Pers	Précision : 80,7%
		Org	Précision : 75,4%
		Loc	Précision : 93,2%
	(Fourour, 2002)		Rappel :84,5 % Précision : 89 %
	(Friburger, 2002)	Pers Loc	Rappel :93% Précision : 94%.
	Abuleil (2004)	Pers	Précision : 90%
		Loc	Précision : 93%
		Org	Précision : 92%
Approche basée sur apprentissage	Benajiba et al (2007)	Pers Org Loc	Rappel : 37.51% Précision : 51.39% F_mesure : 43.36%

machine	Benajiba et Rosso, (2007)	Pers Org Loc	Rappel : 49.4% Précision : 63.21% F_mesure : 55.23%
	Mohammed et Nazlia (2012)	Pers	F_mesure : 69.90%
		Loc	F_mesure : 43.30%
		Org	F_mesure : 59.20%
	Gahbiche-Braham et al(2012) , (2014)	Pers loc Org	
Approche hybride	Zribi et al (2010)	Pers loc Org	
	Oudah et Shaalan(2013)	Pers loc Org	Rappel : 94,9% Précision : 94.2% F_mesure : 94.5%.
	Mansouri et al (2008)	Pers loc Org,	

Tableau 1: Synthèse des travaux existants sur l'extraction des entités nommées

Discussion

D'après une étude comparative entres les différentes approches utilisées pour EN nous avons remarqué que :

L'approche fondée sur des règles employée avec une grande expertise linguistique a conduit à une mise en œuvre réussie du système NER en surmontant les défis posés par la langue étudié L'approche par apprentissage machine présente l'avantage d'être plus flexible et plus robuste sur les corpus difficiles ou (bruités). Cette approche nécessite des corpus d'entrainement annotés qui ne sont pas toujours disponibles. D'autre part la construction des corpus annotés pour un nouveau domaine est une tache longue et nécessite des efforts de la part des experts humains pour la produire.

Enfin, au-delà de ces deux types d'approches, il existe une troisième voie qui consiste à combiner l'approche symbolique et l'approche statistique en une approche qualifié mixte ou hybride. Cette dernière, rendue possible grâce à la maturité acquise par les deux autres, est sans doute la plus prometteuse.

1.8 Métriques d'évaluation des entités nommées

Le rappel, la précision et la F-mesure sont des mesures largement utilisées dans les évaluations en TALN. La précision est le pourcentage des résultats corrects parmi les résultats obtenus. Le rappel est le pourcentage des résultats corrects parmi les résultats qu'on doit trouver. La F-mesure est la combinaison de la précision et du rappel et leur pondération.[30]

$$Rappel = \frac{Nombre\ d'EN correctement\ reconnus}{Nombre\ d'EN\ dans\ le\ corpus} \quad (1)$$

$$Précision = \frac{Nombre\ d'EN correctement\ reconnus}{Nombre\ d'EN\ reconnues} \quad (2)$$

$$F\_mesure = \frac{2(précision * rappel)}{précision + rappel} \quad (3)$$

Une autre façon de voir la précision et le rappel est :

$$Rappel = \frac{TP}{TP+FN} \quad (4)$$

$$Précision = \frac{TP}{TP+FP} \quad (5)$$

Alors que :

FN (false negatives) : sont des documents pertinents que votre système a manqués ne les ont pas retournés à l'utilisateur.

TP (true positives) : il s'agit des informations qu'on a besoin et le système les a également trouvés, donc c'est bien.

FP (false positives) : sont les documents que vous n'auriez pas dû retourner leurs non pertinents mais votre système les inclus.

1.9 Conclusion

La reconnaissance des entités nommées **(NER)** apparaît comme une composante essentielle dans plusieurs domaines du Traitement Automatique des Langues Naturelles **(TALN)** : analyse syntaxique, résolution de coréférence, traduction automatique, recherche d'information, etc.

Ce chapitre récapitule les travaux existants sur la reconnaissance des entités nommées et liste les différents outils relatifs aux domaines de la presse en ligne.

Nous présentons dans le prochain chapitre l'analyse d'opinion dans l'actualité ainsi que les approches et les techniques adoptées dans les travaux de recherche réalisés dans ce contexte.

2 Méthodes d'analyse d'opinion dans l'actualité

2.1 Introduction

Il est d'une importance considérable de connaitre l'avis des autres, c'est l'un des aspects indispensables de tout processus de prise de décision (politique, financière et économique). Avec la disponibilité des ressources riches en subjectivité, ce besoin est plus que jamais satisfaisable. Dans ce contexte l'analyse de tonalité a pris une grande ampleur depuis déjà une dizaine d'année.

L'analyse de tonalité dans l'actualité intervient pour résoudre cette problématique : analyser le contenu diffusé par les médias afin de détecter l'orientation de l'information envers certaines entités cibles et anticiper ainsi les citations de crise. Une difficulté majeure est la nature du texte à analyser, le texte diffusé par certaines sources médiatiques est à prédominance ou à apparence objective contrairement aux réseaux sociaux.

Dans le présent chapitre, tout d'abord, dans la première partie « généralité et concepts », nous commençons par donner quelques définitions dans le cadre de notre travail, nous abordons ensuite les domaines d'application d'analyse de sentiment, nous terminons par présenter les difficultés et les défis rencontrées lors de l'analyse de sentiment.

Dans la deuxième partie « analyse de sentiment », nous entamons les différents niveaux d'analyse d'opinion par la suite, nous discutons les trois grandes approches liées à la détection d'opinion et de polarité, enfin nous citons les différents travaux existants dans ce domaine.

2.2 Généralités et concepts

2.2.1 Définitions et terminologies

L'analyse d'opinion est un domaine récent qui a comme objectif de pouvoir reconnaitre l'orientation d'un texte à travers les mots utilisés. Dans la littérature, plusieurs vocables sont utilisés pour faire référence à ce même domaine entre autres : l'analyse de subjectivité, de tonalité, de sentiment ou d'opinion. Ces termes qui sont étroitement liés représentent les concepts clés de notre étude c'est pour cette raison que nous détaillons, dans cette partie, la définition de chaque terme pour ensuite sortir avec une convention sur la terminologie à utiliser dans le reste du rapport.

2.2.1.1 Subjectivité

Le premier qui a introduit la notion de subjectivité dans le langage était ***Emile BENVENISTE*** [31] Il définit la **« subjectivité »** comme la capacité du locuteur à se poser en

tant que *sujet* dans son énoncé. Selon ***BENVENISTE*** l'homme se constitue en sujet à travers le langage qu'il utilise. Cette définition affirme que les marqueurs de subjectivité de l'auteur sont identifiables et analysables à partir de son discours.

Une définition plus générale, donnée par **(Finegan, 1995)** [32], considère la subjectivité comme une expression de soi et une représentation du point de vue du locuteur. Le locuteur grâce au langage choisi peut s'approprier une identité unique que *FINEGAN* qualifie par l'empreinte du locuteur.

Pour **Wiebe et al** [33], pouvoir distinguer les expressions subjectives des expressions objectives est très important en traitement automatique de langage naturel, vu l'effet qu'a l'attitude de l'auteur sur la pertinence de ses documents. Ils commencent par fixer une définition de la subjectivité qui est une expression linguistique des opinions, sentiments, émotions, évaluations, croyances et spéculations de quelqu'un.

2.2.1.2 Tonalité

Dans la littérature ***ton*** ou ***tonalité*** d'un texte est l'attitude de l'auteur envers le sujet écrit. Elle est exprimée à travers le choix des mots, ou l'opinion prononcée nettement. La tonalité d'un texte a le même effet d'une voix sur le lecteur.

De point de vue TAL, le terme « tonalité » est utilisé dans de nombreux travaux pour faire référence à l'orientation sémantique ou la polarité qui consiste à la classification d'un mot, d'une phrase ou d'un document comme étant positive, négative ou neutre. [34]

Deux autres termes s'imposent dans l'ensemble des concepts clés : *sentiment* et *opinion*. Ces deux derniers reviennent souvent dans le domaine du Traitement Automatique de Langage en tant que synonymes. D'après **Bing** [35] « l'opinion » et les « sentiments » sont interchangeables car ils décrivent le même concept, il utilise cette parité pour définir l'opinion comme une expression subjective décrivant les sentiments d'un individu.

Pang and Lee [36] de leurs côtés distinguent entre les deux termes. Pour eux l'appellation analyse de sentiment est utilisée dans le domaine de traitement automatique de langage naturel alors que l'opinion Mining est favorisé dans la recherche d'information.

2.2.1.3 Opinion

La communauté de l'Opinion Mining catégorise les textes issus de toute source possible en deux grands types : les textes qui rapportent des *faits* et ceux qui expriment des *opinions*. Un

fait est une information avérée et jugée comme vraie alors que l'opinion [3]est une croyance individuelle liée aux convictions de la personne qui l'a exprimé. Pour garder trace de cette liaison et rajouter d'autres dimensions importantes, l'opinion sera représentée en n-uplets où *n* est le nombre de ses propriétés.

La fouille d'opinion ou l'opinion Mining est fondée sur la détermination de ces n-uplets. Cette représentation est la formalisation du problème d'opinion Mining dans un contenu textuel pour permettre par la suite une automatisation partielle ou totale de son processus.

Kim et Hovy [37] proposent une représentation en quadruplet :

Opinion= [Topic, Holder, Claim, Sentiment]

*« The **Holder** believes a **Claim** about the **Topic** and in many cases associates a **Sentiment** »*

- **Topic/Cible :** aussi appelé Target fait référence à l'objet ou le sujet de l'expression,
- **Holder/Porteur d'opinion :** est la personne qui exprime son opinion,
- **Claim/Revendication :** la croyance du titulaire de l'opinion, si le sujet et bon ou mauvais.
- **Sentiment :** le jugement du titulaire de l'opinion sur le sujet.

Exemple : « L'attaque des Etats Unis contre l'Irak est une erreur »

*Opinion= [**Topic :** L'attaque des Etats Unis, **Holder :** l'Auteur, **Claim :** Est une erreur, **Sentiment :** Négatif]*

Kobayashi et al [38] ont conservé le même nombre d'attributs (quatre) mais avec des éléments différents.

Opinion= [Opinion holder, Subject, Aspect, Evaluation]

- Opinion holder : la personne qui fait l'évaluation.
- Subject : une entité nommée.
- Aspect : une partie ou un attribut de l'entité et sur lequel porte l'évaluation.
- Evaluation : une expression subjective qui exprime l'attitude de *l'opinion holder* sur l'aspect.

L'attribut Aspect exprime qu'une opinion peut concerner certaines caractéristiques d'un objet et non pas l'objet en entier et Evaluation combine entre revendication et sentiment utilisés par Kim et Hovy. [37]

[3] Jugement personnel, manière de penser la plus répandue au sein d'une collectivité. Source : Hachette 2013

Exemple : « Les citoyens d'Alger se plaignent des coupures répétées de la connexion Internet »

*Opinion = [**Opinion Holder :** les citoyens d'Alger, **Subject :** Connexion Internet, **Aspect :** coupures, **Evaluation :** Se plaignent]*

a) Différents types d'opinion

Il existe deux types d'opinions le premier appelé opinion régulière [39]. Un autre type est appelé opinion comparative [40]. En fait, nous pouvons également classer les opinions en fonction de la façon dont elles sont exprimées dans le texte, l'opinion explicite et l'opinion implicite.[41]

- Opinion régulière

Est souvent désignée simplement comme une opinion dans la littérature et elle a deux sous-types principaux **:**

- Opinion directe

Fait référence à une opinion exprimée directement sur une entité ou un aspect d'entité.

Exemple : **« le magazine sportif français 'so_foot ' déclare que la sélection algérienne de football est la meilleure sélection du monde en 2019. »**

- Opinion indirecte

Est une opinion exprimée indirectement sur une entité ou un aspect d'une entité en fonction de ses effets sur quelques autres entités. Ce sous-type se produit souvent dans le domaine médical.

Exemple : « Pseudo artisans à Guelma : Des factures salées pour des travaux bâclés »

Une grande partie de la recherche actuelle se concentre sur les opinions directes car ils sont plus simples à manipuler. Les opinions indirectes sont souvent plus difficiles à traiter.

- Opinion comparative

Une opinion comparative exprime une relation de Similitudes ou différences entre deux ou plusieurs entités et / ou une préférence du détenteur de l'opinion sur la base de certains aspects communs des entités.

Exemple « les manifestants ont réclamé la libération des détenus du Hirak et un changement radical du système. Certains ont réitéré leur rejet des résultats de la présidentielle, d'autres ont posé des conditions pour le dialogue ».

Une opinion comparative est généralement exprimée en utilisant la forme comparative ou superlative d'un adjectif ou d'un adverbe, mais pas toujours (par exemple, préférez).

- Opinion explicite

Est une déclaration subjective qui donne une opinion régulière ou comparative.

Exemple : « Djamel Belmadi s'est adjugé le titre de meilleur technicien en charge d'équipe nationale ».

- Opinion implicite

Est une déclaration objective qui implique une opinion régulière ou comparative. Une telle déclaration objective exprime généralement un fait souhaitable ou indésirable.

Exemple : « Le pouvoir algérien est-il prêt à négocier sa propre mort ?»

b) Les classes d'opinion

Les méthodes classiques d'analyse d'opinion considèrent les trois classes de polarités standards :

- Positive,
- Négative,
- Neutre.

Selon **Liu [41]** les classes doivent inclure l'intensité de l'opinion et distinguer les types de ces jugements, il propose une représentation en cinq classes qui est largement adoptée :

- Positif émotionnel.
- Positif rationnel.
- Neutre.
- Négatif rationnel.
- Négatif émotionnel.

2.2.1.4 Sentiment

Les sentiments [4]sont parfois notés en tant qu'une composante de l'opinion comme par exemple dans le modèle de **Kim et Hovy** [37]. Dans d'autres travaux chaque terme est

[4] Tendance affective liée à des émotions, des représentations et des sensations. Source : Hachette 2013

interprété à part. Un sentiment est un jugement personnel exprimé envers une entité. Il est caractérisé par sa polarité et son intensité :

Sentiment = [Polarity, Intensity]

- **Polarity/ Polarité :** l'orientation du jugement s'il est positif, négatif ou parfois neutre,
- **Intensity/ Intensité :** le degré de positivité ou de négativité.

Pak [42] présente ce modèle sous forme de vecteur pour illustrer cette définition (*Figure 4*), l'orientation du vecteur présente la polarité du jugement et sa langueur est l'intensité.

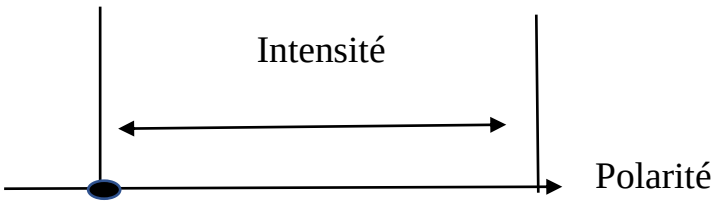

Figure 4:modèle "sentiment analysis comme vecteur" de Pak Alexander,2012 [42]

2.2.1.5 Terminologie adoptée

- ✓ Le sentiment est la composante affective de l'opinion. [37]
- ✓ La subjectivité est une expression linguistique utilisée pour exprimer les opinions de quelqu'un. [33]
- ✓ La tonalité est l'effet produit par le texte dans l'esprit de son lecteur. C'est un terme plus global. Elle est présente dans toute portion de texte : ("What is Tone in Literature?,"). Cette perception est modélisée dans la (*Figure 5).*

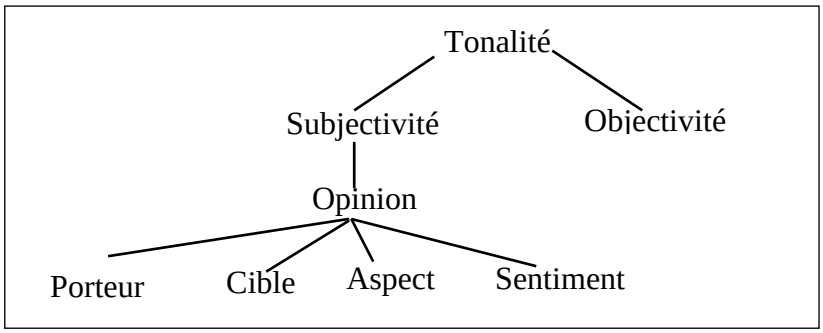

Figure 5:Relation entre les terminologies [33]

Il est à noter que jusqu'à présent il n'y a eu aucune convention sur la terminologie. Les travaux de recherches adoptent la terminologie qu'ils préfèrent. Dans la littérature technique et en particulier dans les domaines « *Sentiment Analysis* » et « *Opinion Mining* » les termes

(opinion, sentiment et tonalité) couvrent le même champ d'étude, ils sont utilisés de manière interchangeable et n'altère en rien le processus de l'analyse automatique qui est l'un des objectifs de ce travail.

2.2.2 Domaine d'application

- **Politique :** Grâce à l'analyse des sentiments, les décideurs de politique pouvant prendre l'avis des citoyens sur certaines politiques, afin de bénéficier de cette information pour améliorer ou créer une nouvelle politique qui convient avec les citoyens.[43]
- **Domaine médical :** Analyse opinion des médicine, patient sur les médicaments et les services hospitalier. Ainsi sur les documents de l'état de patient qui contient le diagnostic et la description du résultat d'examen. [43]
- **Domain éducation :** Développer le niveau d'enseignement à travers l'analyse et l'interprétation de l'opinion de l'étudiant à travers les méthodes d'enseignement est ça se permettre améliorer l'enseignement et l'apprentissage. [43]
- **Marketing :** Du côté entreprises, permet au fournisseur plus de connaissances à propos des besoins des consommateurs, du côté client il peut donner son opinion, s'inspirer des opinions d'autres clients pour l'aider à sa décision et aussi comparer les produits avant de les acquérir. [44]
- **Veille :** La veille est une pratique indispensable pour toute organisation (que ce soit une institution publique, entreprise, une personnalité politique ou une célébrité) qui exerce son activité dans un environnement et veut mesurer l'impact des efforts effectués. L'analyse de tonalité permet de suivre les traces d'apparition de ces organisations dans les différentes plateformes à grande audience (TV, Radio, Réseaux sociaux, journaux, blogs, …). Il s'agit plus exactement de veille de réputatio ou la surveillance de l'image publique. La synthétisation du grand nombre de mentions quotidiennement va permettre aux entités concernées de maitriser davantage leur environnement évaluer leur stratégie. Citons l'exemple d'une personnalité politique : l'analyse de sentiments appliquée à la veille va faciliter la prédiction des résultats des élections, anticiper les crises politiques en cas de constat d'un débat d'opposition sur les réseaux sociaux.

