UNIVERSITE ASSANE SECK-ZIGUINCHOR

L'Excellence, ma référence

****

Ecole Doctorale Sciences, Technologies et Ingénierie (ED STI)

Laboratoire d'Informatique et d'Ingénierie pour l'Innovation (LI3)

THESE DE DOCTORAT

Présentée par :

Lamine FATY

Vers un système de fouille d'opinions dans les commentaires de la presse en ligne : cas du Sénégal

Spécialité : Base de données, Fouille de données et Technologies web

Soutenue le 19 décembre 2020

Devant le jury composé :

M. Amadou COULIBALY	Pr, Université de Strasbourg	Président
M. Cheikh Talibouya DIOP	Pr, Université Gaston Berger de Saint-Louis	Rapporteur
M. Gayo DIALLO	HDR, Université de Bordeaux	Rapporteur
M. Mamadou BOUSSO	MCF, Université de Thiès	Rapporteur
M. Khalifa GAYE	MCF, Université Assane Seck de Ziguinchor	Examinateur
M. Mouhamadou Lamine BA	MA, Université Alioune Diop de Bambey	Examinateur
M. Alassane DIEDHIOU	Pr, Université Assane Seck de Ziguinchor	Directeur de thèse
Mme. Marie NDIAYE	MA, Université Assane Seck de Ziguinchor	Co-Encadrant

DEDICACES

Je dédie ce travail à ma femme et ma fille.

REMERCIEMENTS

Ce mémoire est le fruit d'une longue et lente maturation et nous tenons à remercier tous ceux qui nous ont inspirés et aidés dans cette démarche et particulièrement les personnes qui ont cru en moi et m'ont permis d'arriver au bout de cette thèse.

- ✓ Je remercie le Professeur Alassane DIEDHIOU pour avoir dirigé ma thèse.
- ✓ Je remercie Docteure Marie NDIAYE DIOP pour la confiance en me proposant de travailler avec elle sur ce sujet. Ses critiques constructives ont été d'un apport déterminant pour l'aboutissement de ma thèse.
- ✓ Je remercie le Professeur Amadou COULIBALY d'avoir accepté de présider ma thèse.
- ✓ Je remercie les Professeurs Cheikh Talibouya DIOP, Gayo DIALLO et Mamadou BOUSSO d'avoir accepté de rapporter ma thèse.
- ✓ Je remercie également le Professeur Khalifa GAYE et Docteur Mouhamadou Lamine BA d'avoir accepté de participer au jury de ma thèse.
- ✓ Je remercie les Docteurs Ibrahima DIOP et Khadim DRAME, enseignants chercheurs au département d'Informatique de l'UASZ pour leur soutien et m'ont facilité la compréhension de la problématique de commentaires journalistiques sénégalais.
- ✓ Je remercie aussi le Laboratoire d'Informatique et d'Ingénierie pour l'Innovation (LI3) de l'UASZ d'avoir soutenu ce travail.
- ✓ Je remercie le Professeur Ousmane SALL, Chef de projet Check4Decision de l'Université de Thiès pour m'avoir associé à son équipe-projet de recherche.
- ✓ Je remercie l'administration de l'UFR des Lettres, Arts et Sciences Humaines en particulier d'avoir financé et soutenu ce travail.
- ✓ Je remercie mes collègues doctorants, mes collègues de l'UASZ et de Check4Decision pour leur collaboration dans la réalisation de ma thèse.
- ✓ Je tiens à remercier particulièrement mes parents pour leur soutien.
- ✓ Mention spéciale à ma femme, ma fille, ma famille et mes amis pour leur soutien et leurs encouragements.
- ✓ Je remercie enfin tous ceux qui, de près ou de loin, ont contribué à la réussite de ce travail.

RÉSUMÉ

L'avènement du journalisme web 2.0 offrent aux lecteurs la possibilité de donner leurs avis sur différentes publications. Ainsi, les sites d'informations se transforment progressivement en lieu public de discussion de questions d'actualité concernant les préoccupations des populations.

Par conséquent, les commentaires issus de ces sources contiennent des informations précieuses dont l'analyse peut permettre de déterminer l'opinion globale des lecteurs sur un article ou un aspect d'un article publié.

La fouille d'opinion se présente comme l'outil par excellence pour valoriser les commentaires en ligne. Elle consiste à analyser des contenus textuels issus d'échanges en ligne en vue de mettre en évidence les opinions des internautes par rapport à une entité. Les outils (ressources et méthodes) de fouille d'opinions proposés dans la littérature sont adaptés aux textes rédigés dans des langues officielles comme l'anglais et le français. Cependant, l'hétérogénéité des sources par la différence de leur DOM (Document Object Model) d'une part et l'utilisation du langage urbain d'autre part rendent complexe le traitement des commentaires issus de la presse sénégalaise en ligne par les outils actuels de fouille d'opinions.

Pour faire face à cette complexité, nous proposons des ressources et méthodes permettant de recueillir les commentaires à partir de la presse sénégalaise en ligne, de les stocker et de les analyser tout en nous adaptant au langage urbain utilisé par les internautes. Six contributions sont présentées dans la thèse. La première contribution est une description de la complexité des commentaires issus de la presse sénégalaise en ligne. La deuxième contribution est une architecture d'un système de fouille d'opinions dans la presse sénégalaise en ligne qui structure les autres contributions de cette thèse. La troisième contribution est une modélisation de commentaires journalistiques pour la fouille d'opinions. La quatrième contribution est un outil d'acquisition, de catégorisation et de stockage des données en provenance de la presse en ligne. La cinquième contribution est un lexique bilingue constitué sur la base du langage urbain pour la fouille d'opinions. La dernière contribution est une méthode de fouille d'opinions adoptée pour l'étiquetage et la classification d'opinions sur les commentaires de la presse sénégalaise en ligne.

Mots-clés : Presse en ligne, Data journalisme, Fouille d'opinions, Modélisation de commentaire, Web Scraping, Étiquetage d'opinions, Classification d'opinions

ABSTRACT

Since the advent of web 2.0 journalism, news sites are becoming more and more attractive. These portals offer readers the opportunity to give their opinions on various publications. They are gradually transforming themselves into a public place for discussing current issues about the concerns of the population. As a result, comments from these sources contain promising information, the analysis of which can help determine the overall opinion of readers on an article or one aspect of a published article.

Opinion mining is emerging as the tool of choice for enhancing online commentary. It consists of analyzing textual content from online exchanges in order to highlight the opinions of Internet users in relation to an entity. Opinion mining tools (resources and methods) offered in the literature are adapted to texts written in official languages such as English and French. However, the heterogeneity of the sources due to the difference of their DOM (Document Object Model) on the one hand, and the use of urban language on the other hand make the treatment of comments from the Senegalese online press complex using current opinion mining tools.

To face this complexity, we propose resources and methods to collect comments from the Senegalese online press and analyze them while adapting to the urban language of Internet users. Six contributions are presented in the thesis. The first contribution is a formalization of the complexity of comments from the Senegalese online press. The second contribution is an architecture of a system for searching opinions in the Senegalese online press. The third contribution is a modeling of journalistic comments for opinion mining. The fourth is a tool for the acquisition, categorization and storage of data from the online press. The fifth is a bilingual lexicon based on the urban language for opinion mining. The last is an opinion mining method adopted for the tagging and classification of opinions on comments from the Senegalese online press.

Keywords: Online press, Data-journalism, Opinion mining, Commentary modeling, Web Scraping, Opinion labeling, Opinion Classification

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ILLUSTRATIONS

1 -INTRODUCTION GÉNÉRALE

1.1 - Contexte justificatif

1.1.1 - Contexte général

Pendant longtemps, l'enquête d'opinions est considérée comme la méthode classique pour déterminer l'opinion publique sur une situation particulière. Cette méthode consiste d'abord à interroger de vives voix des personnes sélectionnées dans différentes classes sociales appelées échantillon afin de recueillir leurs points de vue, avis ou opinions sur un phénomène ou un aspect d'un phénomène à travers un questionnaire. Ce dernier se présente sous forme de formulaire constitué d'une série de questions formalisées et adaptées aux circonstances de l'espace et du temps. Ensuite, les questions sont soumises à l'appréciation des populations ciblées. Enfin, les données recueillies sont traitées à l'aide d'outils d'analyse statistique simples afin de donner une image de l'objet d'étude inaccessible par la simple perception du chercheur qui souhaite l'appréhender.