2.2.3 Difficultés et les défis de l'analyse des sentiments

- Difficulté due au contexte on parle de la difficulté de coordination entre plusieurs parties d'une phrase pendant l'analyse syntaxique du texte. [45]

- Difficulté due à la présence d'une négation peut inverser la polarité de celle-ci alors même qu'elle contient un ou plusieurs termes négatifs. [46]
- Difficulté due à la structuration syntaxique et sémantique de la phrase et l'expression de l'opinion qu'elle véhicule. Avec l'utilisation de la conjonction **« mais »** on oppose deux parties d'une phrase.
- Difficulté due à l'ambiguïté de certains mots positifs ou négatifs selon les contextes et qui ne peut pas toujours être levée. [47]
- Difficulté due au vocabulaire qu'on utilise pour exprimer une opinion. Il diffère d'une personne à une autre, comme par exemple un anglo-saxon lorsqu'il exprime ses sentiments utilise des mots bien représentatifs de ce qu'il ressent contrairement aux personnes qui ne connaissent pas ou peu sa langue. [48]

2.3 Analyse d'opinion

L'analyse d'opinion est l'étude computationnelle et sémantique des parties de textes en fonction des opinions, des sentiments et des émotions exprimés dans le texte. Généralement l'expression « analyse des opinions » est utilisée pour désigner la tâche de classification automatique des unités de texte en fonction de leur polarité (positive, négative, neutre).

2.3.1 Différents niveaux d'analyse d'opinion

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

2.3.1.1 Niveau du document

Détermine l'opinion générale de l'ensemble du document. Cette analyse fonctionne bien pour des documents qui présentent un point de vue précis, mais moins pour des comparaisons car elle ne fera pas la différence entre les sujets abordés. [49]

2.3.1.2 Niveau de la phrase

Détermine l'opinion générale d'une phrase (positive, négative ou neutre). Cette analyse peut donner une mesure de la "neutralité" d'un texte par exemple pour analyser des entrées de Wikipédia. Les méthodes utilisées sont celle de l'analyse de subjectivité. [49]

2.3.1.3 Niveau d'entité et d'aspect

Appelé en anglais (Feature level), au lieu de déterminer les entités à analyser en fonction de critère structuraux (phrase, paragraphe, document) ces méthodes se basent sur l'analyse de corrélation entre l'opinion émise et la cible de cette opinion. **Par exemple, la phrase « Le sujet du cours me passionne mais le professeur est ennuyeux. »** présente deux

sentiments sur l'entité **« cours »** : le sujet qui est perçu comme positif et le professeur, qui est perçu comme négatif. Ce niveau d'analyse permet de différencier les aspects qui sont aimé ou non par les auteurs des textes et ainsi permet plus facilement de déterminer des remédiations possibles. En revanche il est très difficile à mettre en place car ce type d'analyse est extrêmement complexe. [50]

2.3.2 Les approches d'analyse d'opinion existants

Il existe trois types d'approches, une basée sur le lexique, la seconde sur l'apprentissage machine et la troisième hybride.

- Approches symboliques.
- Approches statistiques.
- Approches mixtes ou hybrides.

2.3.2.1 Approches symboliques (ou linguistiques)

Tend à utiliser des outils du TAL (analyseur syntaxique, analyseur morphologique, analyseur sémantique, etc.), et à définir un ensemble de patrons lexico-syntaxiques qui sont des règles décrivant une expression régulière, formée de mots et de catégories grammaticales. Ces patrons sont souvent porteurs d'un ou plusieurs marqueurs linguistiques. [51]

La principale tâche de cette approche est la conception de lexiques ou dictionnaires d'opinion. L'objectif de ces lexiques (dictionnaires d'opinion) est de répertorier le plus de mots porteurs d'opinion possible. Ces mots permettent ensuite de classer les textes en deux catégories (positif et négatif) ou trois (positif, négatif et neutre).

Cette méthode nécessite donc la construction d'un dictionnaire d'opinion. Pour construire un tel dictionnaire, trois genres de techniques sont possibles :

- Une constitution manuelle.
- Une constitution automatique en utilisant des corpus.
- Une constitution à base des dictionnaires existants.

2.3.2.1.1 La méthode manuelle

Cette méthode consiste à enrichir le lexique de mots d'opinions de façon manuelle (sans utiliser aucun outil particulier), par la création d'un ensemble de mot et expression

porteurs d'opinions. Cet ensemble de mot est appelé graine.il est utilisé par la suite afin de trouver d'autre mots et expressions porteurs d'opinions.

2.3.2.1.2 La méthode à base de corpus

Cette méthode consiste à utiliser une de conjonctions de coordination suivantes : **AND, OR, BUT**, **EITHER-OR, et NEITHER-NOR**, afin d'associé à chaque mot du lexique, un ensemble de scores d'opinion. Cette coordination doit être apparait entre un mot déjà classé et un mot non classé. Par exemple, si la conjonction **AND** sépare un mot classé positif dans le dictionnaire d'opinion et un mot non classé, alors le mot non classé sera considéré comme étant positif. À l'inverse, si la conjonction **BUT** sépare un mot classé positif et un mot non classé, alors le mot non classé sera considéré comme étant négatif. [52]

2.3.2.1.3 La méthode à base des dictionnaires

Cette méthode utilise des dictionnaires de synonymes et antonymes existants tels que WordNet. **WordNet** : est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton, Dans cette base les mots sont organisés sous forme d'arbres le principe de cette méthode est de détermine l'orientation sémantique du mot traité, s'il existe dans la liste des synonymes du dictionnaire il affecte la même polarité à ce nouveau mot. Dans wordNet les mots sont organisés sous forme d'arbre (*voir la figure 6*). Par exemple utilisent ces dictionnaires afin de prédire l'orientation sémantique des adjectifs. [45]

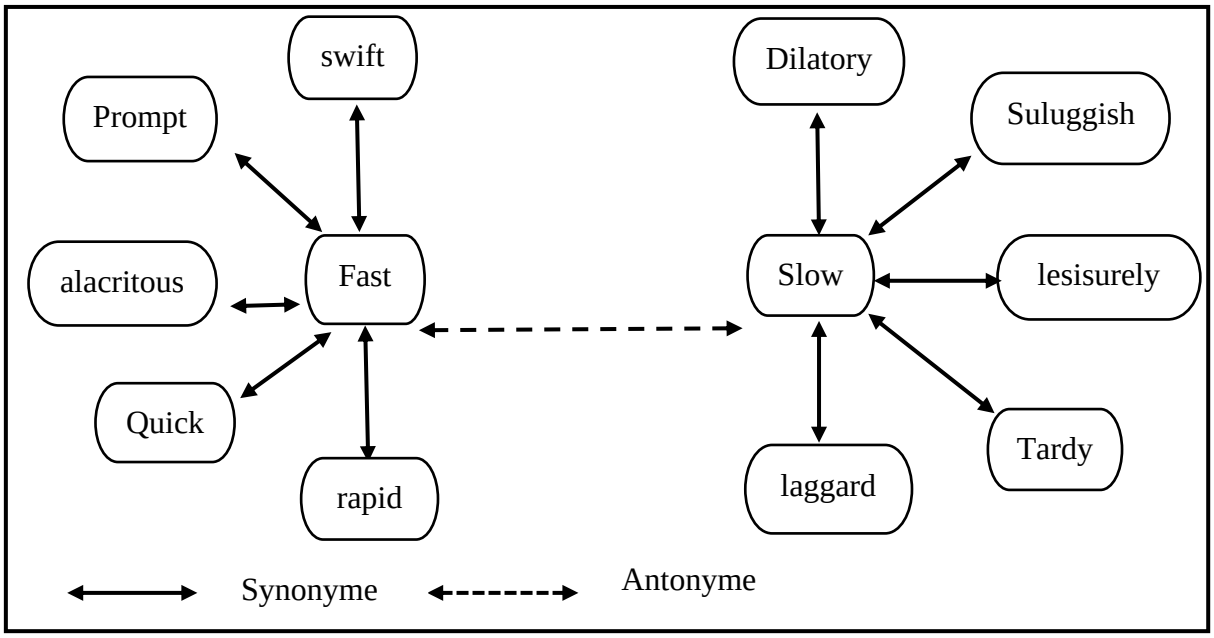

Figure 6:Exemple d'arbre de synonymes et d'antonymes présents dans WordNet [45]

2.3.2.2 Approches statistiques

Ces approches utilisent des données qui sont des phrases subjectives ou des documents avec opinion fournies au classifieur pour l'apprentissage, dans ce type d'approches l'aspect sémantique n'est pas pris en compte, les mots sont considérés généralement comme des variables équivalentes.

Les techniques existantes focalisent sur la classification et l'extraction des opinions dans un format libre (texte en langage naturel).

Ce qui concerne les méthodes d'apprentissage supervisé, il existe des approches s'intéressent à la classification de la polarité des opinions, certaines d'autre s'intéressent aussi à la détection de textes subjectifs.

Donc c'est une classification à deux axes :

- **Positif contre négatif.**
- **Subjectif contre objectif dans l'autre.**

L'objectif de la classification supervisée est de construire un modèle de classification à l'aide d'exemples (corpus d'apprentissage) cette procédure de classification permet de prédire l'appartenance d'un nouvel exemple à une classe. [52]

2.3.2.3 Approches hybrides

La combinaison de ces deux approches semble ; aujourd'hui ; prometteuse dans la mesure où les résultats fournis par les Machines à Support Vectoriel (SVM) donnent de meilleurs résultats que les approches prises individuellement. [51]

La méthode hybride c'est une combinaison de techniques des deux précédentes approches pour aboutir à des résultats très précis. Elles prennent en compte tout le traitement a linguistique puis lancer le processus d'apprentissage, les opinions sont extraites des phrases du document qui sont traitées phrase par phrase, ensuite une valeur globale est attribuée aux classes. Les résultats de cette approche sont plus performants que chacune des approches symboliques et statistiques employées séparément. On résume qu'il existe deux grandes catégories d'analyse des sentiments qui impliquent des méthodes de classification basées sur des algorithmes comme la figure suivante montre.

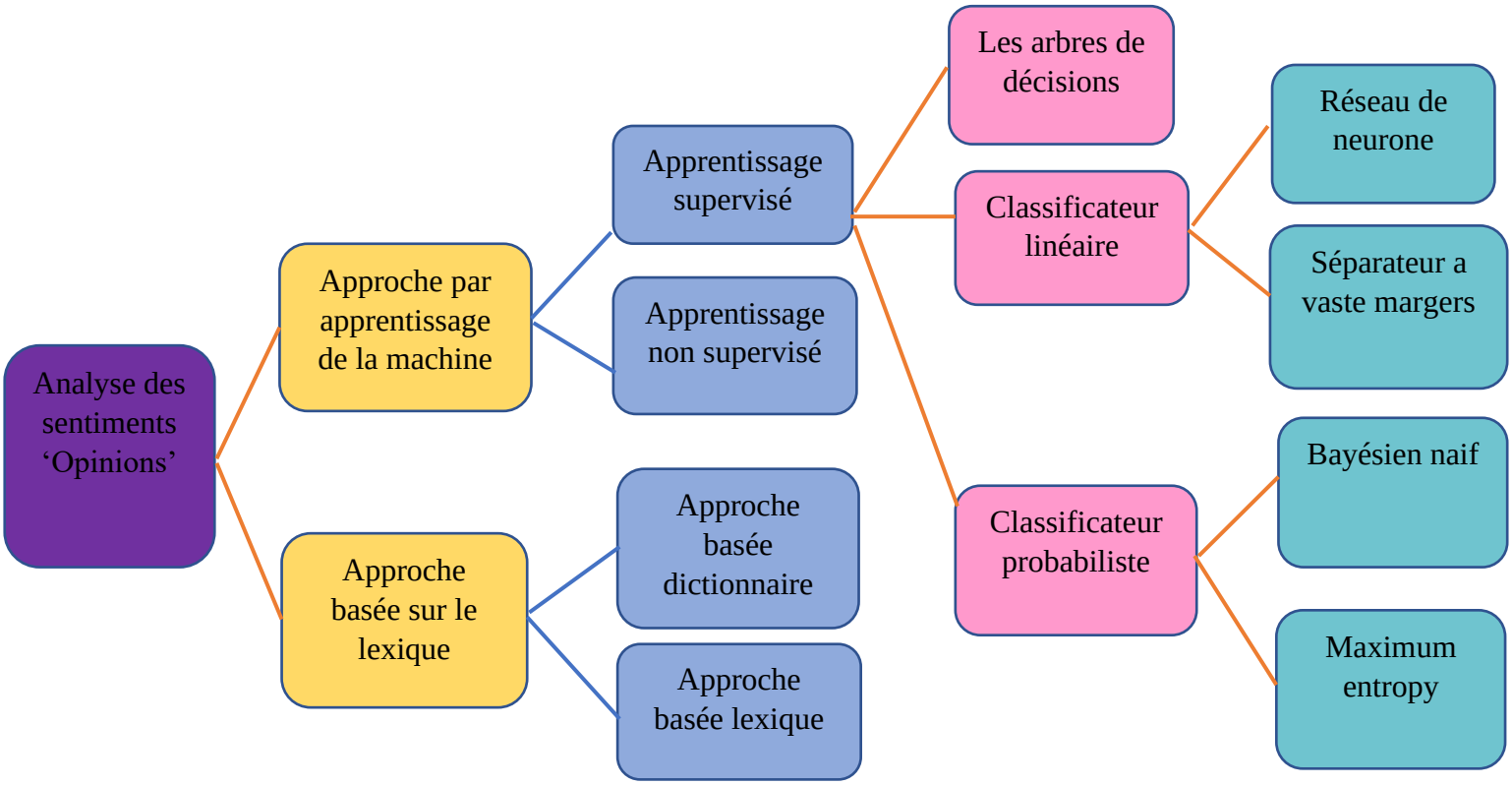

Figure 7:Techniques de classification des sentiments opinion [53]

2.3.3 Synthèse des travaux existants

A l'intérieur de l'analyse de tonalité dans les contenus médiatisés, les réseaux sociaux et les sites d'e-commerce reste un objet de recherche privilégié en termes de nombre de travaux de recherche qui se sont intéressés à l'étude de ces discours. Tandis que la question d'analyse de tonalité des discours de presse, radiophoniques est télévisuels, qui augmentent en intérêt auprès des économistes ces dernières années, n'a pas bénéficié de cet intérêt.

En fixant le thème : analyse de tonalité dans les plateformes d'actualité (articles de presse, blogs journaux en ligne) comme premier critère de sélection nous avons collecté un ensemble de publications. Certains de ces travaux ont été mené à utiliser l'approche lexicale avec des choix différents dans la génération de lexique d'autres se basaient entièrement sur des méthodes ML.

- **Ku et al [54]** ont opté pour une classification lexicale basée dictionnaire en justifiant leur choix par le fait que le classifieur lexical est le seul à pouvoir descendre jusqu'au niveau sémantique des mots et seule l'analyse des mots qui permet de déterminer leur orientation. Une base d'évaluation est préparée en classant des documents textuels, collectés en ligne, en quatre catégories d'opinion {positive, négative, neutre, sans opinion}. Ceci est assuré à

l'aide de trois personnes qui s'occupent du tri. En cas de désaccord entre ces trois ; l'étiquette de la classe qui a plus de vote sera attribuée au document. Si les trois étiquettes sont différentes le document en question est exclu.

- Le vocabulaire de sentiment est formé avec deux dictionnaires *General Inquire* et *Chinese Network Sentiment Dictionary* pour collecter les mots graines qui seront ensuite élargis avec *Academia Sinica Bilingual Ontological Wordnet.* Les auteurs calculent la tendance de polarité d'un mot Swi en prenant en considération son contexte d'apparition comme suit :

$$S_{w_i}=P_{w_i} - N_{w_i} \qquad (6)$$

Avec *Pwi* et *Nwi* indiquent le poids de wi comme étant un mot positif, respectivement négatif.

- La classification se fait en suivant l'hypothèse « *l'opinion de l'ensemble est une fonction de l'opinion de ses composants* ».

$$S_{p_j}= S_{opinion\ holder} \sum_{i=1,m} S_{w_i} \qquad (7)$$

$$S_{d_k}=\sum_{j=1,n} S_{p_j} \qquad (8)$$

Avec :

S_{p_j} Est la polarité de la phrase j.

$S_{opinion\ holder}$ Est le poids du détenteur de l'opinion.

S_{d_k} Est la polarité du document k.

La précision aboutie avec le classifieur proposé dans cet article est de 61.06% tandis qu'un classifieur basé sur l'algorithme SVM de l'apprentissage automatique n'a réalisé que 46,81%.

- L'un des travaux les plus référencés dans l'analyse de sentiments appliquée à l'actualité presse est celui de **Godbole et al [55]** Dans leur article, ils utilisent l'approche lexicale pour extraire les opinions des articles de journaux.

− La génération du vocabulaire de sentiment se fait séparément par domaine (général, santé, criminalité, sports, affaires, politique, médias) à l'aide du dictionnaire WordNet afin d'assurer un vocabulaire spécifique selon la catégorie thématique de l'article.

Dans le but de contourner les limites de l'approche basée dictionnaire citées ci-dessus, ils proposent une amélioration de l'algorithme d'enrichissement de l'ensemble des mots graines.

L'idée consiste à exécuter l'algorithme en deux itérations :

Dans la première ; la distance qui le sépare un mot donné (W) de son mot graine est calculée ; pour associer une valeur de signifiance aux chemins qui les relient, le score de W et la somme des valeurs de signifiance de chaque chemin, la formule de calcul du score est la suivante :

$$\text{Score (W)} = \sum {}^{1}/_{C^{d}} \qquad (9)$$

d : la distance entre W et le mot graine ; c : constante > 0.

La deuxième itération est ajoutée pour repérer et éliminer les mots qui changent de polarité entre positive et négative ou ceux qui sont ambigus. Pour ce faire, seuls les chemins dont le nombre d'alternance ou *flips* est inférieur à seuil fixé sont maintenus. Et puisque WordNet ordonne les mots selon le sens couramment utilisés la suppression des mots ambigus est faite en prenant seulement les X premiers mots cités ; avec X un entier à fixer.

− La classification de l'opinion dans une phrase est basée sur le lexique généré tout en tenant compte du contexte du mot pour calculer son score final de polarité.

Score(W) est inversé si W est précédé d'une négation, augmenté ou diminué si précédé d'un adverbe. **Exemple** : *score* (Intéressant) = 3 ; *score* (pas intéressant) = -3 ; *score* (très intéressant) = *score* (Intéressant) +1= 4 ; *score (peu* intéressant*)* = *score* (Intéressant) – 1 = 2.

- **Balahur et al [56]** synthétisent les résultats obtenus dans plusieurs expériences dans lesquelles ils ont testé la pertinence de quatre dictionnaires de sentiments. Grace à ces expériences ils ont abouti à divers résultats concernant les facteurs qui influent la classification des opinions. Comme première étape, la composition du fragment de texte à analyser est mise en examen ; les résultats ont montré qu'il est nécessaire, avant d'entamer L'analyse de polarité, de définir la source et la cible de chaque expression d'opinion, séparer entre une bonne/mauvaise *information* et une *opinion* positive/négative en éliminant les mots de la catégorie thématique (crise, catastrophe, dégâts, …) et surtout appliquer l'opération d'étiquetage des mots de sentiment (positif, négatif ou neutre) seulement aux opinions directes et explicites qui n'exigent ni des interprétations

personnelles ni des bases de connaissances antérieures dans un domaine. Avec tous ces critères réunis une précision de 81% est enregistrée.

La deuxième série d'expériences évaluent l'impact du lexique d'opinion et la longueur de l'expression d'opinion à analyser sur la précision. Malgré l'utilisation de dictionnaires réputés dans la génération du vocabulaire de sentiments tel que WordNet Affect [5]et SentiWordNet[6]. La valeur de la précision n'a pas dépassé 61% ce qui montre qu'avoir un lexique élargi et riche n'augmente pas nécessairement les performances du système. La plus haute valeur de précision (83%) est enregistrée lorsque le lexique est construit à partir d'une combinaison de deux sources externes : Micro WNOp et JRC Tonality. Ce dernier est un lexique pas très large et créé manuellement par les auteurs de l'article. De plus limiter le nombre de mots autour de la cible d'opinion à six mots de sentiment donne de meilleurs résultats que de prendre toute la citation.

Le score d'opinion d'une citation (phrase entre deux côtes) est égal à la somme des scores de mots de sentiments repérés.

$$\text{Score(citation)} = \sum_i Score(Wi) \qquad (10)$$

- **Europe Media Monitor (EMM)** est un projet de veille média multilingue lancé par le *Centre Commun de Recherche de la Commission Européenne* en 2002, Le système collecte, catégorise, résume et produit des statistiques d'en moyenne 300 000 articles de presse par jour rédigés en 70 langues. Dans l'intention de compléter les fonctionnalités d'EMM une étude est réalisée par **Steinberger et al [57]** qui propose une technique d'analyse de l'opinion exprimée dans les articles colletés. L'objectif est de connaitre l'orientation des avis envers certaines personnes ou organisations et la tendance des opinions selon les sources d'informations et les pays. La réalisation d'un tel système d'analyse de sentiments à grande échelle présente plusieurs difficultés. La plus complexe est le nombre important de langues à couvrir qui réduit le choix de techniques en obligeant le système d'adopter des méthodes faiblement liées à la langue. Dans ce contexte, l'approche lexicale basée dictionnaire est écartée étant donné que le vocabulaire proposé est en une seule langue et la traduction d'un aussi grand lexique en 70 langues consomme un temps énorme. La solution proposée pour éviter ce problème est de construire un

[5] http://wndomains.fbk.eu/wnaffect.html
[6] http://ontotext.fbk.eu/sentiwn.html

vocabulaire réduit composé des dix mots de sentiment les plus fréquemment utilisés. Cet ensemble de seulement dix mots couvre 12.3% de tous les mots de sentiments identifiés dans les articles presse.

La méthode consiste à identifier en premier lieu l'entité cible du jugement qui est dans la plupart des cas une entité nommée (repérée par les lettres en majuscule) en suite extraire les mots de sentiments en se limitant à un intervalle de six mots autour de la cible pour des soucis d'optimisation. Et en fin, ces mots sont comparés aux termes présents dans l'ensemble de dix éléments et ainsi, la polarité et attribué à la phrase.

Cette approche a atteint une précision de 74% en identifiant comme défis majeurs la dominance des articles à apparence objective (72%) et le fait d'employer souvent le style passif dans la rédaction ce qui crée une ambiguïté lors de l'association du mot de sentiment à l'entité la plus proche.

- **Le système *AzFinText* (Schumaker and Chen, 2012) [58] :** qui analyse la tonalité des articles de finance dans le but de modéliser son impact sur les prix des actions, implémente le module externe *OpinionFinder* **(Wilson et al., 2005a) [59]** pour extraire et catégoriser opinion exprimée dans les articles. *OpinionFinder* couvre tout le processus de classification partant du prétraitement : Tokenisation avec *OpenNLP*[7] et *Abney Stemmer* [8]pour la racinisation jusqu'à la classification des phrases en subjective /objective avec un classifieur *Naïve Bays*.

  Les chercheures estiment que le module *OpinionFinder*, pris à part, a pu atteindre une exactitude de 76% tandis que son intégration dans le système global *AzFinText* avec un contexte financier le système à réaliser une exactitude de 59.0%

- **Li et al [60]** opèrent une analyse de tonalité sur les articles de l'actualité financière afin d'étudier l'impact du sentiment exprimé par le rédacteur sur les décisions d'investissement prises par les lecteurs qui se traduisent en variation des prix des actions. L'annotation des données d'entrainement est semi-automatique à l'aide de deux dictionnaires : *Harvard IV-4* 25[9]*(HVD)* et *Loughran and McDonald (LMD) Financial Sentiment Dictionary* conçus manuellement. L'apprentissage et la classification sont effectués avec l'algorithme SVM.

[7] http://opennlp.apache.org/
[8] http://www.vinartus.net/spa/
[9] http://www.wjh.harvard.edu/~inquirer/

Nous présentons dans ce qui suit une synthèse des travaux les plus signifiants en termes de nombre de référencement et la date de publication *tableau 2* :

Référence	Méthode	Niveau d'analyse	Domaine	Résultats
(Wilson et al., 2005a)	Apprentissage automatique Naïve Bayes	Phrase	Multi_domaine	Exactitude = 76% Precision = 79% Rappel = 76% F-measure = 77.5%
(Ku et al., 2006)	Lexical blasé Dictionaries	Phrase	Article scientifique et de blogs sur le Colonage des animaux	Precision = 61.06%
(Godbole et al., 2007)	Lexicale basée Dictionnaire	Phrase	Articles de journaux et de blogs	–
(Schumaker and Chen, 2012)	Apprentissage automatique Naïve Bayes	Document	Articles de Finance	Exactitude = 59.0%
(Balahur et al., 2013)	Lexicale basée Dictionnaire	Phrase	Citations et discours rapportés dans les articles de presse	Precision = 83%
(Li et al., 2014)	Apprentissage automatique SVM	Document	Articles de Finance	Exactitude =65.05%
(Steinberger et al., 2017)	Lexicale basée Dictionnaire	Document	Articles de presse en ligne	Precision = 74%

Tableau 2: Synthèse des travaux existants sur la classification de tonalité

Discussion

D'après le tableau de synthèse sur les travaux existants dans le domaine d'analyse de tonalité qui s'intéresse à l'actualité presse et suite d'une comparaison entre les différentes approches présentées nous avons constaté que :

Ce qui se dégage des travaux basés sur l'approche lexicale est qu'ils travaillent avec la méthode dictionnaire tandis que le vocabulaire conçu à partir d'un corpus n'est pas vraiment apprécié.

Dans leur généralité, les meilleurs résultats, dans les travaux basés sur l'approche lexicale, sont obtenus quand l'analyse est effectuée au niveau phrase où une connaissance préalable de la cible d'opinion est nécessaire.

Pour ce qui est de l'approche basée apprentissage automatique, les résultats dépendent fortement des données à analyser : le pourcentage des passages subjectifs, la langue et le niveau de détail choisi. Nous avons remarqué que le classifieur Naïve Bayes utilisé avec une représentation TF-IDF construit une bonne combinaison.

La plupart des travaux ont privilégié la méthode basée sur le lexique de sentiment pour sa facilité, puisqu'elle ne demande pas des données annotées et les résultats obtenus qui sont plus au moins satisfaisants. Cependant, les inconvénients de cette approche, restent non-négligeables.

L'approche Machine Learning, fondée sur des algorithmes qui ont fait preuve d'efficacité dans plusieurs domaines, présente à son tour des limites entre autres la nécessité d'une annotation manuelle d'un nombre important de données d'entrainement, la difficulté de collectionner des documents « avec tonalité explicite » surtout dans le domaine du journalisme, sans oublier l'inconvénient majeur qui est l'analyse de la sémantique sans connaissance parfaite du vocabulaire.

De nombreux travaux de recherche se réalisent dans le but de perfectionner l'analyse de tonalité avec la méthode basée apprentissage automatique afin d'atteindre le même niveau de la classification thématique qui a réalisé de très bons résultats. Une approche hybride peut combler les limites des deux méthodes en accordant les points forts de chacune afin d'atteindre des meilleures performances.

2.4 Conclusion

Ce chapitre présente brièvement le domaine de l'analyse d'opinion, en commençant par la définition et terminologie adopté, les concepts utilisés tel que subjectivité, tonalité, opinion, et sentiment. Les facteurs qui rendent difficile la fouille d'opinion comme la dépendance du domaine et du contexte. Nous avons étudié les approches existantes.

Un état de l'art sur les méthodes existantes : les méthodes basées sur la construction de lexique d'opinions, les méthodes basées sur l'apprentissage automatique et les méthodes hybrides qui utilisent les deux précédentes.

L'identification d'opinions a connu beaucoup d'applications importantes : extraction et classification des critiques des clients sur un produit commercialisé en ligne, le suivie des opinions publiques sur un politicien par la fouille en ligne dans les forums et les blogs … etc.

Partie 2. Contribution

3 Conception

3.1 Introduction

Dans ce travail nous nous situons dans un contexte de veille média et nous nous intéressons au problème d'analyse d'opinions dans l'actualité. Dans ce chapitre, nous présentons dans un premier temps notre approche d'analyse d'opinion dans l'actualité à travers une architecture générale qui sera détaillée point par point.

Par la suite, nous proposons dans la conception l'application de l'analyse d'opinion dans l'actualité en utilisant le langage de modélisation UML. Nous présenterons trois types de diagrammes :

- Diagramme de cas d'utilisation (use case),
- Diagramme de séquence,
- Diagramme de classe.

3.2 Approche d'analyse d'opinion dans l'actualité

L'analyse d'opinion dans l'actualité c'est une tâche qui consiste à extraire les opinions exprimées pour un sujet précis. D'une autre façon C'est une classification d'opinion dont le but de déterminer si l'opinion est positive, négative ou neutre afin d'attribuer une étiquette au texte selon l'opinion qu'il exprime.

3.2.1 Architecture du système

L'architecture globale de notre système d'analyse d'opinion dans l'actualité en ligne, est représentée dans *la figure 8*.

Le processus de l'application de système s'étale sur plusieurs étapes, après la collecte des données à partir des différentes sources RSS, une série des prétraitements sera effectuer (la tokenization, suppression des mots vides et la lémmatisation), ensuite le traitement sur les données agrégées en commancons par l'identification des entités nommées suivie par l'extraction de toalité,et enfin visualiser les résultats aux utilisateurs .

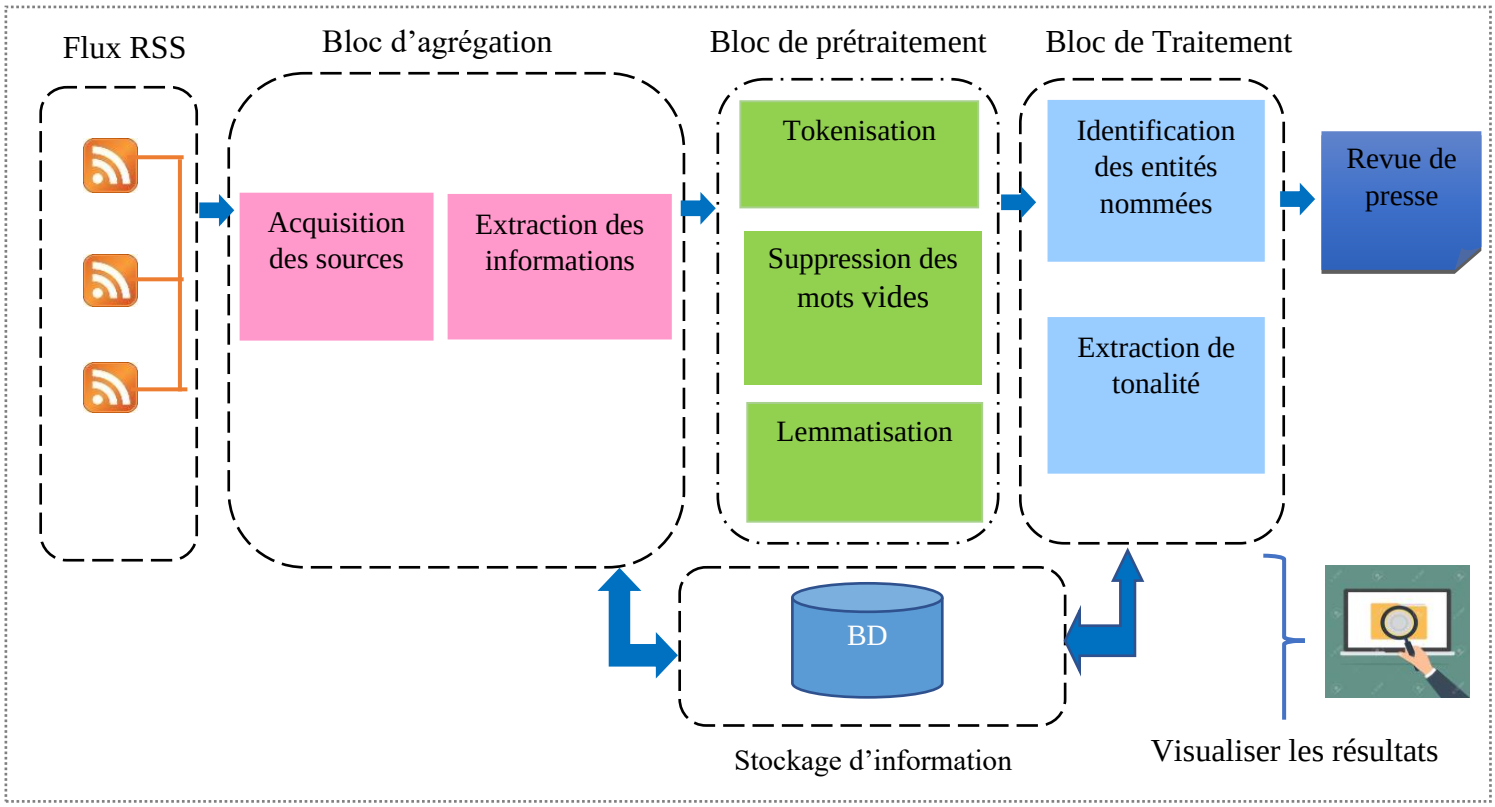

Figure 8 : Architecture général du système

3.2.2 Présentation générale du système

Notre système comprend les composants suivants :

3.2.2.1 Identification des sources RSS

Cette étape consiste à ajouter les flux RSS des sources pour le traitement. Par exemple pour la source <**APS**> il ajoute le flux : http://feeds.aps.dz/aps-algerie pour obtenir le lien, il faut aller au site de la source et cliquer gauche sur l'icône spéciale du flux, ensuite copier le lien comme illustre la figure suivante.

Figure 9 Les étapes d'obtention d'un flux RSS pour un site web

Les flux RSS sont des fichiers XML respectant un formalisme et contenant des éléments obligatoires comme le titre de l'information, lien vers une page web, description et des éléments optionnels comme le langage de flux, auteur, date de publication et date de la dernière modification, comme illustre la figure 10.