L'avènement des technologies web 2.0 a entrainé de nombreuses mutations dans le processus de production d'informations. Le web 2.0 désigne l'ensemble de techniques, fonctionnalités et usages du web permettant la mise en ligne de l'information et la participation des utilisateurs à la production d'informations. Dans ce processus, les mutations s'opèrent à plusieurs niveaux. Entre autres, nous avons le changement de supports (du format papier au format numérique), l'accélération de la production et de la circulation de l'information à travers l'internet, l'accès immédiat à l'information et qui est quasi incontrôlable et la démocratisation de la production d'informations avec la participation des lecteurs.

Avec cette possibilité offerte, nous assistons à un foisonnement de commentaires issus des échanges à travers les réseaux sociaux, les sites de commerces, les sites dédiés à l'information, etc. Les discussions constituent d'immenses opportunités pour une organisation moderne du fait de la richesse de l'information qu'elles véhiculent et de leur abondance. Elles sont souvent exploitées pour déterminer l'opinion majoritaire des intervenants. Dans cette lancée, la fouille d'opinions [1][2] constitue l'outil le plus approprié pour valoriser les commentaires en ligne. La fouille d'opinions est un processus composé de trois grandes phases à savoir l'acquisition de données, le prétraitement de données et l'analyse d'opinions. L'acquisition de données consiste à interroger des sources web identifiées afin d'en extraire des informations de manière automatique. Le prétraitement regroupe les tâches de nettoyage, de formatage et d'uniformisation des données dans le but de pouvoir appliquer des techniques d'analyse. Enfin, l'analyse d'opinions permet de catégoriser un point de vue en fonction de

« favorable » ou « défavorable » à l'égard d'une entité ciblée (produit, service, phénomène, évènement, article journalistique, etc.).

1.1.2 - Contexte applicatif

Le travail que nous avons mené dans le cadre de cette thèse trouve son contexte d'application dans le domaine de la presse sénégalaise en ligne.

La presse en ligne s'inscrit pleinement dans la tradition journalistique consistant à aller chercher de l'information brute pour la présenter de manière adéquate au public. Ce type de médias diffuse l'information à travers des portails web dédiés. Nous avons plusieurs types de presse en ligne au Sénégal, entre autres les sites dédiés à l'information, les radios et les télévisions numériques. L'avantage de la presse sénégalaise en ligne réside dans la diffusion d'informations en *streaming* (diffusion d'informations en temps réel) ou *podcasting* (c'est-à-dire enregistrer et laisser l'information à la disposition du public pendant un certain temps). En plus, ces portails offrent aux lecteurs la possibilité de donner leurs avis sur les publications. Au Sénégal, les sites dédiés à l'information sont nombreux et présentent des informations dans diverses structures arborescentes.

En outre, les lecteurs ne sont plus de simples consommateurs d'informations, mais participent, de façon dynamique, à la génération d'informations qui pourraient intéresser le public. En effet, la participation des lecteurs à la discussion a rendu ces portails journalistiques de plus en plus attrayants. Ce type de communication favorise l'expression d'opinions, de goûts ou d'attentes souvent contrôlés ou réprimés dans le cadre d'enquêtes d'opinions [3]. En plus, les commentaires issus des échanges entre internautes contiennent beaucoup d'informations intéressantes qui peuvent être exploitées pour l'aide à la décision. En réalité, les opportunités d'extraction de connaissances utiles à partir des commentaires en ligne sont innombrables. Par conséquent, ces commentaires sont considérés comme une source privilégiée pour déterminer l'opinion d'internautes.

Cependant, les lecteurs rédigent leurs commentaires dans un langage naturel libre qu'on appelle langage urbain sénégalais. Le langage urbain sénégalais modifie les caractéristiques orthographiques, voire grammaticales d'une langue afin de réduire sa longueur. D'une part, ce langage comporte des expressions issues de plusieurs langues (étrangères et nationales). D'autre part, il possède une syntaxe et un vocabulaire propre, différents du langage écrit « standard ». Depuis quelque temps, la nouvelle génération sénégalaise a adopté l'utilisation de ce langage

dans les conversations quotidiennes tant dans le monde réel que virtuel. Malheureusement, il n'existe pas encore d'outils (ressources et méthodes) pour le traitement automatique des données textuelles contenant ce langage urbain.

Dans la littérature, des solutions de fouille d'opinions ont été proposées et ne cessent d'évoluer [4][5][6]. Elles sont destinées à analyser des textes écrits dans des langues officielles (françaises, anglaises, portugaises, etc.). La fouille d'opinions sur les commentaires issus de la presse sénégalaise en ligne fait face à la complexité de ces commentaires qui constitue la problématique de cette recherche.

1.2 - Problématique et objectifs de recherche

1.2.1 - Problématique

La complexité des commentaires issus de la presse sénégalaise en ligne est liée à plusieurs problèmes tels que l'ambiguïté des données, le multilinguisme, le multi-domaine et l'hétérogénéité des structures DOM (Document Object Model).

- ***Problème de l'ambigüité*** : L'ambigüité est la propriété d'un mot, groupe de mots, d'un concept ou d'une phrase ayant plusieurs significations ou plusieurs analyses grammaticales possibles. Elle peut être considérée comme des propriétés intrinsèquement associées à une phrase prise hors contexte et qui provoque diverses interprétations. L'ambigüité dans les commentaires de la presse sénégalaise en ligne est liée à la structuration syntaxico-sémantique de l'énoncé qui peut être regroupée en trois (03) types à savoir les ambigüités syntaxique, sémantique et lexicale.
- ***Problème de multilinguisme*** : À côté de l'ambiguïté, nous avons constaté aussi dans ces commentaires le multilinguisme. En parlant de multilinguisme, nous faisons allusion de la cohabitation de langues étrangères et nationales dans les commentaires. Ce phénomène résulte en particulier, la croissante composition multi-ethnique et multiculturelle de notre pays. Dans ce pays, vingt-cinq (25) langues nationales cohabitent avec le français, l'arabe et d'autres langues étrangères[1]. Ces langues nationales s'imposent de plus en plus dans les commentaires en ligne.
- ***Problème de multi-domaine*** : la notion de multi-domaine renvoie à la notion de « hors-contexte » ou « hors-sujet ». Pour nous, tout commentaire qui ne porte pas sur la même

[1] Selon la Direction de l'Alphabétisation et des Langues Nationales au Sénégal

thématique que l'article auquel il est associé est considéré comme un commentaire « hors-contexte ». En effet, l'analyse d'opinions sur des données contenant plusieurs entités nécessite une attention particulière, car il existe des mots ou groupes de mots dont les sens dépendent de leur domaine. Certains domaines ont des vocabulaires intrinsèquement positifs ou négatifs.

- ***Hétérogénéité des structures DOM (Document Object Model)*** : Le DOM est un modèle d'objets de document (page web) qui est utilisé par W3C[2] comme le modèle standard de structuration d'un document HTML[3] ou XML[4]. Ce modèle définit la composition d'une page web sous forme d'arborescence. Compte-tenu de l'hétérogénéité des structures DOM, l'acquisition des articles et des commentaires associés, devient l'une des principales difficultés qui nécessite une solution adaptée.

1.2.2 - Objectifs de recherche

Face à cette situation complexe, notre recherche a pour objectif général de proposer un système de fouille d'opinions doté de ressources et de méthodes performantes afin de rendre les commentaires de la presse sénégalaise en ligne accessibles et intelligibles. Pour atteindre cet objectif général, nous l'avons subdivisé en plusieurs sous objectifs spécifiques :

- la formalisation de la complexité des commentaires issus de la presse sénégalaise en ligne ;
- la proposition d'une architecture de système de fouille d'opinions pour la représentation et l'évolution des connaissances, la recherche d'opinions, l'intégration des ressources et l'aide à la décision ;
- la modélisation d'un commentaire journalistique et d'un réseau de commentaires pour faciliter la collecte, la fusion, la catégorisation et le stockage, mais aussi pour éviter la complexité de la fouille d'opinions basée sur les aspects ;
- la proposition de méthodes d'acquisition de données pour constituer le corpus ;
- la mise en place de ressources linguistiques notamment un lexique d'opinions, une ontologie d'évènements et un corpus d'entrainement pour permettre l'apprentissage automatique et approfondi ;

[2] http://www.standard-du-web.com/world_wide_web_consortium.php
[3] [3] https://fr.wikipedia.org/wiki/Hypertext_Markup_Language
[4] https://fr.wikipedia.org/wiki/Extensible_Markup_Language

- la proposition de méthodes de traitement automatique pour la fouille d'opinions, la catégorisation selon les entités nommées, etc. ;
- la construction d'une base de connaissances pour la recherche sémantique d'informations ;
- la proposition de méthodes de visualisation pour rendre les résultats accessibles et attrayants.

1.3 - **Organisation du manuscrit**

Ce document est structuré en deux grandes parties à savoir l'état de l'art et nos contributions.

Dans la partie état de l'art, nous présentons les méthodes d'analyse de données textuelles de manière générale et ses domaines d'applications avec l'abondance de données non structurées. Dans le même sillage, nous mettons en évidence des approches de fouille d'opinions et leurs limites vis-à-vis de commentaires issus de la presse sénégalaise en ligne. Cette partie s'étend sur deux (02) chapitres : de la fouille de textes à la fouille d'opinions et l'état de l'art de la fouille d'opinions.

Nos contributions vont dans le sens de la mise en place d'un système de fouille d'opinions pour valoriser les commentaires issus de la presse sénégalaise en ligne. Pour cela, nous avons proposé des ressources et méthodes pour analyser ces commentaires. Cette partie décrit les méthodes et les implémentations de nos solutions c'est-à-dire les procédures utilisées, les expérimentations et les résultats de l'analyse. Les contributions sont étalées sur quatre (04) chapitres à savoir : (i) la modélisation de commentaires journalistiques pour la fouille d'opinions, (ii) l'architecture d'un système de fouille d'opinions dans la presse sénégalaise en ligne, (iii) l'acquisition de commentaires journalistiques en vue de la fouille d'opinions et (iv) vers un lexique (bilingue) français-wolof pour l'étiquetage d'opinions.

En somme, ce présent travail est subdivisé hormis l'introduction générale et la conclusion générale en six (06) chapitres répartis comme suit :

- **Chapitre 2 : De la fouille de textes à la fouille d'opinions**

Ce chapitre traite le processus général de fouille de textes c'est-à-dire l'acquisition de données, le prétraitement et l'application des méthodes d'analyse de données. Il montre aussi les nouvelles disciplines qui sont apparues pour le traitement automatique des données non structurées en général et de données textuelles en particulier. Ce chapitre a un double objectif :

d'une part, il définit les concepts et la démarche méthodologique pour analyser une collection de documents textuels ; d'autre part, il met en exergue une démarcation de la fouille d'opinions par rapport aux autres disciplines issues de la fouille de textes.

- **Chapitre 3 : État de l'art sur la fouille d'opinions**

Ce chapitre met l'accent sur les approches de fouille d'opinions qui sont l'approche lexicale, l'approche par apprentissage et l'approche hybride. Il montre également différentes solutions proposées dans la littérature basées sur ces approches. La complexité des commentaires issus de la presse sénégalaise en ligne y est mise en lumière. Nous proposons, d'une part, une discussion dans laquelle nous expliquons comment cette complexité constitue des obstacles face aux solutions existantes. D'autre part, nous comparons les approches afin de guider notre choix.

- **Chapitre 4 : Modélisation de commentaires journalistiques pour la fouille d'opinions**

Ce chapitre expose la modélisation d'un commentaire journalistique et celle d'un réseau de commentaires. En d'autres termes, il s'agit essentiellement de proposer une représentation formelle d'un commentaire journalistique et une description des relations entre les commentaires. L'objectif de la modélisation de commentaires est double. Elle permet, d'une part, de faciliter la représentation de données et d'autre part, de faire de la fouille d'opinions basée sur les aspects.

- **Chapitre 5 : Architecture d'un système de fouille d'opinions dans la presse sénégalaise en ligne**

Le chapitre portant sur l'architecture globale de notre système de fouille d'opinions décrit les différents modules du système et l'interaction entre ces modules. Dans ce chapitre, nous exposons la cartographie de sites dédiés à l'information au Sénégal. L'objectif visé dans ce travail est de circonscrire les périmètres de notre système afin de proposer des solutions idoines.

- **Chapitre 6 : Acquisition de commentaires journalistiques en vue de la fouille d'opinions**

Le chapitre 6 présente OpinionScraper qui est un outil d'acquisition et de catégorisation de commentaires journalistiques en vue de la fouille d'opinions. La mise en place de cet outil

s’inscrit dans le but de construire une base de données de manière synthétique et cohérente à partir de commentaires journalistiques en séparant les entités et leurs aspects.

- **Chapitre 7 : Vers un lexique bilingue (français-wolof) pour l’étiquetage d’opinions**

Ce chapitre présente notre outil SenOpinion qui est un lexique bilingue d’opinions (bilingue-wolof) destiné à l’étiquetage d’opinions de commentaires journalistiques sénégalais. Dans ce chapitre, nous décrivons d’abord le processus de conception de SenOpinion. Ensuite, nous proposons des procédures pour son utilisation et un modèle mathématique pour le calcul de score. Enfin, nous expliquons les résultats d’expérimentation avec cet outil.

2 -DE LA FOUILLE DE TEXTES A LA FOUILLE D'OPINIONS

2.1 - **Introduction**

Généralement, l'analyse de grandes quantités de données est l'apanage d'outils de fouille de données [7][8]. La fouille de données, terme que l'on traduit en anglais par *Data Mining*, est un ensemble de techniques issues de Statistiques et du Machine Learning qui met en jeu un processus automatisé d'exploitation de données dans l'optique de découvrir des connaissances jusqu'alors inconnues. Elle permet de restituer l'essentiel de l'information utile. Ici, il est important de préciser que les données habituellement exploitées en fouille de données sont bien structurées c'est-à-dire qu'elles sont présentées sous forme de tableaux dont les colonnes et les lignes représentent respectivement des attributs et individus statistiques.

L'informatisation croissante a entrainé une production fulgurante des données qu'on appelle données non structurées [9][10]. La plupart de données non-structurées nous proviennent du web [11][12]. La fouille du contenu du web peut réellement représenter un atout précieux pour les acteurs socio-économiques et politiques qui sauront en tirer profit. Elle consiste à extraire des informations utiles à partir du contenu réel de pages web. En d'autres termes, la fouille du contenu du web a pour but d'exploiter des opportunités d'extraction de connaissances utiles à partir de documents en ligne. Dans ce contexte, un document correspond à un élément de la collection (ensemble de documents) et peut être de différentes natures (texte, image, audio, vidéo, etc.). Il peut s'agir d'un document dans sa totalité ou d'une partie d'un document. Nous nous intéressons dans ce chapitre plus particulièrement aux documents de nature textuelle.

La branche de fouille de données qui s'applique aux données textuelles est la fouille de textes (*Text Mining* en anglais) [13][14]. Cette dernière regroupe l'ensemble de méthodes destinées à extraire automatiquement de nouvelles connaissances à partir de textes écrits en langage naturel. C'est un domaine de l'intelligence artificielle qui allie la linguistique, les statistiques et l'informatique. De nos jours, plusieurs disciplines héritent de la fouille de textes parmi lesquelles nous avons le Traitement Automatique du Langage Naturel (TALN), la Recherche d'Information (RI) et la Fouille d'opinions qui sont très utilisées à l'heure actuelle.

L'objectif de ce chapitre est de décrire la fouille de textes en général et la fouille d'opinions en particulier. Dans l'atteinte de cet objectif, d'abord nous décrivons le processus général de l'analyse statistique de données textuelles. Ensuite, nous présentons les disciplines qui héritent de la fouille de textes. En conclusion, nous dressons une synthèse.

2.2 - **Processus de fouille de textes**

Globalement, le processus de fouille de textes comprend trois grandes phases à savoir l'acquisition de données, le prétraitement de données et l'application de méthodes d'apprentissage. La Figure 1 schématise ces phases qui seront décrites en détail par la suite.