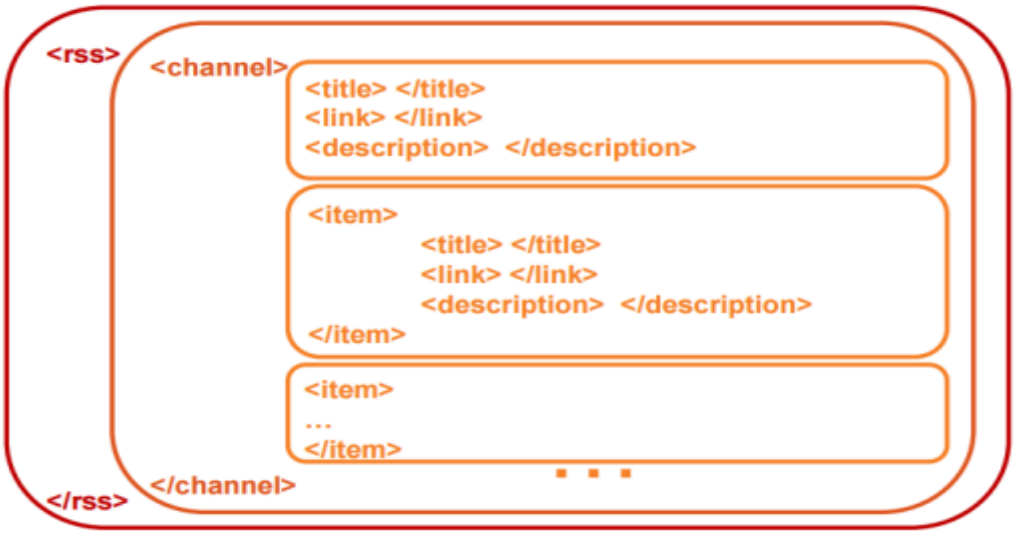

Figure 10 : La structure générale de flux RSS

3.2.2.2 Agrégation des données

Cette étape consiste à agréger les données des sources identifiées, pour cela, un programme sera lancé automatiquement pour faire l'agrégation des données qui seront stockées par la suite dans une BD.

Les informations agrégées sont : Titre, Description, Auteur, Date de publication, Lien.

Ci-dessous le pseudo algorithme qui décrit l'étape de l'agrégation des données.

Pseudo-algorithme agrégation des données

1. Entrées : sources d'actualité.

2. Sorties : « titre, description, auteur, date, heure, url ».

DEBUT

Pour chaque source faire :

Extraire les données « titre, description, auteur, date, url » à partir des flux RSS

Fin pour

FIN

Figure 11:pseudo algorithme de l'agrégation des données

3.2.2.3 Prétraitement des données agrégées

Notre système prend en considération la langue de la source de données, car chaque langue a ses propres caractéristiques. Nous avons choisi le champ titre de l'information pour le traitement, car ils signifient le contenu de l'article. Les différentes techniques utilisées sont :

- **Tokenisation** : Dans cette partie, l'opération de Tokenisation est l'acte de décomposer une séquence de chaînes en éléments appelés jetons (tokens). Le processus de tokenisation consiste à transformer le texte en liste des tokens**, exemple** la phrase " Reprise timide des commerces à Alger" devient " Reprise "," timide ", "des"," commerces ", " à ", "Alger ".
- **Suppression des mots vides (stop words)** : dans cette étape nous supprimons des mots qui ne portent également aucune information. Par exemple la phrase précédente devient " Reprise "," timide ", " commerces ", "Alger ".
- **Lemmatisation :** c'est l'opération qui consiste à Transformer les flexions en leur lemme ; **Exemple : pris, prend, prisse → prendre**

3.2.2.4 Détection des entités nommées

La détection des entités nommées est une tâche essentielle dans notre système, après l'agrégation de flux d'actualités **(RSS)**, on effectue la tâche d'extraction des entités nommées afin de pouvoir identifier **les entités cibles d'une opinion**.

Pour cela, nous utilisons une méthode de combinaison entre deux approches symboliques et statistiques.

On s'intéresse à identifier les différentes entités personne morale ou physique (noms de personne et d'organisation).

Par exemple : Condamnation de Khaled Drareni : L'Union européenne discute avec l'Algérie.

Dans cet exemple deux entités nommées seront identifier par notre système :

Organisation : ['L'Union', 'européenne']

Personne : ['Khaled', 'Drareni']

Ci-dessous le pseudo algorithme qui permet de détecter les entités nommées.

Pseudo algorithme extraction des EN

1. Entrées : sources d'actualité.

2. Sorties : classes des entités nommées (personne), (organisation).

DEBUT

Pour chaque source faire :

    Extraire la paire d'EN (token, type d'EN)

Fin pour

FIN

Figure 12 : pseudo algorithme d'extraction des EN

3.2.2.5 Classification et Extraction de tonalité

Cette étape représente la partie principale du système, il s'agit d'extraire la tonalité globale exprimé dans l'article d'actualité. Pour cela nous utilisons la méthode de **classification à base de règles**. Elle est basée en fait sur un dictionnaire de mots, qui prend les scores de mots individuels de SentiWordNet [10]Ou chaque mot est annoté par sa polarité et sa subjectivité (objective, subjective).

SentiWordNet : est une ressource lexicale pour l'extraction d'opinion. SentiWordNet attribue à chaque synset de WordNet trois scores de sentiment : positivité, négativité, objectivité.

Pour chaque information d'actualité (titre) nous avons une métrique qui donne la polarité sous forme d'un tuple nommé de la forme Sentiment **(polarité, subjectivité)**.

Exemple1 : <word form="condamné" pos="JJ" polarity="-0.25" subjectivity="0.50" />

Soient :

Pos : l'étiquetage grammaticale de mot.

Le score de polarité est un flottant dans la plage **[-1.0, 1.0].**

La subjectivité est un flottant dans la plage **[0.0, 1.0]** où **0.0** est très objectif et 1.0 est très subjectif.

[10] http://ontotext.fbk.eu/sentiwn.html

Dans cet exemple la polarité du mot **« condamné »** est **négative (-0.25)** ce qui engendre son positionnement dans la classe négative.

Exemple 2:<word form="adorable" pos="JJ" polarity="0.70" subjectivity="0.80"/>

Dans cet exemple la polarity du mot **« adorable »** est **positive (0.70)** ce qui engendre son positionnement dans la classe positive.

Exemple 3 : <word form="administratif" pos="JJ" polarity="0.00"subjectivity="0.00" />

Dans cet exemple la polarité du mot **« administratif »** est **neutre (0.00)** ce qui engendre son positionnement dans la classe neutre.

En général, un texte donné sera représenté par un sac de mots. Après avoir attribué des scores individuels à tous les mots, le sentiment final est calculé par une opération de mise en commun comme la moyenne de tous les sentiments.
Ci-dessous le pseudo algorithme qui permet l'analyse de tonalité :

Pseudo algorithme extraction de tonalité

1. Entrées : sources d'actualité.

2. Sorties : classes de tonalité (positive, négative, neutre).

DEBUT

Pour chaque source faire :

    Polarité = Analyser les données (subjectivité)

    Si (polarité>0) alors Retourner (positive)

    Sinon Si (polarité<0) alors Retourner (négative)

    Sinon Retourner (neutre)

 Fin pour

FIN

Figure 13 : pseudo algorithme analyse de tonalité

3.2.2.6 Visualisation des résultats

C'est la dernière étape de notre système, sert à visualiser les résultats de traitement aux utilisateurs. Pour cela le système doit envoyer périodiquement une synthèse de l'actualité à la cible.

3.3 Conception de l'application d'analyse d'opinion dans l'actualité

Dans cette partie, nous décrivons la conception de notre application d'analyse d'opinion dans l'actualité.

3.3.1 Identification des besoins

C'est la première phase et la plus importante dans le développement d'une application, car elle permet la compréhension du domaine et du problème en collectant le maximum d'informations. Notre but est d'analyser de manière automatique les articles de presse en ligne pour l'analyse d'opinion. Le système doit permettre notamment de :

- Agrégation de flux d'actualité qui sont enfaite des fichiers XML souvent utilisées par les sites d'actualité et les blogs,
- Identification des entités nommées,
- Extraction des tonalités.

3.3.2 Identification des acteurs

Un acteur représente un rôle d'un utilisateur qui interagit avec le système. L'utilisateur peut être un utilisateur humain, une organisation, une machine ou un autre système externe. Notre système interagisse avec les acteurs suivants :

Acteur	Le rôle
Agent (déclencheur automatique)	Un acteur non humain, Il peut être une machine ou un robot qui déclenche les différentes fonctionnalités du système, il collecte les informations à partir des différentes sources, détecte les tonalités et stocke les informations dans la BD.
Utilisateur	Il s'agit de la personne qui visualise les informations.
Administrateur	Son rôle est de gérer les sources et les utilisateurs.

Tableau 3: Les acteurs du système

3.3.3 Présentation de cas d'utilisation

Les diagrammes de cas d'utilisation modélisent le comportement d'un système et permettent de capturer les exigences du système. Chaque usage que les acteurs font du système est représenté par un cas d'utilisation. Afin de répondre aux problèmes liés à l'analyse d'opinion

pour une veille informationnelle, nous présentons le diagramme de cas d'utilisation suivant (figure 14) :

- Accès au système,
- Gérer les utilisateurs,
- Gérer les sources,
- Analyser les informations,
- Effectuer l'analyse de l'actualité.

3.3.4 Diagramme de cas d'utilisation

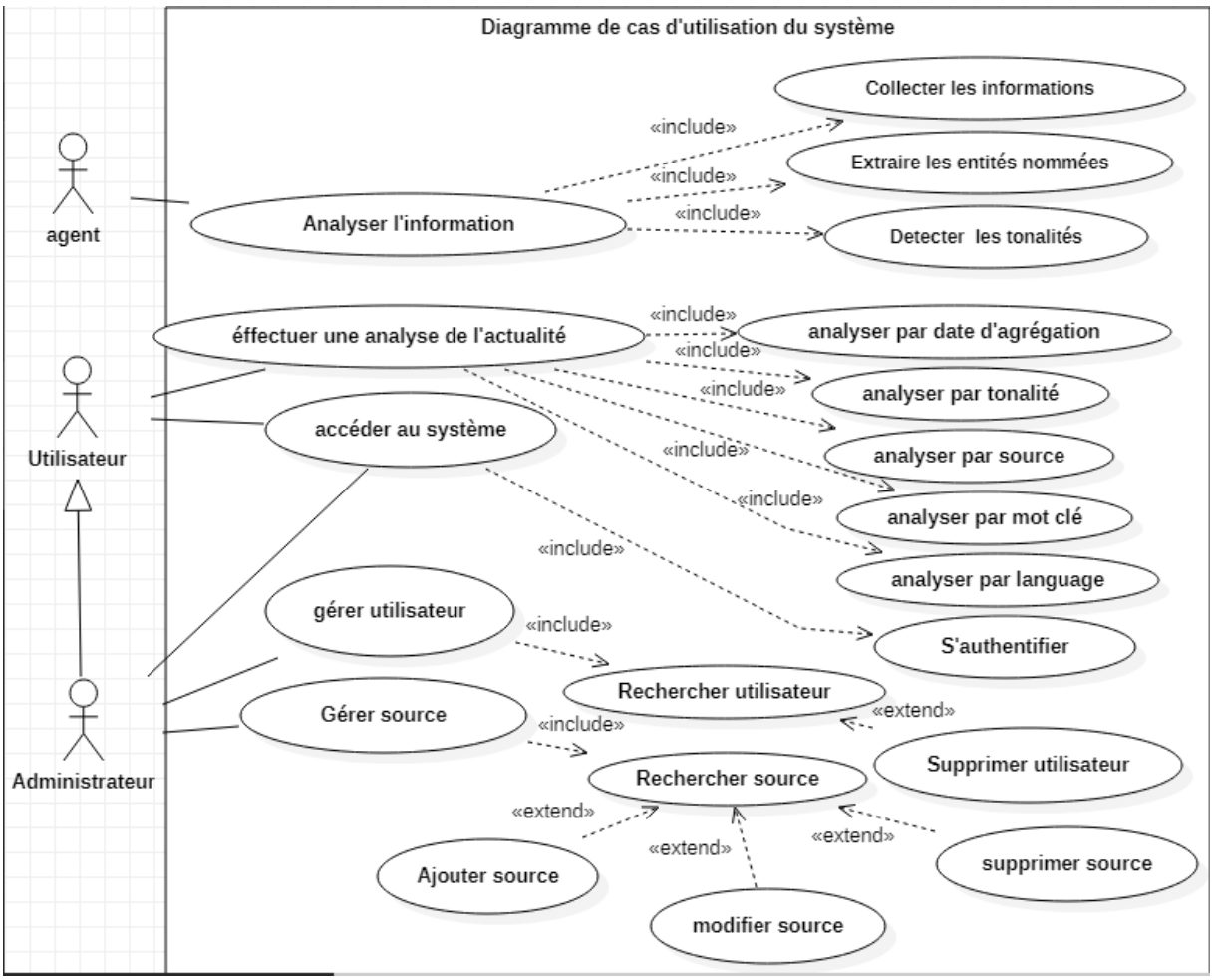

Figure 14: Diagramme de cas d'utilisation

3.3.5 Description des cas d'utilisation

Cas d'utilisation accès au système

- L'utilisateur demande le formulaire d'authentification,
- Le système affiche le formulaire d'authentification,
- L'utilisateur saisit son identifiant et le mot de passe,

- Le système vérifie si les informations sont valides,
- Le système affiche la page d'accueil.

Cas d'utilisation collecter les données

- Le système collecte des données issues du web à partir des différentes sources de flux RSS choisis par l'administrateur et extrait les informations nécessaires (titre, description, auteur, date de publication, lien) à partir des données collectées,
- Stocke les données agrégées dans la BD.

Cas d'utilisation extraire les EN

- Le système récupère les données agrégées,
- Effectue un prétraitement (tokenisation, suppression des mots vides, stemming) sur les données agrégées,
- Extrait les EN détectées,
- Envoi les résultats pour les stockés dans la BD.

Cas d'utilisation détecter les tonalités

- Le système analyse les données agrégées,
- Extrait la tonalité (positive, négative, neutre),
- Stocke les résultats dans la BD.

Cas d'utilisation gérer les utilisateurs

Administrateur peut effectuer les tâches suivantes :

- La recherche d'un utilisateur
- La suppression d'un utilisateur

Cas d'utilisation gérer les sources

Administrateur gère les sources à travers :

- L'ajout des nouvelles sources
- La modification des informations de la source
- La suppression des sources

3.3.6 Diagrammes de séquences

Les diagrammes de séquences sont la représentation graphique des interactions entre les acteurs et le système selon un ordre chronologique. Le diagramme de séquence permet de

montrer les interactions d'objets dans le cadre d'un scénario d'un diagramme des cas d'utilisation.