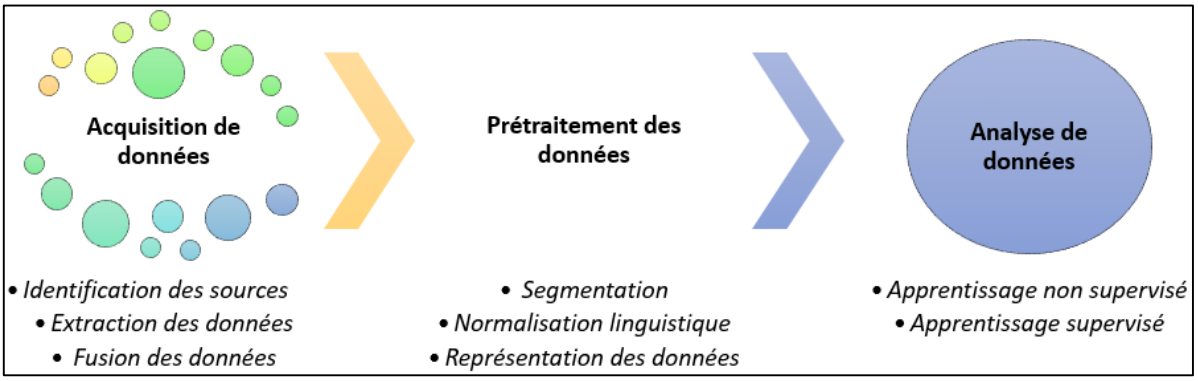


Figure 1 : *Processus de fouille de textes*

2.2.1 - Acquisition de données

Dans ce contexte, nous nous intéressons aux données provenant du web. L'acquisition de données est composée de tâches d'identification de sources, d'extraction et de fusion de données.

2.2.1.1 - Identification de sources

L'identification de sources est l'étape préliminaire indispensable surtout dans un contexte où les sources d'informations sont multiples et variées. Elle consiste à cibler des sources pertinentes et accessibles. La pertinence fait référence aux sources propices et intéressantes du point de vue des objectifs d'étude. Autrement dit, les sources ciblées doivent être suffisamment à jour par rapport aux objectifs de l'étude dans le seul but de disposer d'informations susceptibles d'apporter une réponse à la problématique posée. À cet effet, le web est organisé de telle manière que toutes les informations ne sont pas accessibles de manière équivalente. Une source techniquement accessible doit également l'être économiquement et légalement. En outre, il existe des types de données qui ne peuvent être extraits qu'à condition de s'acquitter de droits payants auprès de l'éditeur et d'autres sont jugés sensibles, leurs accès sont juridiquement encadrés.

L'intérêt de l'identification de sources est d'éviter les écueils de la surinformation ou de la désinformation, car la qualité et la fiabilité des sources ont un impact sur la qualité des résultats. Après avoir identifié les sources, vient l'extraction de données.

2.2.1.2 - Extraction de données

L'extraction de données est une opération consistant à interroger les sources ciblées afin d'en extraire des informations utiles. L'extraction peut être manuelle, semi-automatique ou automatique. L'enjeu de cette opération réside dans son automatisation. Toutefois, le web scraping est considéré comme une des techniques privilégiées pour extraire des informations désirées à partir des pages web [15][16]. Le scraper aspire des contenus de pages web via un programme dans le but de les structurer pour permettre son utilisation dans un autre contexte [17]. Le web scraping nécessite l'utilisation de parsers (ou analyseurs) qui incorporent des API (Application Programming Interface), c'est-à-dire un ensemble de classes, méthodes ou fonctions qui spécifient comment des programmes consommateurs se servent de fonctionnalités du programme fournisseur. Ces parsers reposent souvent sur les chemins DOM (Document Object Model) ou XPath pour parcourir le document et en extraire les informations qu'il contient [18] [19]. De nos jours, plusieurs langages informatiques proposent des librairies notamment Java [20], Python [21] et R [15] afin d'extraire des contenus de pages web. Les données collectées comportent souvent des incohérences et des bruits. Il est donc utile de procéder à une phase de fusion.

2.2.1.3 - Fusion de données

La fusion consiste à unifier des données issues de plusieurs sources dans un seul format en détectant celles qui ne respectent pas les règles établies. En d'autres termes, il s'agit d'agréger des données dans un format unique. Cette opération intègre les premiers niveaux de nettoyage qui consiste à sélectionner les données en fonction de besoins définis en éliminant certains caractères spéciaux, des sauts de lignes, espaces inutiles, redondances, etc. L'objectif est d'obtenir une collection de documents propre et exploitable afin d'archiver correctement l'ensemble des éléments d'intérêt dans une base de données.

2.2.1.4 - Travaux sur l'acquisition de données en ligne

Dans la littérature, des solutions ont été développées pour extraire et fusionner des contenus à partir des sites dont la structure des pages est basée sur du HTML [22][21], des sites

d’informations [23][24]. Sous cet angle, nous nous focalisons sur des travaux qui s’appliquent sur les sites d’informations.

Dans cet ordre d’idées, Sundaramoorthy et al. [23] ont proposé une solution qui extrait et fusionne des articles publiés dans plusieurs sites d’informations dans le but de les résumer et de les visualiser. Cette solution prend en entrée les différentes URL et récolte toutes les données associées. Elle s’assimile à un agrégateur c’est-à-dire une plateforme qui visualise en agrégeant les données de plusieurs sources. Pour détecter les nouvelles publications, ces auteurs ont activé des fonctions de flux RSS. Le flux RSS est un outil par excellence pour détecter les dernières mises à jour sur les plateformes web.

Cependant une telle pratique est coûteuse en termes de ressources et alourdit le prétraitement du fait que toutes les données associées sont extraites alors qu’elles ne sont pas toutes pertinentes pour l’étude visée. En plus, le flux RSS est loin d’être pertinent comme source d’information à des fins d’analyse ; car il est contrôlé par les propriétaires. Par ailleurs, nous avons mené des tests sur 132 sites de presse en ligne au Sénégal. Les résultats obtenus abondent dans le même sens sur la non-pertinence des informations issues des flux RSS. En effet, nous avons constaté que ces flux RSS sont rarement à jour. Donc ils ne fournissent pas d’informations fiables concernant les nouvelles publications.

À côté de ces auteurs, Sarr et al. [24] ont proposé un extracteur automatique d'articles journalistiques dans le but de vérifier des faits journalistiques. Cette solution est basée sur la librairie Newspaper de python. En effet, Newspaper a été spécifiquement développée pour la presse en ligne, ce qui fait son avantage. À partir d’un URL principal, l’outil est capable d’extraire les différentes parties d’un article journalistique à savoir les liens, les auteurs, la date de publication, le titre, le résumé, les mots-clés, le contenu (texte) de la page et même l'image principale. Cependant, cet outil ne demeure pas sans inconvénient. En effet, il ne prend pas en compte le formatage des commentaires.

En résumé, les problèmes liés à l’acquisition de données sont, d’une part, les limites des flux RSS et d’une part la non prise en compte des commentaires. À côté de l’acquisition de données, il y a la valorisation. Pour cela, le prétraitement devient une étape cruciale.

2.2.2 - Prétraitement de données textuelles

Le prétraitement est l’étape intermédiaire entre l’acquisition et l’analyse de données. Elle consiste à transformer l’ensemble de données non structurées collectées en tableau de

valeurs afin d'adapter les textes aux méthodes d'analyse de données. Le prétraitement est composé de plusieurs tâches que nous décrivons ci-dessous.

2.2.2.1 - Normalisation linguistique

Par normalisation linguistique, nous faisons allusion à la segmentation, la lemmatisation (ou la racinisation) et la suppression d'éléments vides.

La segmentation (*tokenisation*) consiste à découper un texte en unités plus petites séparées par des signes particuliers. Dans les textes écrits en langue française, les signes de ponctuation (espaces, apostrophes, etc.) sont souvent utilisés pour segmenter le texte en plusieurs tokens. Un *token* est un constituant syntaxique de la phrase (mots, séquences contiguës de caractères, ponctuation, etc.). Dans l'Illustration 1, nous proposons une phrase et une segmentation basée sur les signes de ponctuation.

Illustration 1 : Segmentation basée sur les espaces

Texte brut : Macky Sall est le président du Sénégal.
Texte segmenté : «Macky », « Sall », « est », « le » « président », « du » « Sénégal », « . »

À l'issue de cette opération, une liste de tokens est produite. Ces tokens doivent ensuite faire l'objet de lemmatisation.

La lemmatisation d'un texte consiste à regrouper les mots qui ont les mêmes variations au sein de la même entrée. Il s'agit de regrouper les mots d'une même famille dans un texte afin de les réduire à leur forme canonique (le lemme). La lemmatisation passe par l'étiquetage morphosyntaxique qui consiste à identifier la structure grammaticale (nom, verbe, adjectif, adverbe, …) de chaque token. Le résultat obtenu à l'issue de cette opération est l'infinitif pour les verbes, le singulier pour les noms, etc. (voir Illustration 2).

Illustration 2 : Lemmatisation

Phrase 1 : Les bonnes mangues viennent de la Casamance
Lemmatisation 1 : « le » « bon » « mangue » « venir » « de » « le » « Casamance »
Phrase 2 : Les enfants viendront à la fête avec un bon cadeau
Lemmatisation 2 : « le » « enfant » « venir » « à » « le » « fête » « avec » « un » « bon » « cadeau »

À côté de la lemmatisation, la racinisation est aussi un procédé utilisé pour réduire la dimension de la liste. Elle consiste donc à nettoyer la liste obtenue en regroupant l'ensemble des déclinaisons autour d'une même racine comme nous le présentons dans l'Illustration 3.

Illustration 3 : Racinisation

Les mots chanterons, chantâmes, chantait donnent **chant**

Par ailleurs, les textes écrits en français comportent certains éléments qui ont un sens beaucoup moins précis. Ces éléments interviennent en particulier dans la construction des phrases. En général, ils n'ont pas de sens en eux-mêmes. Il s'agit de ponctuations, prépositions, déterminants, conjonctions, pronoms, verbes auxiliaires. Ils ont une distribution uniforme et apparaissent dans tous les textes. L'élimination de ces éléments permet de nettoyer la liste afin de faciliter l'analyse.

En pratique, plusieurs outils sont proposés dans la littérature pour la normalisation linguistique de texte écrit en français. Les plus populaires sont TreeTagger[25], UNIGRAM [26], Talismane [27], Qtag [28], Brill [29] et MEltfr [30]. Pour obtenir les groupes de mots, les fonctions d'extractions de *n-grammes* sont utilisées avec *n* pouvant varier entre 1 à 6 mots par exemple. Un *n-gramme* est une sous-séquence de *n* éléments construits à partir d'une séquence donnée. L'utilisation de la méthode de *n-grammes* permet une analyse de texte bien plus approfondie que la seule analyse de mots. Elle permet aussi de prendre en compte le contexte du token. Selon les besoins d'application, il est possible de supprimer les mots les plus courts fournis par la méthode de *n-grammes*. La liste de termes (mots et groupes de mots) obtenue à l'issu de la normalisation fait l'objet de pondération et de représentation pour appliquer les méthodes d'analyse de données.

2.2.2.2 - Pondération et représentation de termes

2.2.2.2.1 - Pondération de termes

La pondération permet de déterminer le poids d'un terme dans chacun des documents d'une collection. Ce poids permet de dire qu'un terme quelconque est discriminant ou pas, par

rapport à un document donné ; autrement dit, si le terme apparait souvent ou pas du tout dans le document en question. Il existe différentes manières de calculer le poids d'un terme dans une collection de documents. Par la suite, nous présentons les méthodes TF (*Term Frequency*) et TF-IDF (*Term Frequency-Inverse Document Frequency*) qui sont généralement utilisées.

- ***TF*** : On désigne par TF la fréquence d'un terme (descripteur) dans un texte donné. Nous dénombrons plusieurs manières de calcul de la TF:
    - TF absolue est le nombre de fois qu'un terme apparait dans un document donné.
    - TF relative est le rapport entre le nombre de fois qu'un terme est apparu dans le texte sur le nombre de tous les termes du texte.
    - TF booléenne se contente juste de la présence ou de l'absence du terme dans le texte.

Le calcul de TF est très simple et s'avère efficace en pratique. Mais, le principal inconvénient est le fait qu'un terme qui apparait avec une fréquence assez grande dans tous les documents d'un corpus n'a pas de pouvoir discriminant. La méthode TF-IDF est définie pour prendre en compte ce cas particulier.

- ***TF-IDF*** : Cette pondération montre le caractère discriminant d'un terme et est calculée souvent par la formule suivante :

$$tfidf_{t,d,D} = tf_{td} * \log_{10} \frac{N}{n_t}$$

Où

- TF est relative ou absolue,
- N: nombre de documents dans la collection,
- n_t: nombre de documents dans lesquels le terme *t* est apparu

Un terme est discriminant s'il est fréquent dans certains documents et rare dans d'autres. C'est un principe de rapport entre l'abondance particulière d'un terme et sa rareté générale dans la collection. Les travaux basés sur cette méthode consistent à extraire seulement les termes pertinents considérés comme discriminants [31]. Après cette étape, ces termes sont représentés pour permettre l'application des méthodes d'analyse de données directement.

2.2.2.2.2 - Représentation de termes

Il existe plusieurs types de représentations en vue de l'analyse de données. Ici, nous présentons la représentation en « sac de mots » et celle matricielle qui sont fréquemment utilisées en fouille de textes.

La représentation par sac de mots (*bag of words*) repose sur le principe qu'une collection de documents peut être décrite au moyen d'un dictionnaire (de « mots »). Un document est donc représenté par un vecteur de la même taille que le dictionnaire, dont la composante *i* indique le poids du $i^{ème}$ mot du dictionnaire dans le document. Pour l'appliquer les méthodes d'analyse de données, la collection de documents est représentée sous forme de matrice.

Les données stockées dans des matrices contiennent des documents en lignes et les termes en colonnes (ou inversement). Chaque cellule $\boldsymbol{C_{ij}}$ d'une matrice est le poids du terme $\boldsymbol{j}$ dans le document $\boldsymbol{i}$. La représentation matricielle permet une analyse simultanée de plusieurs variables (voir Tableau 1). En guise d'illustration, nous appliquons cette représentation sur la collection de documents suivante (voir Illustration 4) :

Illustration 4 : Présentation du jeu de données

Doc1 *:* Le Sénégal est membre de l'Union Africaine.
Doc2 *:* Macky Sall est le Président du Sénégal et le Président de l'Union Africaine.
Doc3*:* Moustapha Niasse est Président de l'Assemblé nationale du Sénégal et conseiller de l'Union Africaine.

Tableau 1 : Représentation matricielle

Doc/ Terme	Sénégal	Membre	Union Africaine	Macky Sall	Président	Moustapha Niasse	Assemblée nationale	conseiller
Doc1	1	1	1	0	0	0	0	0
Doc2	1	0	1	1	2	0	0	0
Doc3	1	0	1	0	1	1	1	1

L'objectif du prétraitement est de fournir une représentation sous un format exploitable par les méthodes d'analyse de données.

2.2.3 - Méthodes d'apprentissage automatique

En général, des algorithmes développés pour la fouille de textes sont issus de l'apprentissage supervisé [32] et non supervisé [33].

2.2.3.1 - Apprentissage supervisé

La classification automatique supervisée est composée de méthodes qui cherchent à affecter un nouvel élément dans une classe d'un ensemble de classes prédéfini sur la base de critères validés [34][35]. Le processus définit deux (2) phases à savoir la phase d'apprentissage et celle de test :

- ***Phase d'apprentissage*** : La phase d'apprentissage a pour objectif de déterminer un modèle à partir de données annotées ou non appelées aussi données d'apprentissage ou d'entrainement. Il existe de nombreux jeux de données issus des campagnes d'évaluation internationales comme le Défi de Fouille de Textes (DEFT), Semantic Evaluation (SemEval) ou Social Book Search (SBS), CLEF, etc. Ces corpus construits et distribués à la communauté scientifique permettent de développer de nouveaux modèles à partir des techniques d'apprentissage existantes afin de pouvoir les appliquer à tout type de contexte.
- ***Phase de test*** : La phase de test consiste à prédire avec le modèle la classe ou l'étiquette d'une nouvelle donnée issue de la base de test.

Aujourd'hui, nous assistons à un grand engouement autour de la fouille de textes. Il faut préciser que ces corpus d'entrainement fournis lors des campagnes d'évaluation internationales ne sont adaptés qu'aux textes rédigés en anglais ou en français. Par ailleurs, la mise en place d'un corpus d'entrainement nécessite un travail laborieux. Pour pallier à cette contrainte, l'apprentissage non supervisé est aussi utilisé.