3.3.6.1 Scénario d'authentification

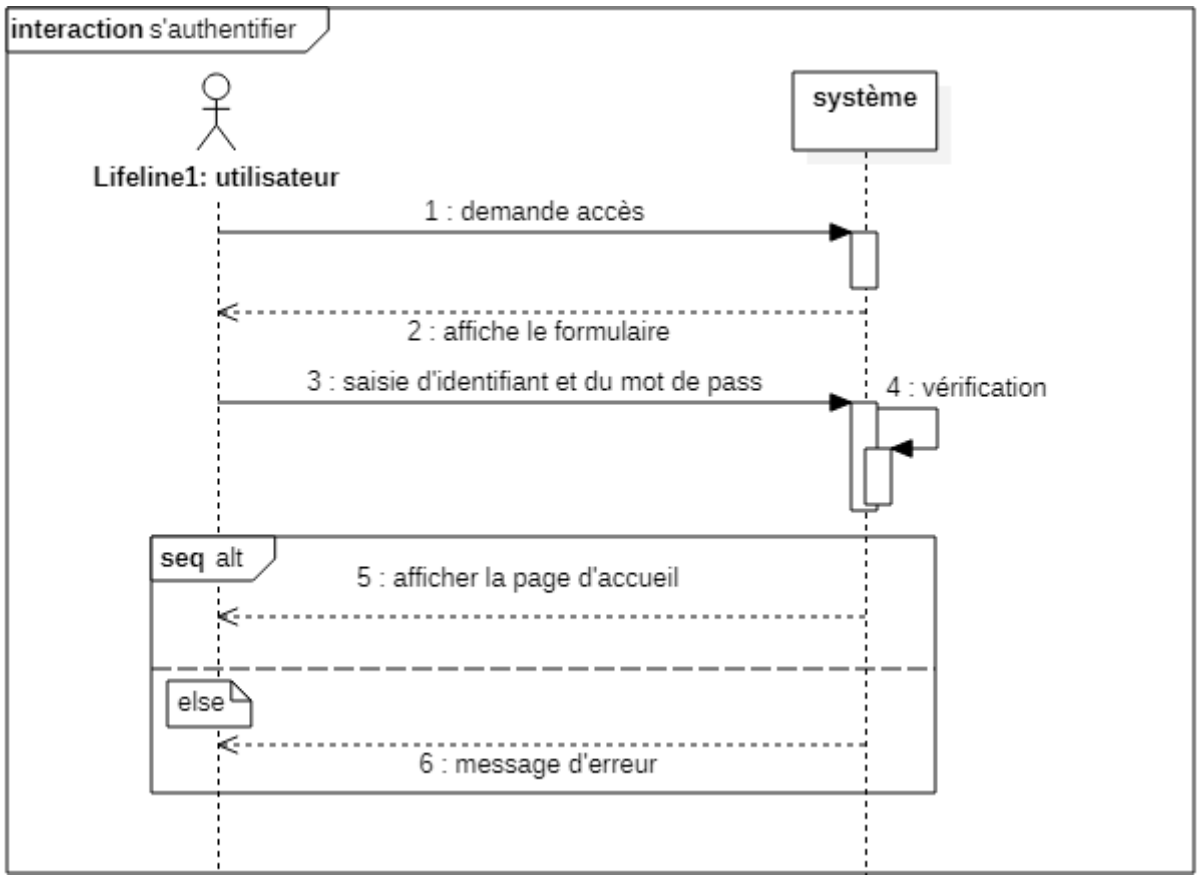

Figure 15: Diagramme de séquence pour s'authentifier

3.3.6.2 Scénario d'agrégation des données

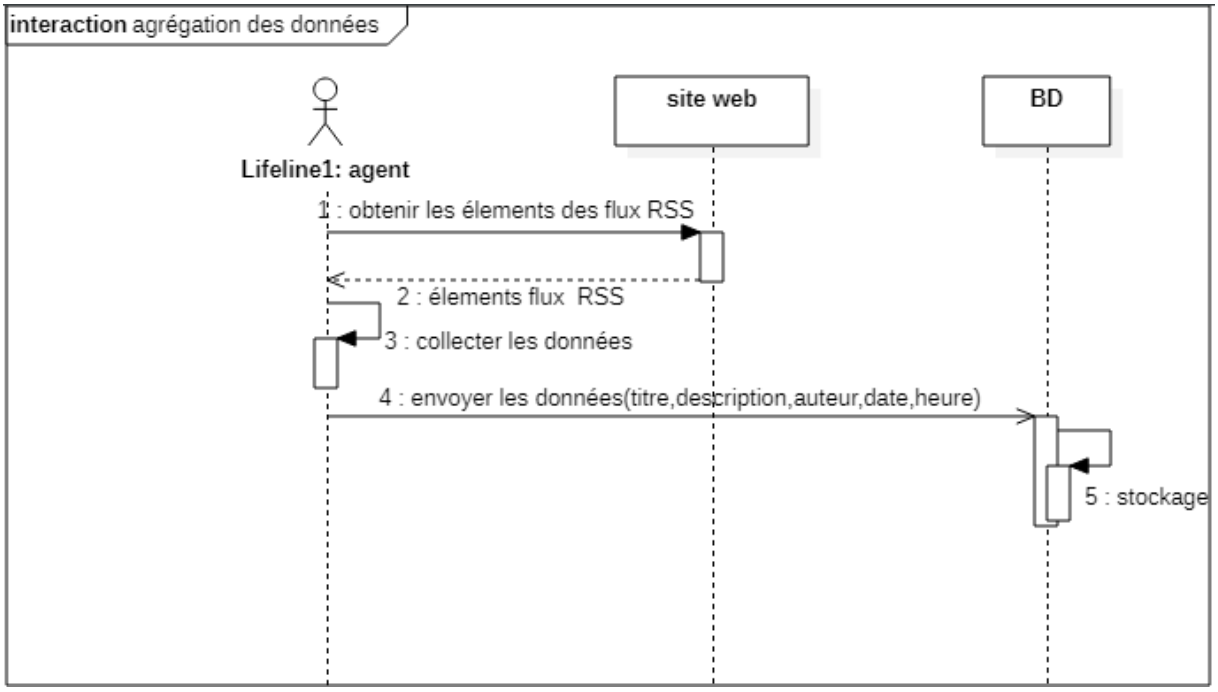

Figure 16: Diagramme de séquence agrégation des données

3.3.6.3 Scénario d'extraction des entités nommées

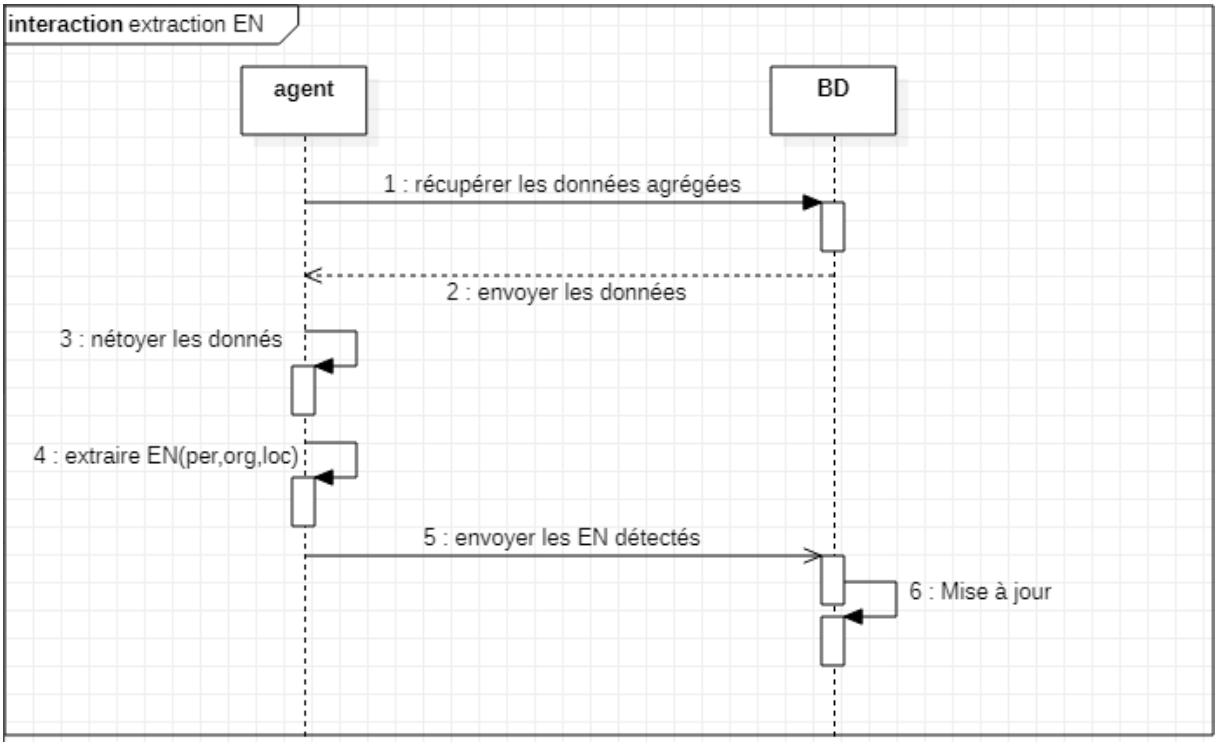

Figure 17: Diagramme de séquence extraction des entités nommées

3.3.6.4 Scénario extraction des tonalités

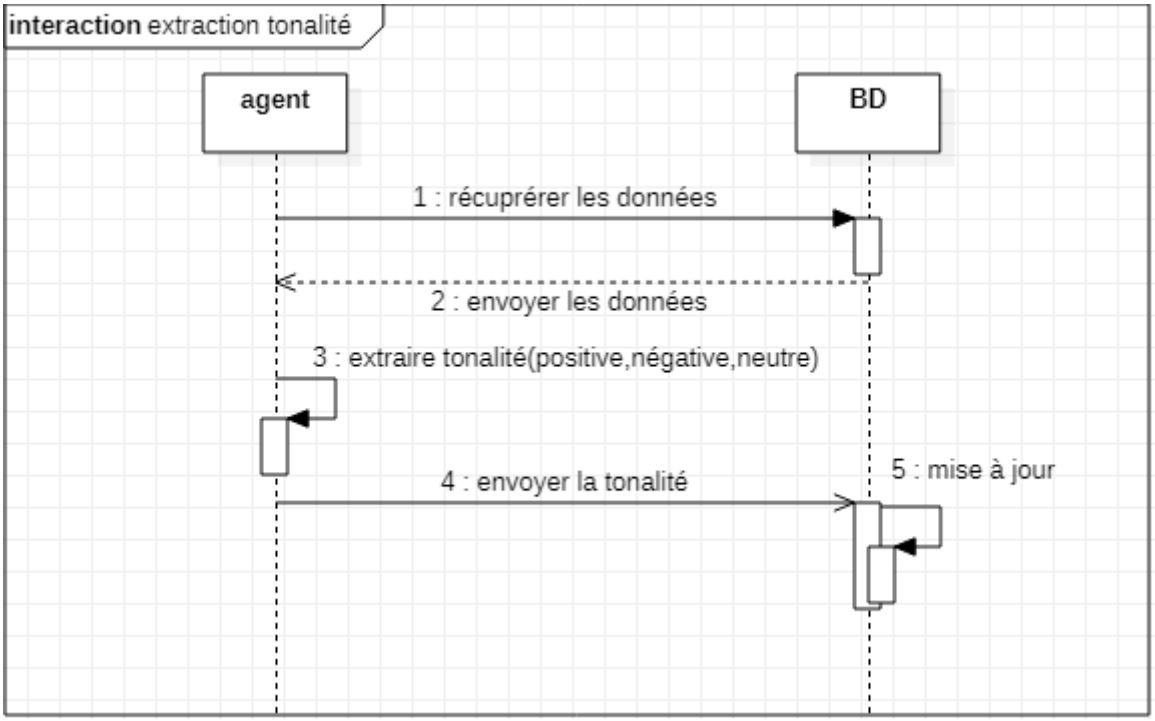

Figure 18: Diagramme de séquence extraction de tonalité

3.3.7 Diagramme de classes

Le diagramme de classe est considéré comme le plus important dans la modélisation orientée objet, il permet de fournir une représentation abstraite des objets du système qui vont interagir pour réaliser les cas d'utilisation. Après l'identification des besoins concernant notre projet, nous avons proposé le diagramme de classes suivant :

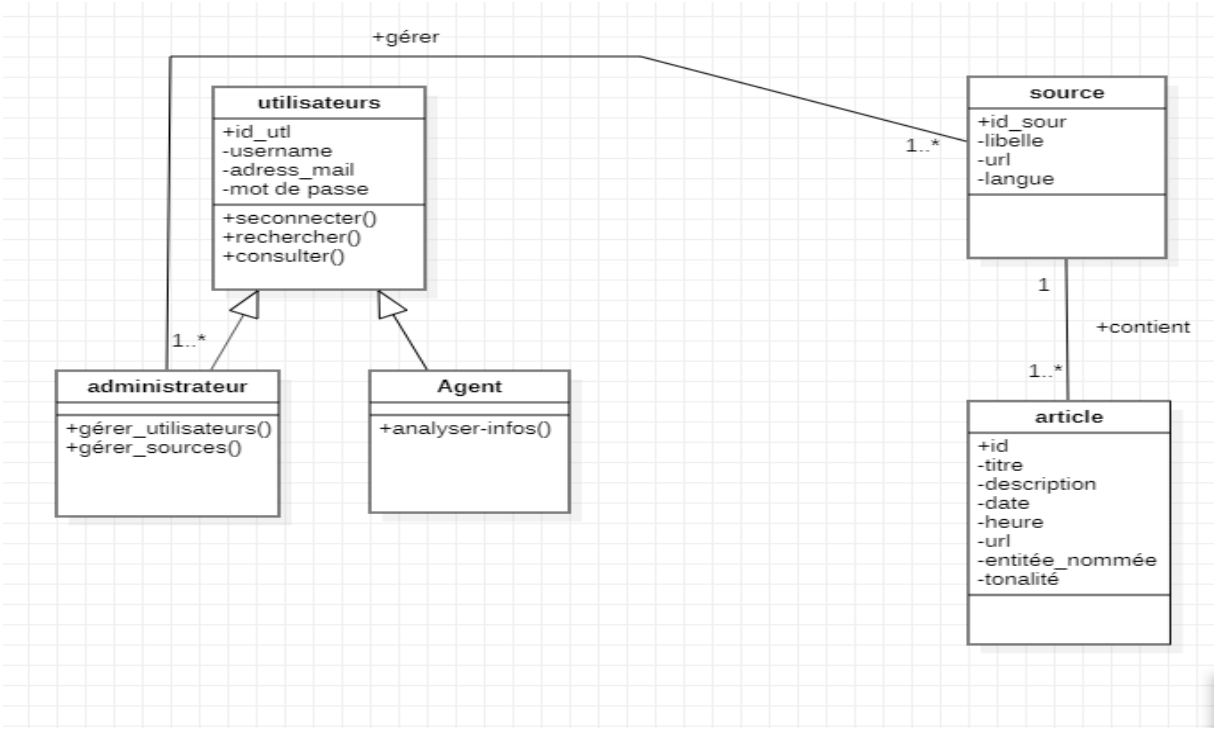

Figure 19: Diagramme de classes

Le diagramme de classes spécifie également les relations entre les objets de notre système (utilisateurs gérer les sources), aussi il permet de définir les différents opérations (**fonctions**) associées aux objets de la classe par exemple : (les fonctions associées à l'administrateur sont : gérer les utilisateurs et gérer les sources).

3.4 Conclusion

Dans ce chapitre nous avons présenté notre méthode de conception ainsi que l'architecture générale de notre système.

Un modèle d'apprentissage a été conçus pour répondre à l'objectif principale de notre projet : pour la prédication de tonalité.

Dans le chapitre suivant nous allons décrire l'implémentation ainsi que l'évaluation de notre Système.

4 Implémentation et évaluation

4.1 Introduction

Dans cette partie et après avoir présenté notre conception et expliqué les déférentes étapes de notre système, nous allons aborder dans un premier temps, les déférents outils, bibliothèques et sources utilisées, ensuite une présentation de notre application sera donnée. Enfin, nous clôturerons ce chapitre par l'évaluation des performances de notre solution.

4.2 Les outils utilisés

4.2.1 Langages de programmation

- **Python :** est un langage de programmation interprété, multiparadigme et multiplateformes.[61] On a le choisi à cause de la simplicité et la rapidité de développement, encore un langage facile à apprendre et à utiliser, fonctionne sur tous les principaux systèmes d'exploitation, est utilisé pour crée des logiciels de qualité professionnelle.

- **PHP (HyperText Préprocesseur) :** est un langage de script utilisé côté serveur. Il est utilisé dans le développement web ainsi que comme langage de programmation général.[62] On a l'utilisé à cause de ces avantages tel que sa flexibilité et sa grande compatibilité avec d'autres bases de données, écriture générique et classe et encore facile à maîtriser.

4.2.2 Serveur WEB

- **Apache 2 :** est un logiciel de serveur web open-source et gratuit même pour un usage commercial, il permet de servir un contenu sur le web, fiable et stable, flexible grâce à sa structure basée sur des modules, facile à configurer et plateforme-Cross (fonctionne sur les serveurs Unix et Windows).[63]

4.2.3 Base de données

- **MySQL :** est un serveur de bases de données relationnelles Open Source développé et supporté par Oracle. Simple à utiliser que la plupart des serveurs de bases de données commerciaux. Il stocke les données dans des tables, permet de manipuler ces bases de données et de diriger l'accès à leur contenu il emploie pour cela SQL (Structured Query Language) (un langage de requête) aussi on peut effectuer diverses opérations en utilisant des interfaces écrites en C, C++, Java, Python, PHP.[64]

4.2.4 Outils de développement

- **HTML (HyperText Markup Language) :** est un langage de balisage, a été construit pour décrire les documents hypertexte sur internet. Il est simple à utiliser, sa conception lui permet de rester indépendant vis à vis des plates-formes et de pouvoir être échangé sur les réseaux, il peut être composé sur n'importe quel système avec un simple éditeur de textes.[65]

- **CSS (Cascading Style Sheets) :** est un système destiné à mettre en forme les contenus de pages Web comme HTML et XML. L'utilisation de CSS permet facile d'entretenir votre site internet et de le mettre à jour, Une plus grande cohérence lors de la Conception de votre site, le temps de téléchargement de votre site sera plus rapide. [66]

- **Java script :** est un langage de programmation utilisé notamment lors de la conception de sites web et d'applications. Il est particulièrement utile pour concevoir des sites dynamiques. Ces avantages sont : compatible avec tous les supports numériques, prise en compte de l'expérience utilisateur et gain de temps.
- **Bootstrap :** Bootstrap est le Framework HTML, CSS et JavaScript le plus populaire pour le développement de sites Web réactifs et adaptés aux mobiles. [67] On utilise bootstrap car il nous assure d'avoir une certaine cohérence et une certaine robustesse dans l'ensemble de notre design.

4.3 Les bibliothèques utilisées

- **NLTK (Natural Language Toolkit) :** est une plate-forme pour la création des programmes Python utilisant des données en langage humain, contenant un ensemble de fonctions destinées au traitement automatique du langage naturel. Il fournit des interfaces faciles à utiliser avec plus de 50 corpus et ressources lexicales tels que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la création de token, le stemming, le balisage, l'analyse et le raisonnement sémantique etc. on a utilisé la bibliothèque NLTK[11] dans la phase des prétraitements car elle supporte plusieurs langues .
- **Feedparser :** est une bibliothèque Python facile à utiliser qui analyse les flux dans tous les formats connus, y compris Atome, RSS et RDF.[68]

[11] https://www.nltk.org/

- **Polyglot :** est un pipeline de langage naturel prenant en charge d'énorme application multilingue. il prend en charge la tokenisation pour 165 langues, la détection de la langue pour 196 langues, la reconnaissance d'entités nommées pour 40 langues,, l'étiquetage vocal pour 16 langues , l'analyse des sentiments pour 136 langue , les enveloppements de mots pour 137 langues, l'analyse morphologique pour 135 langues, la translittération pour 69 langues, dans notre programme on a utilisé la reconnaissance d'entités nommées pour la langue arabe et la langue française offert par polyglot.[69]
- **SpaCy :** SpaCy est une bibliothèque python open-source pour NLP écrite en Python et Cython. Il propose des modèles pré-formés pour les NER multilingues, Les modèles SpaCy sont conçus pour être prêts pour la production. Parmi ces modèles on a utilisé fr_core_news_lg de taille 545 MB qui fournit des résultats beaucoup mieux que les autres modèles de français tel que (fr_core_news_sm, fr_core_news_md).[70]
- **TextBlob** : TextBlob est une bibliothèque Python pour le traitement de données textuelles. Il fournit une API cohérente pour plonger dans les tâches courantes de traitement du langage naturel (NLP) telles que part-of-speech tagging, noun phrase extraction, sentiment analysis .[71]

4.4 Les sources RSS exploitées

Pour évaluer notre système, nous avons exploité une quarantaine de sources RSS qui sont des sources d'actualité en arabe et d'autres en français, ou dans chaque source on retrouve plusieurs articles d'actualité.

a) Sources en langue arabe

1. Elbilad, http://www.elbilad.net/feed/article .
2. Dzair-tube, https://dzair-tube.com/feed/.
3. El khabar, https://www.elkhabar.com/feeds/.
4. Algeriatimes, http://www.algeriatimes.net/xmlfile/news.xml .
5. Sabq, https://www.sabqpress.net/feed/.
6. Ennahar, https://www.ennaharonline.com/feed/.
7. Aljazair alyoum, https://www.aljazairalyoum.com/feed/.
8. TSA_A, https://www.tsa-algerie.com/ar/feed/
9. APS_A, http://feeds.feedburner.com/APS-Algerie-ar.
10. Radio Algérie_A, http://www.radioalgerie.dz/news/ar/rss.xml.
11. Algérie part_A, https://algeriepart.com/ar/feed/.