2.2.3.2 - Apprentissage non supervisé

La classification automatique non supervisée appelée aussi ''clustering'' est une approche qui vise à identifier des ensembles d'éléments (individus ou variables) qui partagent certaines similarités afin de les regrouper dans la même classe [36] et dissocie en même temps des classes qui ont des éléments dissimilaires [37]. Ce regroupement d'enregistrements doit vérifier l'homogénéité intra-classe qui fait appelle à la cohésion et l'hétérogénéité inter-classe qui fait à l'allusion à la séparation. En fouille de textes, la classification non supervisée est souvent basée sur la détermination des co-occurrences. Les co-occurrences sont des termes qui apparaissent ensemble dans plusieurs documents et dans le même contexte [38][39][40].

Compter le nombre de co-occurrences entre des documents permet d'estimer s'ils sont sémantiquement liés ou non.

Nous complétons la description des méthodes d'analyse de données par la présentation des algorithmes les plus utilisés en fouille de textes.

2.2.4 - Algorithmes d'apprentissage utilisés en fouille de textes

En pratique, les systèmes de fouille de textes les plus efficaces sont basés sur des algorithmes d'apprentissage supervisé [41][42]. Dans cette partie, nous allons présenter brièvement quelques algorithmes d'apprentissages supervisés qui sont fréquemment utilisés en fouille de textes.

2.2.4.1 - Classifieur Naïve Bayes

La classification par la méthode naïve bayésienne propose un algorithme basé sur le théorème de Bayes qui est fondé sur les probabilités conditionnelles. Cette approche très utilisée dans la classification de textes se base sur la fréquence d'occurrences de termes (TF) dans les documents. Il s'agit de prédire la classe d'un document connaissant ses caractéristiques [43] [44].

2.2.4.2 - Régression logistique

Dans son expression mathématique, l'analyse par régression examine la relation entre une variable dépendante (la variable de réponse) et des variables indépendantes particulières (les prédicteurs). Les méthodes de fouille d'opinions basées sur la régression logistique tentent d'expliquer cette relation à travers la polarité (positive=1 ou négative=0) et les caractéristiques des documents. La polarité représente la variable expliquée (variable à prédire) obtenue à partir du corpus annoté et les caractéristiques correspondantes aux variables explicatives (variables prédictives) obtenues à l'aide de la fréquence d'occurrences de termes des documents de la base d'analyse. Lors du concours Deft 2007[5], Charton et al. [45] ont expérimenté cette approche qui avait donné de bons résultats.

2.2.4.3 - Machines à vecteurs de support

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais *support vector machine*, SVM) [46] constituent un ensemble de techniques d'apprentissage supervisé

[5] https://deft.limsi.fr/2007/

visant à maximiser la marge de séparation entre deux classes afin que tous les points d'une même classe soient du même côté de l'hyperplan. SVM est un classificateur dit linéaire, c'est-à-dire dans le cas parfait, les données (documents textuels dans notre cas) doivent être linéairement séparables. Toutefois, si les données ne sont pas linéairement séparables, la SVM peut être modifiée pour tolérer un minimum d'erreurs. D'ailleurs, le but est de maximiser la marge et de minimiser l'erreur de classification. Une autre alternative pour parer à la non-séparabilité des données consiste à passer à un espace de dimension supérieure. Manek et al. [47] ont expérimenté cet algorithme dans le domaine de l'analyse de sentiments.

2.2.4.4 - Principe d'entropie maximale (MaxEnt ou ME)

Un modèle de maximum d'entropie est un classificateur probabiliste linéaire et discriminant. La classification ME est une technique qui a fait ses preuves dans le traitement du langage naturel (NLP) [48]. Il est utilisé dans ce contexte pour organiser les documents par catégorie. Pour ce faire, le classifieur convertit l'ensemble des caractéristiques en un vecteur codé et associe à chaque caractéristique une pondération afin de déterminer les distributions maximales qui forment les classes. Les solutions basées sur ce modèle donnent souvent des résultats satisfaisants [49].

2.2.4.5 - Réseau de neurones

Un réseau neuronal est l'association d'objets élémentaires appelés neurone formel en un graphe plus ou moins complexe [50]. Un neurone formel est un modèle qui se caractérise par un état interne de données d'entrées et une fonction d'activation. La fonction d'activation opère la transformation d'une combinaison de données d'entrée en termes constants, appelée le biais du neurone. Cette combinaison est déterminée par un vecteur de poids associé à chaque neurone et dont les valeurs sont estimées dans la phase d'apprentissage. Les réseaux de neurones connaissent un regain d'intérêt et même un énorme avantage en analyse des données textuelles avec l'avènement du Deep Learning [51].

De nos jours, il y a principalement deux domaines d'applications de la fouille de textes que nous allons décrire par la suite.

2.3 - **Domaines d'applications de la fouille de textes**

Dans ce document, nous présentons essentiellement la Recherche d'Informations (RI) [52][53] et l'Extraction d'Informations [54] comme deux domaines phares d'application de la fouille de textes (voir la Figure 2).

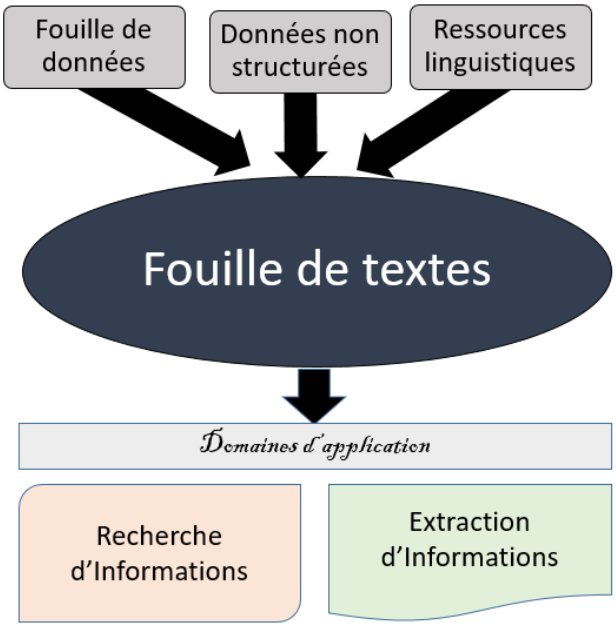


Figure 2 : *Fouille de textes et ses domaines d'applications*

2.3.1 - Recherche d'Informations (RI)

La RI est un ensemble de méthodes permettant de ressortir des informations cachées dans un ensemble de documents. Le but est de retrouver un ou plusieurs document(s) pertinent(s) dans une collection, à l'aide d'une requête plus ou moins informelle [55]. Le principe d'un système de recherche d'informations (SRI) est de fournir aux utilisateurs les documents dits pertinents correspondant à leurs besoins. La pertinence est mesurée à partir du degré de correspondance entre la requête et chaque document. Pour ce faire, le SRI compare la requête aux documents disponibles pour y répondre. Ce processus comprend les phases d'indexation, de recherche et d'appariement.

L'indexation consiste à traiter un document afin d'identifier un ensemble de descripteurs significatifs représentant son contenu. Les descripteurs, communément appelés éléments clés, entrées d'index ou termes d'indexation peuvent être des mots, groupes de mots ou concepts d'une ontologie. L'importance d'un descripteur dans un document est déterminée grâce à des méthodes telles que le TF, le TF-IDF ou encore la position du terme dans le document.

Dans la phase de recherche, l'utilisateur exprime son besoin en information par une requête. Comme les documents, la requête est aussi transformée et représentée par un ensemble

de descripteurs. Elle peut également être reformulée ou étendue pour mieux exprimer le besoin en information de l'utilisateur.

L'appariement documents-requête consiste à estimer la pertinence de chaque document de la collection par rapport à la requête de l'utilisateur afin de classer les documents. Cette correspondance est souvent basée sur les mots que partagent la requête et les documents. Les documents les plus pertinents sont ainsi retournés à l'utilisateur.

2.3.2 - Extraction d'informations (EI)

L'EI consiste à rechercher dans des textes en langue naturelle des informations à propos d'une cible à l'aide de critères prédéfinies [56]. En d'autres termes, les systèmes d'extraction d'information détectent les phrases pertinentes et en extraient les informations voulues concernant une entité sélectionnée. Ce processus repose sur trois grandes étapes à savoir d'abord l'identification de fragments de textes contenant une information, ensuite la définition de la structure de représentation de l'information et enfin le développement de règles permettant d'extraire l'information d'intérêt [57]. Les systèmes d'extraction d'information s'appuient sur les outils du Traitement Automatique du Langage Naturel (TALN) [58][59].

Le TALN est une discipline connexe à la fouille de textes qui a pour objectif de modéliser, grâce à l'informatique, le langage humain écrit ou parlé [60]. Il est utilisé pour effectuer une analyse lexicale et morphosyntaxique d'un texte. Il s'intéresse à la syntaxe et à la sémantique du texte à analyser pour aider à reconnaître les constituants du texte (phrases, mots), leur nature et leurs relations. À cet effet, l'étiquetage morphosyntaxique (Part Of Speech Tagging en anglais) est une impérieuse nécessité. Il consiste à assigner à chaque mot d'un texte sa catégorie grammaticale.

Aujourd'hui, plusieurs sous domaines d'applications de l'EI émergent avec le foisonnement de données textuelles. Comme sous domaine d'application nous pouvons par exemple citer le résumé automatique, la reconnaissance d'entités nommées, etc. Le résumé automatique de texte consiste à faire une reformulation et paraphrase, une extraction du contenu pertinent d'un texte, une détection des informations les plus importantes, des redondances, afin de générer un texte cohérent humainement crédible. La reconnaissance d'entités nommées permet de déterminer dans un texte des expressions linguistiques qui font référence à un objet précis du monde [61] [62][63]. Il s'agit notamment de noms de personnes, de produits, de maladies, d'organisations et d'emplacements géographiques, ainsi que les concepts de temps, de monnaie, de pourcentage, d'événement, de quantité, de distance, de valeur, etc.

2.4 - **Fouille d'opinions**

La fouille d'opinions est aussi une application de la fouille de textes. Elle consiste à traiter des contenus textuels souvent issus d'échanges sur le web afin de découvrir l'opinion majoritaire d'internautes. Dans ce contexte précis, la grande question qui mérite d'être posée est la suivante :

Qu'est-ce qu'une opinion ?

Pour répondre à cette question, nous allons d'abord définir la notion d'opinion, ensuite décrire les différents niveaux de granularité de l'opinion, après parler de RI et EI dans les systèmes de fouille d'opinions et enfin, nous mettrons en exergue les opportunités de la fouille d'opinions.

2.4.1 - Opinion

2.4.1.1 - Définition de l'opinion

Ici, la définition consiste à une description de la notion d'opinion dans sa particularité et spécificité pour faciliter la reconnaissance d'un vocabulaire d'indices d'opinions dans un texte. Ce type de travail n'est pas chose aisée ; néanmoins nous le tenterons dans le but de distinguer l'opinion par rapport au sentiment de manière générale.

Le sentiment et l'opinion émanent tous de la subjectivité d'un discours dont les propos ne peuvent pas être vérifiés pour leur véracité ou exactitude. En d'autres termes, la subjectivité est un ressenti personnel de l'individu. Il s'agit de toute information qui relève de l'avis, de l'évaluation, de la croyance ou du jugement personnel. Ces deux concepts sont sémantiquement proches ; néanmoins il existe des nuances.

Le sentiment prend sa source de l'émotion c'est à dire toute information qui relève de la joie, la peur, la tristesse, la colère, la surprise, le dégout, l'attirance et l'anticipation [64][65][66]. Il est considéré par certains auteurs [67][68] comme une combinaison d'émotions c'est-à-dire la somme de deux ou plusieurs émotions comme nous le montrons dans le Tableau 2.

Tableau 2 : Sentiments issus de la combinaison d'émotions[6]

[6] https://fr.wikipedia.org/wiki/Robert_Plutchik

Combinaison d'émotions	Sentiments
joie et attirance	Amour
attirance et peur	Soumission
Peur et surprise	Crainte
Surprise et tristesse	Désappointement
Tristesse et dégoût	Remords
Dégoût et colère	Mépris
Colère et anticipation	Agressivité
Anticipation et joie	Optimisme

Contrairement au sentiment, l'opinion est un vocabulaire qui exprime un point de vue, une évaluation sur un objet du monde réel [69].

Pour déterminer la subjectivité d'un document, la fouille d'opinions (ou Opinion Mining) [1][2] et l'analyse de sentiments (Sentiment Analysis) [70][71] sont les deux concepts utilisés. Bien que ces deux soient employés de manière interchangeable, il existe une différence entre eux dans la pratique. Le vocable « Analyse de sentiments » est utilisé lorsqu'on identifie dans une collection les vocabulaires subjectifs et les catégorise en fonction d'émotions tandis que celui de « Fouille d'opinions » est utilisé lorsqu'on veut classer les textes suivant l'opinion qu'ils expriment c'est-à-dire favorable, défavorable ou neutre. Dans cette thèse, nous utilisons le terme « Fouille d'opinions » que nous jugeons plus approprié pour notre thème de recherche.

2.4.1.2 - Types d'opinions

Dans les documents textuels, les opinions sont exprimées de différentes manières. Dans la littérature, ces opinions sont catégorisées en trois types à savoir les opinions implicite, comparative et régulière (ou explicite) [72][73] comme le montre la figure 3.

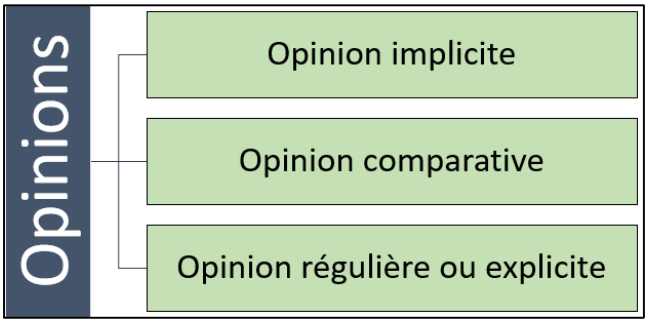


Figure 3 : *Type d'opinions*

- **Opinion implicite** : Les termes que nous qualifions d'opinions implicites n'impliquent pas directement la subjectivité [74]. Ce type d'opinions est souvent entrainé par l'usage de figures de style ou l'emploi d'expressions au sens figuré et non au sens propre des mots. Dans l'exemple ci-dessous, le terme "bon marché" désigne le prix abordable (voir Illustration 5).

Illustration 5 : Opinion implicite

Ce téléphone portable est à bon marché

- **Opinion comparative** : L'opinion comparative est le type d'avis qui met en comparaison l'entité ciblée avec d'autres entités [75]. Elle crée une relation de similarité ou dissimilarité entre deux ou plusieurs éléments de la phrase. Ce type d'opinion se caractérise par la présence d'un prédicat exprimant l'évaluation (voir Illustration 6).

Illustration 6 : Opinion comparative

Sadio Mané est plus efficace que Neymar devant les buts

- **Opinion régulière ou explicite** : Une opinion régulière ou explicite exprime un avis directement sur des caractéristiques ou des aspects spécifiques de l'entité ciblée. Dans la littérature, ces termes sont des vocabulaires naturellement appelés opinions. Ce type de vocabulaire vise à exprimer des opinions en termes non ambigus. Il montre le niveau de subjectivité qui n'implique pas l'expression d'une évaluation (voir Illustration 7).

Illustration 7 : Opinion régulière

Phrase 1 : Sadio Mané est un attaquant efficace **Phrase 2** : Le son de ce mobile est bon

L'extraction d'opinions implicite ou comparative à partir d'un ensemble de documents constitue un grand défi contemporain. Cependant, la plupart de travaux réalisés en fouille d'opinions s'intéressent à l'opinion explicite [76].

2.4.1.3 - Formalisation de l'opinion

Dans la littérature, Liu et al. [77] ont proposé une représentation de l'opinion autour de cinq dimensions comme le montre la Figure 4.