12. Bourse-dz_A, http://bourse-dz.com/ara/feed/.
13. RT_A, https://arabic.rt.com/rss/ ,ar.
14. Elmihwar, http://elmihwar.com/ar/.
15. Jeune-independant, http://www.jeune-independant.net/spip.php?page=backend.
16. Echorouk, https://tv.echoroukonline.com/echorouk-tv/feed/.
17. Le quotidien_algerie, https://lequotidienalgerie.org/feed/.

b) Sources en langue française

1. Le Soir d'Algérie, https://www.lesoirdalgerie.com/article/feed.
2. Liberté, https://www.liberte-algerie.com/article/feed.
3. Algérie-focus, https://www.algerie-focus.com/feed/.
4. TSA, https://www.tsa-algerie.com/feed/.
5. APS, http://feeds.aps.dz/aps-algerie.
6. Algérie part, https://algeriepart.com/feed/.
7. Observ Algérie, https://www.observalgerie.com/feed/.
8. Algerie360, http://www.algerie360.com/feed/.
9. Elwatan, https://www.elwatan.com/feed.
10. Express-dz, https://www.express-dz.com/feed/.
11. Casbah, http://casbah-tribune.com/feed/.
12. Bourse-dz, http://bourse-dz.com/feed/.
13. Eco, https://www.algerie-eco.com/feed/.
14. Ntic, https://www.ntic-dz.com/feed/.
15. Emergent, https://maghrebemergent.info/feed/.
16. Radio-Algerie, http://www.radioalgerie.dz/news/fr/rss.xml.
17. 24, https://www.alg24.net/feed/.
18. RT, https://francais.rt.com/rss.
19. Lexpressiondz, http://www.lexpressiondz.com/feeds/index.rssi.
20. France 24, https://www.france24.com/fr/rss.
21. Tvmaghreb, https://tvmaghreb.com/feed/.
22. Dia, http://dia-algerie.com/feed/.
23. Dzair Daily, https://www.dzairdaily.com/feed/.
24. Ouest-france, https://www.ouest-france.fr/rss-en-continu.xml.

4.5 Les données agrégées

On a utilisé l'API de Python feedparser pour Agréger les données des sources identifiées, après chaque durée de temps un programme sera lancé automatiquement pour faire l'agrégation des données.

Après l'extraction des données à partir du flux RSS <express-dz>, nous avons obtenu les résultats comme suit :

Titre	Description	Auteur	Date de pub	Lien
Coronavirus : 3271 personnes guéris depuis l'apparition de la pandémie en Algérie	Cent-quatre-vingt-sept (187) cas confirmés au coronavirus, 113 guérisons et 7 décès ont été enregistrés durant les dernières 24 heures en Algérie, a indiqué vendredi à Alger le porte-parole du comité scientifique de suivi d'évolution de la pandémie du Coronavirus.	Mohand NB	2020-05-16	https://www.express-dz.com/...

Tableau 4:Exemple de résultat obtenu par le scraping(feedparser)

4.6 Présentation et description de l'application

Notre système est connecté à la base de données MySQL. On a planifié notre système d'exécuter automatiquement chaque une durée précise pour actualiser les résultats. Les pages de notre site sont :

4.6.1 Page d'accueil

Figure 20 : page d'accueil

Cette page d'accueil permet un utilisateur de se connecter ou de s'inscrire a travers les boutons de l'entête « inscription » et « connexion ».

4.6.2 Page d'inscription :

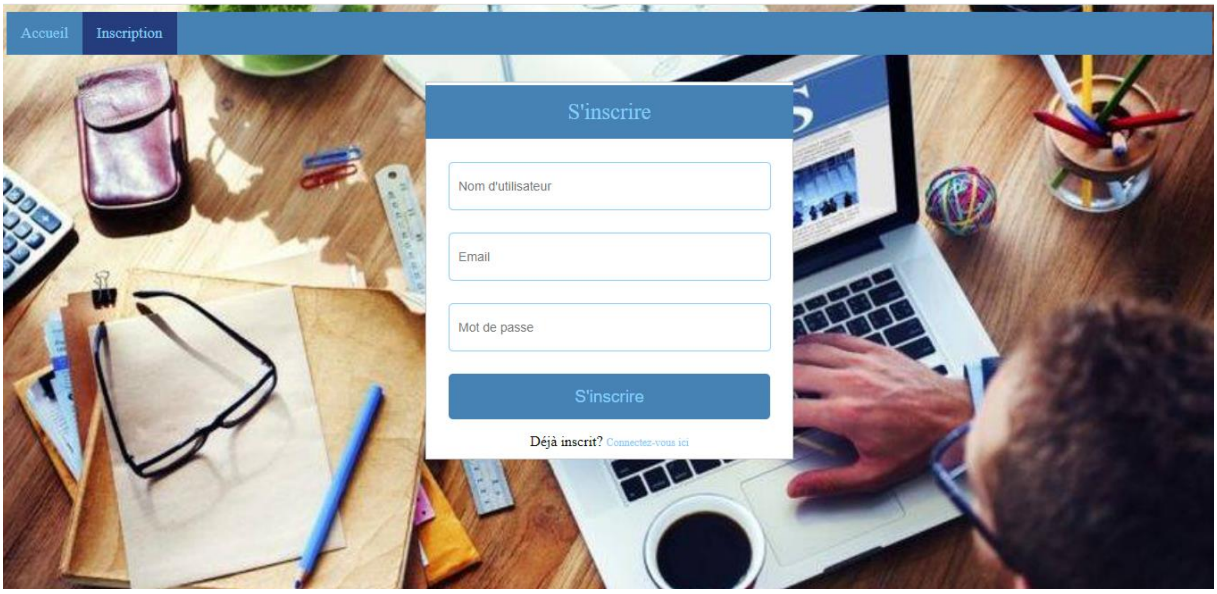

Figure 21 : Page d'inscription

Pour créer un compte il faut remplir le formulaire, tous les champs sont obligatoires, doit respecter les déférentes formes des champs, par exemple l'E-mail s'écrit de cette façon : "exemple@mail.com".

4.6.3 Page d'analyse de l'actualité

Figure 22 : page d'analyse de l'actualité

Page de consultation de l'actualité contient un tableau des informations agrégées (Titre, Auteur, Description, Date, Source, entité nommée, tonalité et Lien vers l'article), les informations affichées en bleu sur les lignes du tableau sont les titres des articles de tel source (7). On peut spécifier les informations par le remplissage des champs au-dessous de l'entête chaque zone nous permet de :

1 : affiche les informations par une date d'agrégation

2 : par un terme ou un mot clés

3 : par source

4 : par une langue (Arabe ou Française)

5 : par tonalité (Positive, négative ou neutre)

6 : valider les choix

Voici des exemples d'affichage des informations spécifiées, on a choisi d'afficher les informations qui sont agrégées à la date 23/09/2020 avec :

(a) La tonalité "negative"

Figure 23 : Exemple d'affichage des informations par tonalité

(b) La source "Express-dz"

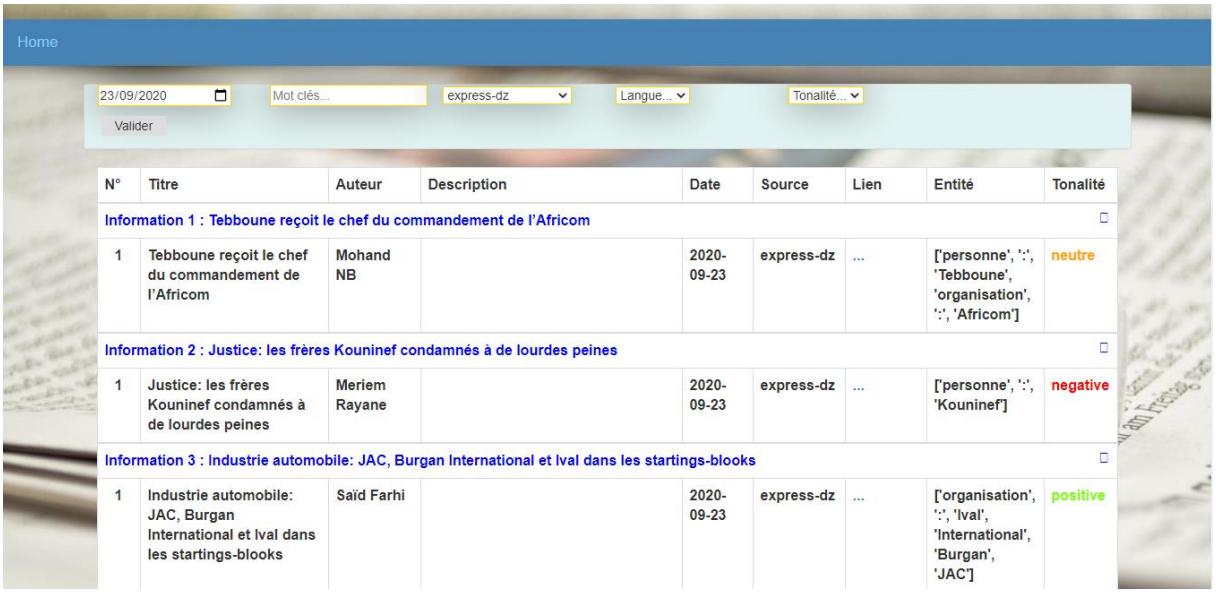

Figure 24 : Exemple d'affichage des informations par source

(c) La langue "Arabe"

Figure 25 : Exemple d'affichage des informations par langue

(d) Le mot clés "covid-19"

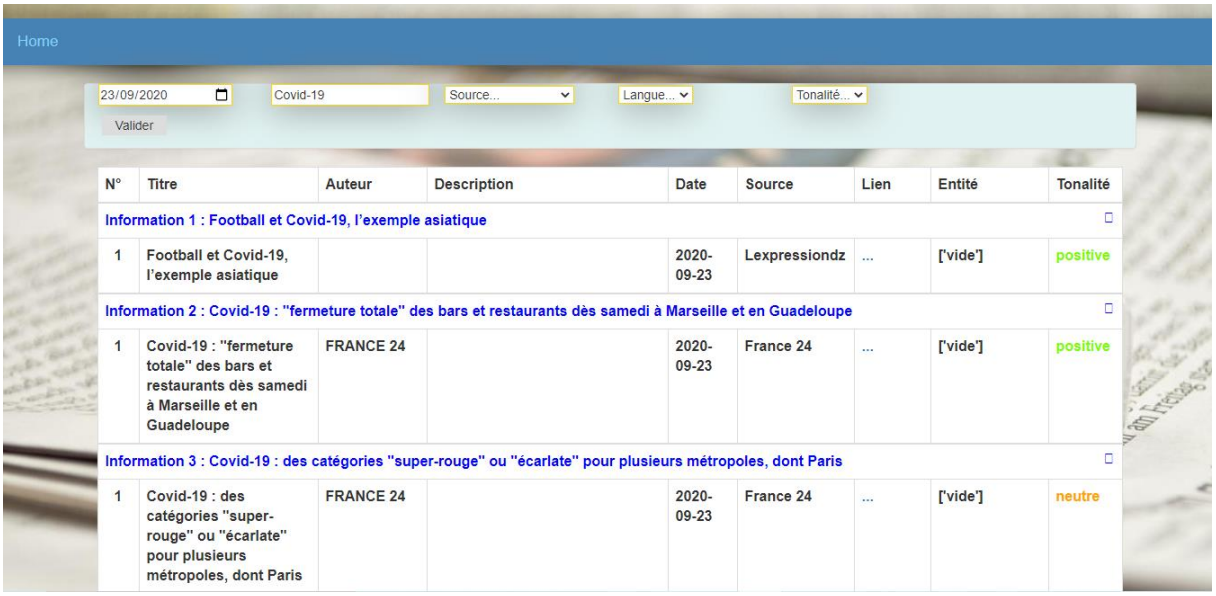

Figure 26 : Exemple d'affichage des informations par mot clés

4.6.4 Page de gestion des sources (profile administrateur)

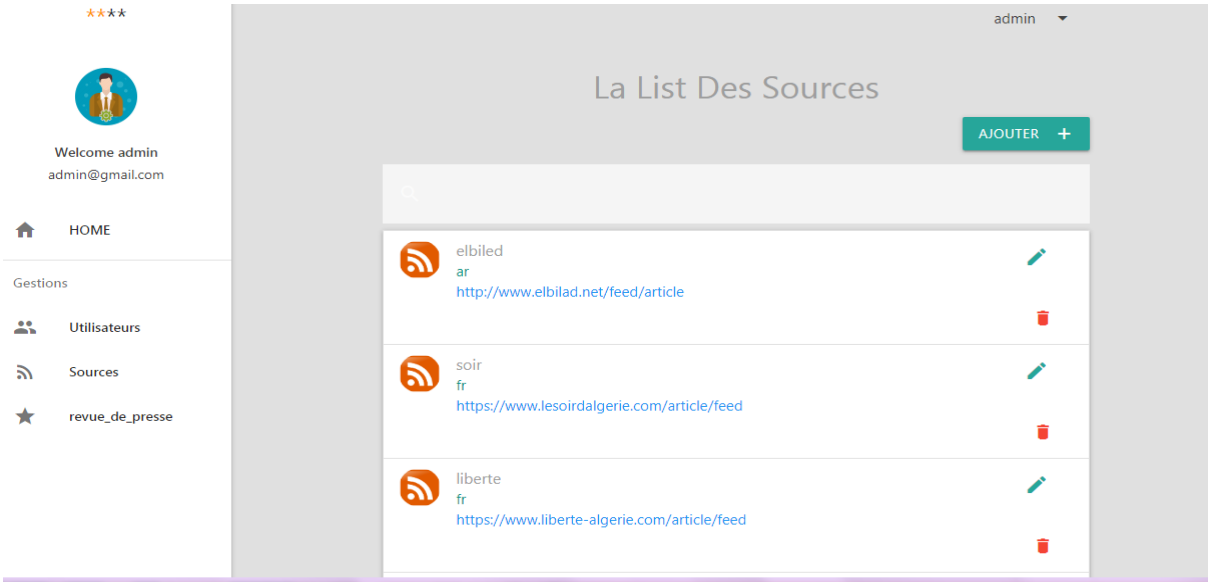

Figure 27 : Page de gestion des sources

La liste de toutes les sources est affichée ici. Pour ajouter une source, il suffit d'entrer son libelle, son lien du RSS et sa langue et puis on clique sur "Ajouter". La source sera affichée à la fin de la liste. Pour supprimer une source cliquez sur le bouton rouge de la colonne des actions. Si on clique sur le bouton vert de la modification, les informations de la source seront affichées sur les champs du formulaire et le bouton "Ajouter" va changer vers "Modifier", nous pouvons aussi rechercher la source par libelle et langue.

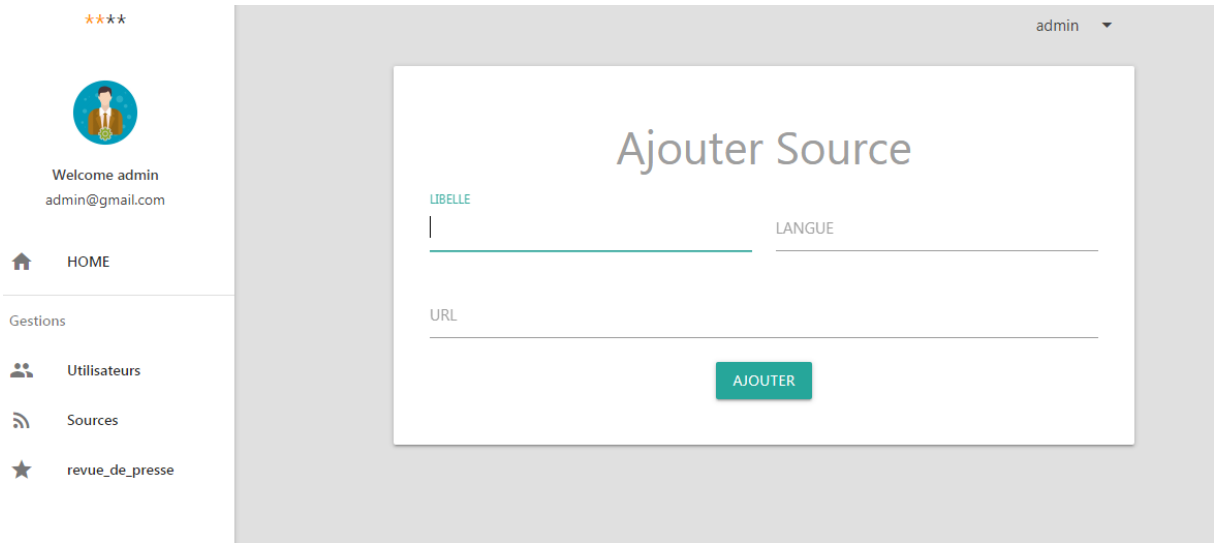

Figure 28 : Exemple d'ajouter une source

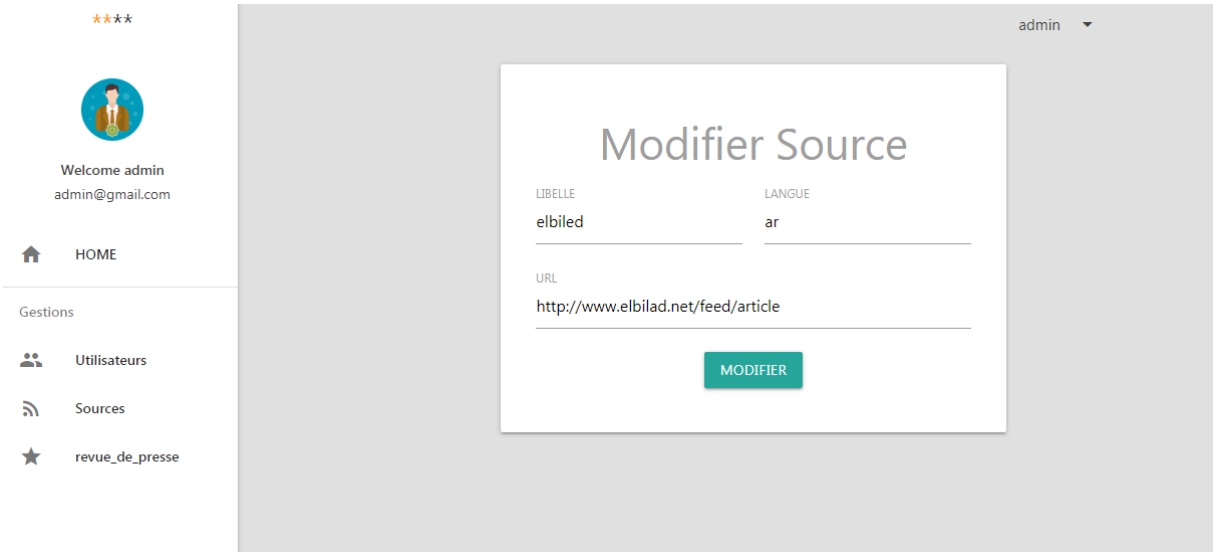

Figure 29 : Exemple de modifier une source

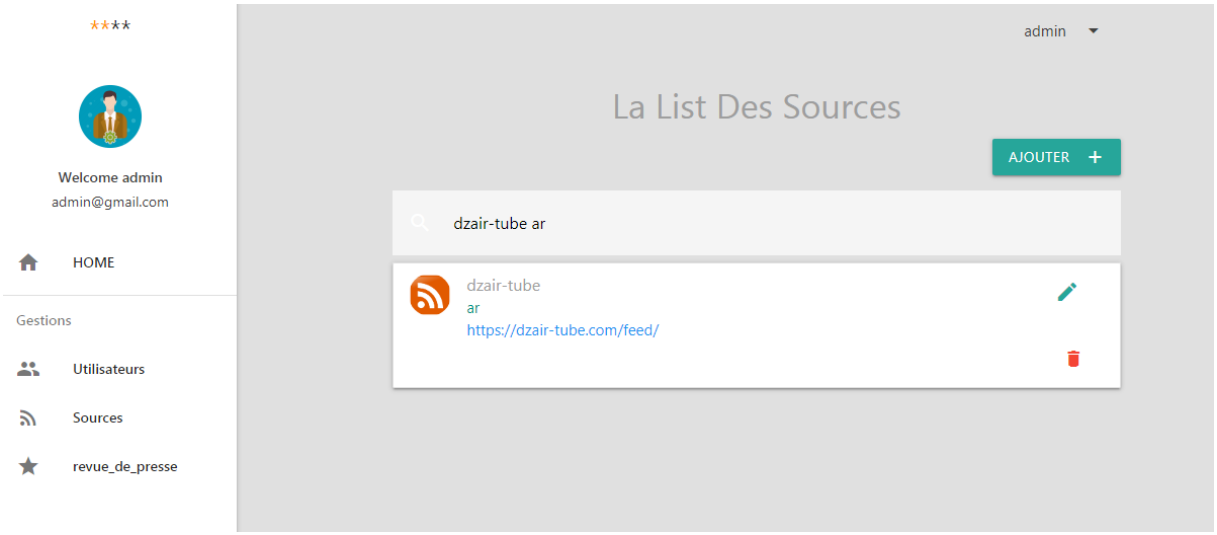

Figure 30 : Exemple de rechercher une source

4.6.5 Page de gestion des utilisateurs (profil administrateur)

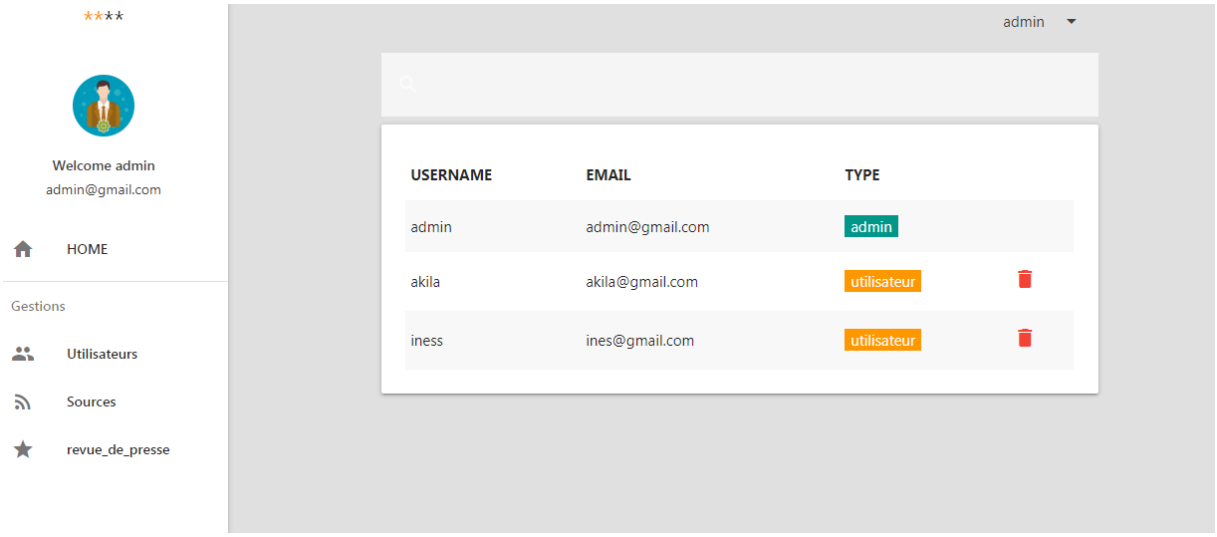

Figure 31 : Page de gestion des utilisateurs

La liste de tous les utilisateurs est affichée ici. Pour supprimer un utilisateur, il suffit de cliquer sur le bouton rouge de la colonne des Actions ; Vous pouvez également rechercher l'utilisateur qui doit être supprimé via son nom d'utilisateur ou bien son e-mail.

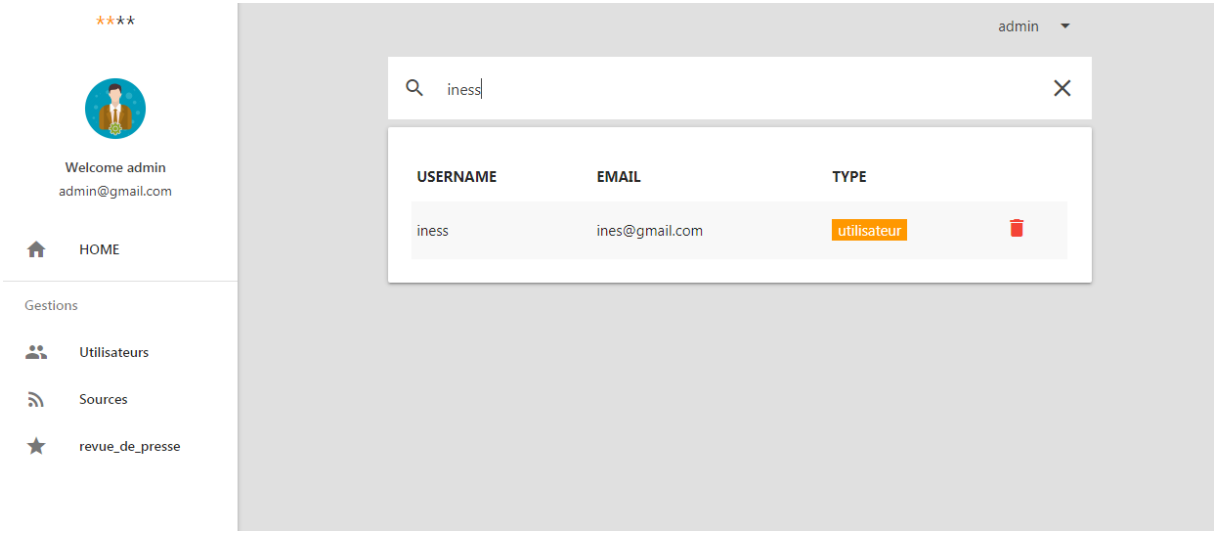

Figure 32 : rechercher utilisateur par Username

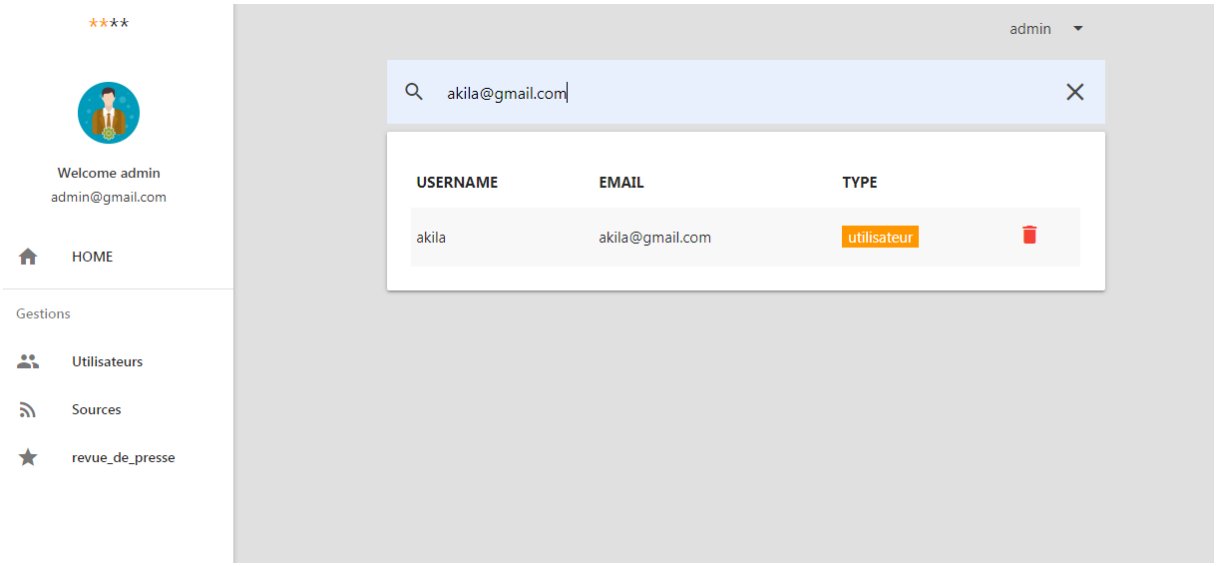

Figure 33 : rechercher utilisateur par E-mail

4.7 Évaluation du système

Due au manque de benchmark spécialisé dans notre domaine étudié nous avons donc construire notre propre corpus de test.

Notre corpus d'évaluation, est formé d'une collection de documents qui sont collectés de l'actualité diffusée par les sources média algériennes (arabe et français).

Nous avons fait une évaluation semi-automatique sur un corpus qui contient 130 documents dans lequel il y'a 50 docs en langue arabe et 80 docs en langue française. Evaluation sert à comparer la liste des EN et l'ensemble des classes d'opinion retournées par le système et ceux définit par l'utilisateur.

4.7.1 Évaluation du système NER

Il existe une bonne gamme de modèles préformés de reconnaissance d'entités nommées **(NER)** fournis par les bibliothèques **NLP** open source populaires (par exemple **NLTK, Spacy, StanfordCoreNLP**) et certains moins connus (par exemple **Polyglot**).

Pour évaluer notre système NER, nous avons fait une étude comparative entre ces outils en termes des mesures de performances notamment rappel, précision et F_mesure nous avons obtenu les résultats illustrés dans *(les figures 34 ,35).*

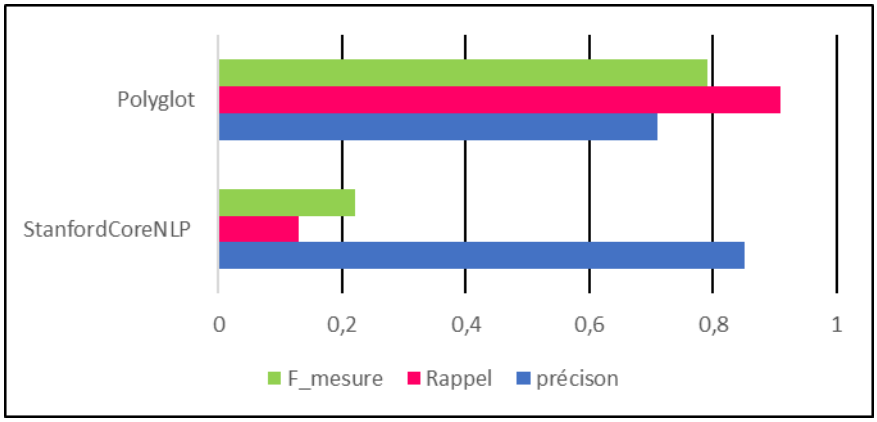

Figure 34 : résultats d'Evaluation sur le corpus arabe

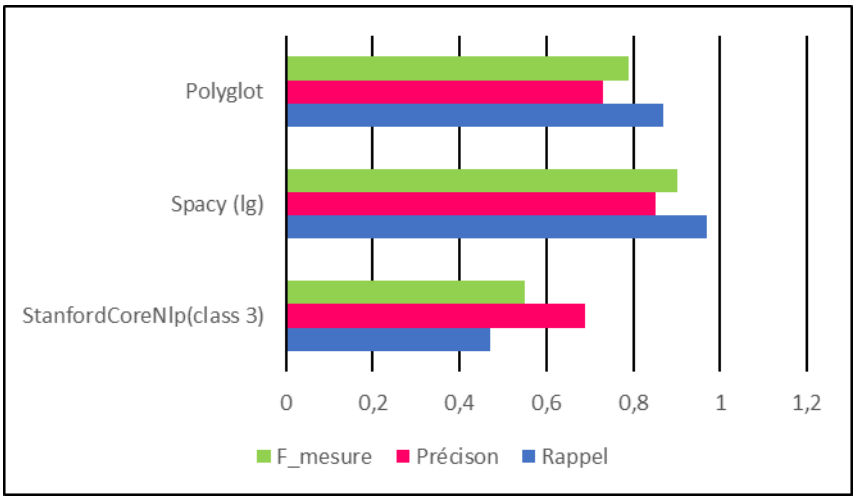

Figure 35 : Evaluation des méthodes sur un corpus français

Discussion des résultats

D'après les résultats d'évaluation sur le corpus arabe **Polyglot** a un taux de performance (F_mesure) très élevé par rapport à **StanfordCoreNLP**,aussi on a pas eu l'occasion de tester **SpaCy** car il ne prend pas en charge la langue arabe (il n'existe pas un modèle SpaCy pour l'arabe). Ce qui nous a fait le choisir pour faire extraction d'entités nommées sur les articles arabe.

D'autre part les résultats d'évaluation sur le corpus français, **Polyglot** et **Spacy(lg)** ont donné le meilleur résultat en fonction des indicateurs de performance par rapport à **StanfordCoreNLP**. Alors on a proposé de faire une combinaison entre ces deux modules, afin de pouvoir améliorer les résultats en langue française et on a obtenu les résultats présentés dans *le tableau (5).*

Les résultats de cette combinaison sont plus performants que chacune employé séparément.

	Rappel	Précision	F_mesure
(SpaCy& Polyglot)	0.97	0.89	0.92

Tableau 5 : résultats d'évaluation de la combinaison

4.7.2 Évaluation de TextBlob pour l'analyse des sentiments

Notre évaluation concernera de TextBlob qui est une bibliothèque fournit une API cohérente pour sentiment analysis. Dans le but améliorer les résultats de ce module on a proposé d'enrichir le lexique de mots d'opinion avec plus de mots et des expressions porteurs d'opinion (verbes, noms et les adverbes d'opinion).

Pour évaluer la performance de notre méthode basée sur **la classification à n** classes (positives, négatives et neutres). Celons **(Sebastiani, 2005)** [72] les moyennes globales de la précision et du rappel sur l'ensemble des n classes peuvent être évaluées de 2 manières :

La micro-moyenne fait d'abord la somme des éléments du calcul (a, b, et c) sur l'ensemble des n classes, pour calculer la précision et le rappel globaux.

$$précision = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i} \quad (11)$$

$$Rappel = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} c_i} \quad (12)$$

La macro-moyenne calcule d'abord la précision et le rappel sur chaque classe i, puis en fait la moyenne sur les n classes.

$$précision = \frac{\sum_{i=1}^{n} \left(\frac{a_i}{a_i + b_i}\right)}{n} \quad (13)$$

$$Rappel = \frac{\sum_{i=1}^{n} \left(\frac{a_i}{a_i + c_i}\right)}{n} \quad (14)$$

Soient :

– ai = nombre de documents correctement attribués à la classe i.

– bi = nombre de documents faussement attribués à la classe i.

– ci = nombre de documents appartenant à la classe i mais non attribués à la classe i.

– n = nombre de classes.

Nous avons utilisé aussi la F-mesure pour évaluer la performance d'une classification à n classes, on peut utiliser soit la micro-moyenne des précisions et rappels, soit leur macro moyenne. Nous avons choisi de calculer la F-mesure globale avec la macro-moyenne pour que les résultats sur chaque classe comptent de la même manière quelle que soit la taille de la classe.

Les résultats obtenus sont présentés dans le tableau ci-dessous :

	Rappel	Précision	F_mesure
TextBlob ar	0.64	0.63	0.63
TextBlob fr	0.80	0.77	0.78
TextBlob fr (avec un lexique enrichi)	0.90	0.82	0.85

Tableau 6 : résultats d'évaluation de TextBlob pour macro-moyen

Discussion des résultats

On utilise le module NLP sur python **Textblob** (ar /fr) pour faire l'analyse de tonalité.

Lors de traitement avec Textblob on a remarqué que parfois une phrase a une opinion positive, elle est détectée comme négative et vice versa, le problème est dû au module pattern (fr |ar) de Textblob qui regroupe un lexique d'adjectifs (par exemple, bon, mauvais, étonnant, irritant, ...) qui apparaissent fréquemment dans les critiques de produits, annotés avec des scores pour la polarité des sentiments (positif ↔ négatif) et la subjectivité (objectif ↔ subjectif).

La fonction sentiment () renvoie un tuple (polarité, subjectivité) pour la phrase donnée, basé sur les adjectifs qu'elle contient, donc la polarité dépend directement de l'adjectif utilisé.

Les logiciels professionnels utilisent généralement des outils complexes basés sur des réseaux de neurones et des classificateurs associés à une analyse lexicale. Alors que, TextBlob essaie juste de donner un résultat basé sur un résultat direct de l'analyse grammaticale. Dans notre cas, la polarité des adjectifs est la source du problème.

Pour éviter ce problème et améliorer les résultats d'analyse de ce dernier on a pu alimenter le lexique de **Textblob fr** en ajoutant plus de ***2525*** mots d'opinion (verbes, adverbes et d'autre adjectives), les résultats obtenus sont présentés dans le tableau *(5)*.

En ce qui concerne de **Textblob ar** on a trouvé des difficultés pour l'améliorer à cause de son implémentation qui est un peu différente par rapport à **Textblob fr** (tout d'abord, le texte arabe est traduit en texte anglais à l'aide de GoogleTranslate API ensuite le sentiment est alors calculé par Pattern Analyzer).

4.8 Conclusion

A travers ce chapitre nous avons présenté les outils de développement, les bibliothèques utilisées et la collection des sources que nous avons utilisées ainsi que les différentes interfaces graphiques à travers lesquelles nous pouvons superviser les différentes fonctionnalités du système. Le résultat obtenu est satisfaisant mais il reste certaines améliorations à faire surtout pour la langue arabe.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Dans ce mémoire notre objectif a été de développer un système de classification basé sur l'exploration de texte par les techniques de Data Mining et de TAL. Cela dans le but de réaliser une analyse de la tonalité exprimée envers des entités cibles. Pour cela, nous avons en premier lieu exposé une étude bibliographique concernant la reconnaissance des entités nommées du texte, ou nous avons énoncé les différentes méthodes appliquées dans ce contexte et une synthèse des travaux existants. Ensuite, nous avons effectué une étude bibliographique sur les différentes approches d'analyse de tonalité et une synthèse des travaux réalisés dans le cadre de la tonalité véhiculée dans le texte. Enfin, nous avons étudié les systèmes de classification et d'analyse de tonalité existants pour une analyse de l'actualité.

Après l'étude théorique, nous somme passées à la phase de contribution. Nos travaux dans cette étape peuvent être divisés en deux grandes parties : la première partie englobe la phase de conception ou nous avons proposé l'architecture globale du système en décrivant les différentes fonctionnalités proposées. Dans la seconde partie, nous avons utilisé le langage python pour implémenter notre système. Enfin, des expérimentations et des évaluations du système sur un échantillons de données des différentes sources d'actualité écrite en arabe et d'autres en français ont été faites en vue de tester l'efficacité de notre système.

A l'issu de ce travail, nous avons rencontré quelques problèmes principalement lors de la phase de la recherche documentaire sur l'analyse de tonalité des textes de l'actualité. La plupart des travaux de recherche se focalisent sur tout ce qui est fouille d'opinion/sentiment dans les réseaux sociaux ou la nature des textes est complétement différente de ceux de l'actualité en ligne. L'analyse des sentiments est une tâche complexe et difficile à l'étudier à cause de la subjectivité et transparence de certains discours (humour, ironie, sens caché, etc..).

La plupart des modèles pré-entraines sont basés sur la langue anglaise alors que nous travaillons à la fois sur l'arabe et le français, nous avons fait face à un manque de ces derniers afin de pouvoir tester les méthodes basées sur l'apprentissage supervisé. A travers ce travail, nous avons pu atteindre nos objectifs fixés au préalable. Nous avons obtenu des résultats encourageants et qui peuvent être considérablement améliorés. Par ailleurs, arrivés à terme de ce projet, nous pouvons présenter quelques améliorations possibles à notre solution :

- Créer des corpus annotés basés sur la langue arabe pour la détection des entités nommées.

- Enrichir le lexique de mots d’opinion avec plus de mots et des expressions porteurs d’opinion (verbes et les adverbes d’opinion) dans le but de mieux analyser les informations diffusées par les sources médias algériennes.
- Expérimenter des approches hybrides pour l’analyse de tonalité dans le but d’avoir de meilleurs résultats.

Bibliographie

[1] Entités nommées.Technolangue.net. URL http://www.technolangue.net Consulté le : 27/12/2019.

[2] POIBEAU, Thierry et NAZARENKO, Adeline. L'extraction d'information, une nouvelle conception de la compréhension de texte ? TAL. Traitement automatique des langues, vol. 40, no 2, p. 87-115,1999.

[3] Flitti sarah, identification automatique des entités nommées, mémoire de fin d'étude Pour l'obtention du diplôme master, 2016.

[4] Mohamed Hatmi. *Reconnaissance des entités nommées dans des documents multimodaux*, Informatique. THÈSE DE DOCTORAT, 2014.

[5] MCDONALD, D. D. Internal and external evidence in the identification and semantic categorization of proper names. In Corpus processing for lexical acquisition, pages 21–39. MIT SPress, Cambridge, MA, USA. (1996).

[6] Farber B; Freitag D; Habash N and RambowO; Improving NER in Arabic using a morphological tagger. Proceedings of workshop on HLT & NLP within the Arabic world ,2008.

[7] AbdelRahman S; Elarnaoty M; Magdy M and Fahmy A.Integrated ,machine Learning techniques for Arabic named entity recognitin , vol. 7, p. 27-36,2010.

[8] lakel kheira ; les annotations sémantiques dans les documents web : application au textes psychologiques et langue arabe ; thèse doctorat université Oran ;2017 .

[9] Meryem Talha ,Siham Boulaknadel and Driss Aboutajdine, Système de reconnaissance des entités nommées amazighes ; université Mohammed V-Agdal Rabat 4. 2014.

[10] Azeddine Zidouni, « Modèle graphique discriminants pour l'étiquetage de séquences, Application à la reconnaissance d'entités nommées radiophonique », thèse de doctorat en l'informatique, 2010.

[11] ZRIBI, Inès, HAMMAMI, Souha Mezghani, et BELGUITH, Lamia Hadrich. L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe. In : TALN.p. 59,20210.

[12] Poibeau Thierry. «Deconstructing Harry, une évaluation des systèmes de repérage d'entités nommées». *Revue de la société d'électronique, d'électricité et de traitement de l'information,2001*.

[13] Mansouri Alireza, Affendey Lilly-Suriani et Mamat Ali. «Named Entity Recognition Approaches». IJCSNS International Journal of Computer Science and Network Security, Vol.8

P.339-344, 2008.

[14] Shaalan, K. and Raza, H. NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology 2009.

[15] Zaghouani, W., Pouliquen, B., Ebrahim, M. and Steinberger, R. Adapting a resource-light highly multilingual named entity recognition system to Arabic.conference on International Language Resources and Evaluation ,2010.

[16] Al-Ahmari, S. S. and Al-Johar, B. A. Cross domains Arabic named entity recognition system, Proceedings of SPIE, First International Workshop on Pattern Recognition,2016.

[17] Friburger, N,Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques. Thèse de doctorat. Université François Rabelais Tours, 2002.

[18] Abuleil Saleem. «Extracting Names from Arabie Text For QuestionAnswering Systems». In Proceedings of the 7th International Conference on Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval. Avignon (Vaucluse), France,P.638-647,2004.

[19] Abuleil Saleem et Evens Martha. «Extracting an Arabie Lexicon from Arabie Newspaper Text». Computers and the Humanities, Vol.36(2), P.191-221. 2002.

[20] FOUROUR, N. Apport du web dans la reconnaissance des entités nommées. In Revue Québécoise de Linguistique (RQL), pages 41–60. (2003).

[21] Benajiba Y assine, Rosso Paolo et Benedi Ruiz José Miguel. «ANERsys: An Arabie Named Entity Recognition System Based on Maximum Entropy». In Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, février , P.143-153,2007.

[22] Benajiba Yassine et Rosso Paolo. «ANERsys 2.0: Conquering the NER task for the Arabie language by combining the maximum entropy with POS-tag information». *Indian International Conference on Artificial Intelligence, 2007.*

[23] Benajiba Yassine et Rosso Paolo. «Arabic named entity recognition using conditional random fields». *Conference on Language Resources and Evaluation ,2008.*

[24] Lafferty John, McCallum Andrew et Pereira Fernando. «Conditional random fields: Probabilistic models for segmenting and labeling sequence data». *International Conference on Machine Leaming , 2001* .

[25] Mohammed Naji F. et Nazlia Omar. «Arabic Named Entity Recognition Using Artificial Neural Network ». *Journal of Computer* Science,2012.

[26] Gahbiche-Braham Souhir, Bonneau-Maynard Hélène, Lavergne Thomas et Yvon François. « Repérage des entités nommées pour l'arabe : adaptation nonsupervisée et combinaison de systèmes ». *The conference of JEPTALN- RECITAL, Grenoble, France,2012.*

[27] Gahbiche-Braham Souhir, Bonneau-Maynard Hélène et Yvon François. « Traitement automatique des entités nommées en arabe : détection et traduction ». *TAL (Traitement Automatique des Langues), 2014.*

[28] Lavergne Tomas, Cappé Olivier et Yvon François. «Practical Very Large Scale CRFs». *In Proceedings of 48th Annual Meeting Association for Computational Linguistics (ACL), Uppsala, Sweden, juillet ,2010*

[29] Oudah Mai et Shaalan Khaled. «Person Name Recognition Using the Hybrid Approach». *Natural Language Processing and Information Systems,2013.*

[30] B. Martins, H. Manguinhas et J. Borbinha. « Extracting and Exploring the Geo-Temporal Semantics of Textual Resources». In: Proceedings of the International Conference on Semantic Computing. IEEE Computer Society, août 2008.

[31] Benveniste, E. Subjectivity in language, in: Problems in General Linguistics. pp. 223-230;1971.

[32] Finegan, E. Subjectivity and subjectivisation: an introduction, in: Subjectivity and Subjectivisation: Linguistic Perspectives. Cambridge: Cambridge University Press, pp. 1–15, 1995.

[33] Wiebe J ,Wilson T Bruce, R Bell M et Martin M., Learning Subjective Language. Comput. Linguist ,.2004.

[34] Schumaker RP et Chen H; Evaluating Sentiment in Financial News Articles. J. Am. Soc. Inf. Sci. Technol. 59, 247–255,2012.

[35] LIU, Bing, *et al.* Sentiment analysis and subjectivity. *Handbook of natural language processing*, p. 627-666,2010.

[36] Pang B, Lee L; Opinion mining and sentiment analysis, in: Foundations and Trends® in Information Retrieval. pp. 1–135, 2008.

[37] Kim S-M et Hovy E; Determining the sentiment of opinions, in: Proceedings of the 20th International Conference on Computational Linguistics - COLING '04. Presented at the the 20th international conference, Association for Computational Linguistics, 2004.

[38] Kobayashi N, Inui K, Matsumoto Y; Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. Proc. Jt Conf Empir Methods Nat Lang Process. Comput Nat Lang Learn10, 2007.

[39] LIU, Bing et MINING, W. D. Exploring Hyperlinks, Contents, and Usage Data. *Web Data*

Mining, ser. Data-Centric Systems and Applications. Springer Berlin Heidelberg, 2007.

[40] Jindal Nitin and Bing Liu. Mining comparative sentences and relations. In Proceedings of National Conf. on Artificial Intelligence (AAAI-2006), 2006.

[41] LIU, Bing. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, vol. 5, no 1, p. 1-167,2012.

[42] PAK, Alexander. Automatic, adaptive, and applicative sentiment analysis. Thèse de doctorat.2012.

[43] Beghdad abdlekrim et Ousrir amina, une approche deep learning pour l'analyse des sentiments sur twiter, mémoire fin d'etude université BOUNAAMA_KHemis Miliana, 2017.

[44] Haseena Rahmath P; Opinion Mining and Sentiment Analysis - Challenges and Applications, Dept. of Computer Science and Engineering, May 2014.

[45] Lazhar FAREK, Identification d'opinions dans les textes arabes en utilisant les ontologies, Thèse de doctorat, 2014.

[46] GARCIA-FERNANDEZ, Anne et FERRET, Olivier. Etude de différentes stratégies d'adaptation à un nouveau domaine en fouille d'opinion (Study of various strategies for adapting an opinion classifier to a new domain)[in French]. In : Proceedings of the Joint Conference JEP-TALN-RECITAL , volume 2: TALN. p. 391-398,2012.

[47] BOULLIER, Dominique et LOHARD, Audrey. Opinion mining et Sentiment analysis: Méthodes et outils. OpenEdition Press, 2012.

[48] FAIZA Belbachir, Recherche de l'Université Paul Sabatier Toulouse, Expérimentation de fonctions pour la détection d'opinions dans les blogs, juin2010,

[49] Mekki Abir, Détection automatique des sentiments dans les réseaux sociaux, mémoire présenté pour l'obtention du diplome de master académique, 2017.

[50] Liu, Bing. "Sentiment Analysis and Opinion Mining Morgan & Claypool Publishers." Language Arts & Disciplines 167 ,2012.

[51] Akila GHERSEDINE, Patrice BUCHE, Juliette DIBIE-BARTH_EL_EMY, Nathalie HERNANDEZ et Mouna KAMEL, IRIT-IC3, Extraction de relations n-aires inter phrastiques guidée par une RTO, 2012.

[52] Chabbou fatima zohra et BAKHOUCHE souhaila, fouille d'opinions méthodes et outils, mémoire master, université de Larbi Tébessi_Tébessa, 2016.

[53] MEDHAT, Walaa, HASSAN, Ahmed, et KORASHY, Hoda. Sentiment analysis algorithms and applications : A survey. Ain Shams engineering journal, vol. 5, no 4, p. 1093-1113,2014.

[54] Ku L.-W, Liang Y-T et Chen H.-H, Opinion Extraction Summarization and Tracking in News and Blog Corpora. Proc. AAAI 100–107,2006.

[55] GODBOLE, Namrata, SRINIVASAIAH, Manja, et SKIENA, Steven. Large-Scale Sentiment Analysis for News and Blogs. Icwsm, vol. 7, no 21, p. 219-222,2007.

[56] BALAHUR, Alexandra, STEINBERGER, Ralf, KABADJOV, Mijail, et al. Sentiment analysis in the news. arXiv preprint arXiv:1309.6202, 2013.

[57] Steinberger R, Taney H et Della Rocca L, Large-scale news entity sentiment analysis, - Recent Advances in Natural Language Processing Meet Deep Learning, pp. 707–715, 2017.

[58] SCHUMAKER, Robert P., ZHANG, Yulei, HUANG, Chun-Neng, et al. Evaluating sentiment in financial news articles. Decision Support Systems, vol. 53, no 3, p. 458-464,2012.

[59] Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S. OpinionFinder: a system for subjectivity analysis, in: Proceedings of HLT/EMNLP on Interactive Demonstrations pp. 34–35,2005.

[60] Li X, Xie H, Chen L, Wang J et Deng X, News impact on stock price return via sentiment analysis. Knowl.-Based Syst. 69, 14–23, 2014.

[61] Presentation python.opentuto.com URL .http://www.opentuto.com/presentation-de-python/ consulté le 08/8/2020.

[62] PHP.espresso-jobs URL https://espresso-jobs.com/ consulté le 5/8/2020.

[63] Apache HTTP.hostinger.fr URL https://www.hostinger.fr/tutoriels/quest-ce-quapache-serveur-web-apache/ consulté le 4/8/2020.

[64] My sql.oracle.com URL https://www.oracle.com/dz/database/what-is-a-relational-database/ consulté le 4/8/2020.

[65] HTML.journaldunet.fr URL https://www.journaldunet.fr/web-tech/ consulté le 5/8/2020.

[66] CSS.schoolmouv.fr URL https://www.schoolmouv.fr/cours/la-page-web-http-et-langages-html-et-css-/ consluté le 5/8/2020.

[67] Bootstrap.w3tutoriels URL https://w3tutoriels.com/bootstrap/ consulté le 5/8/2020.

[68] Feedparser.journaldunet.com URL https://www.journaldunet.com/ consulté le 5/8/2020.

[69] Polyglot.maelfbien.io URL https://maelfabien.github.io/ consulté le 6/8/2020.

[70] Spacy.labsomars.com URL https://labsonmars.com/python-nlp/ consulté le 6/8/2020.

[71] Textblob.readthedocs.io URL https://textblob.readthedocs.io/en/dev/ consulté le 6/8/2020.

[72] Sebastiani F; Text Mining and its Aplications to Intelligence, CRM and Knowledge Management, pp. 109– 129, WIT Press, 2005.