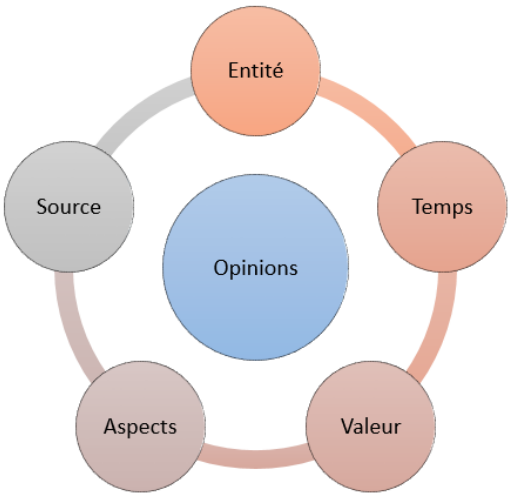
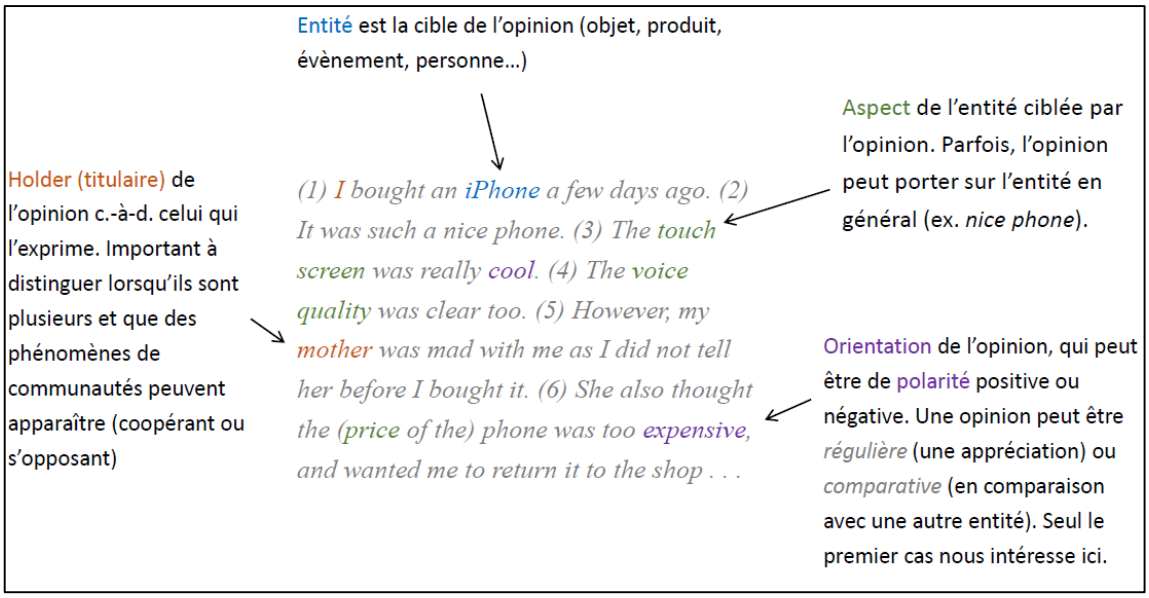


Figure 4 : *Modélisation d'opinions selon Liu et al. [77]*

- **Entité** : l'entité est la cible de l'opinion, autrement dit sur quoi porte l'opinion ;
- **Aspect** : l'aspect est un ensemble d'éléments liés à la cible jugée c'est-à-dire l'entité. Ce sont des caractéristiques qui permettent d'apprécier l'entité ;
- **Source** : la source constitue l'élément énonciateur ou l'auteur de l'opinion. En d'autres termes, c'est celui qui exprime l'opinion ;
- **Temps** : le temps est le moment d'énonciation c'est-à-dire la date à laquelle l'opinion a été exprimée ;
- **Valeur** : la valeur est l'orientation de l'opinion qui peut être de polarité positive, négative ou neutre.

Cette description est basée sur l'opinion régulière et a été formalisée comme la représentation de l'opinion. En guise d'illustration, Rico Rakotomalala [78] a annoté un extrait de texte tiré du livre de Liu et Zhang (page 416) [2]. Dans son exemple, l'auteur cherche à mettre en exergue les différentes dimensions du modèle à l'intérieur d'un texte (voir l'Illustration 9).

Illustration 8 : Extrait de texte annoté (exemple introductif Liu et Zhang, page 416)[2]



Dans la littérature, cette formalisation est considérée comme l'approche la plus performante pour fournir des résultats fiables et proches de la réalité [79][80][81]. Cependant, sa mise en pratique nécessite un processus complexe et difficile à réaliser.

2.4.2 - Niveau de granularité

La fouille d'opinions est aussi une application de la fouille de textes. Elle consiste à traiter des contenus textuels souvent issus d'échanges sur le web afin de découvrir l'opinion majoritaire d'internautes. La fouille d'opinions s'applique à différents niveaux de granularité qui sont les niveaux document, phrase ou aspect [82].

- ***Niveau document*** : L'étude des opinions au niveau du document consiste à déterminer l'opinion générale du document [83][84]. Cette étude cherche à classer des documents en deux catégories selon les avis exprimés : favorable ou défavorable. Elle considère l'ensemble du document comme une unité d'informations de base et suppose que le document est connu pour son opinion.
- ***Niveau phrase*** : Au niveau phrase, la classification d'opinions est appliquée à des phrases individuelles dans un document [85][86] [87][88]. L'étude ne considère que les phrases contenant des mots ou expressions exprimant des opinions. Enfin, ces phrases sont classées en fonction de l'axiologie positive ou négative.
- ***Niveau aspect*** : La fouille d'opinions au niveau des aspects consiste à analyser le texte en tenant compte du contexte dans lequel l'opinion est exprimée [89]. Il s'agit de prendre en considération tous les éléments qui concourent à l'expression d'une opinion c'est-à-dire l'entité et son environnement immédiat.

Bien que la fouille d'opinions au niveau du document et de la phrase soit utile, mais dans de nombreux cas, elle laisse encore à désirer. L'évaluation positive d'un texte d'une entité particulière ne donne pas toujours une opinion positive. De même, un texte d'évaluation négative d'une entité ne signifie pas non plus que l'auteur n'aime pas l'entité entièrement. Pour illustrer ce propos, nous donnons en exemple cette phrase (voir Illustration 8) :

Illustration 9 : Phrase complexe en fouille d'opinions

Le cours de fouille d'opinions est intéressant, mais l'enseignant est ennuyeux

Cette phrase présente deux opinions que sont :

- Le cours de fouille d'opinions qui est perçu comme ***positif*** ;
- L'enseignant qui est perçu comme ***négatif***.

Pour obtenir une analyse plus fine d'une telle phrase, la fouille d'opinions basée sur les aspects a été adoptée comme l'approche la plus efficace et fiable. C'est ainsi que la formalisation de l'opinion a été proposé.

2.4.3 - RI et EI dans les systèmes de fouille d'opinions

Nous pouvons dire sans abus de langage que la recherche d'informations et l'extraction d'informations sont deux domaines de recherche qui ont évolués séparément et même s'ils ont tous deux un ancrage dans l'informatique, l'un est davantage tourné vers le documentaire et l'autre vers la linguistique [90]. Aujourd'hui, les données massives disponibles sur le web posent à ces deux domaines des défis qui les obligent à coopérer progressivement. Cette coopération est mise en œuvre dans les systèmes de fouille d'opinions afin de proposer des solutions performantes.

Les systèmes de fouille d'opinions fonctionnent en trois étapes, à savoir la sélection de l'entité (évènements, services, articles journalistiques, etc.), l'identification de vocabulaires porteurs d'indices d'opinions et la classification de documents selon les orientations favorables, défavorables ou neutres. En fouille d'opinions, un document peut être considéré comme un commentaire ou une phrase d'un commentaire. L'ensemble de commentaires sélectionnés pour l'analyse est appelé la base d'analyses qui est sélectionnée en fonction d'une entité nommée (EN). Les systèmes de fouille d'opinions se servent de la robustesse du processus de RI pour indexer des documents afin de faciliter la sélection de la base d'analyses. Ils se servent aussi

d'EI pour identifier des entités et les termes porteurs d'indices d'opinions. Maintenant, nous allons amplement parler de la fouille d'opinions.

2.4.4 - Intérêt de la fouille d'opinions

Les intérêts que l'on peut tirer de la fouille d'opinions vont de pair avec les applications à cette dernière.

Dans le contexte de forums médicaux ou sanitaires, la fouille d'opinions constitue un outil pour obtenir des taux de suicide et prévoir les cas de suicides, pour comprendre les caractéristiques de la dépression. En plus, elle peut aussi aider à détecter des zones saines et malsaines. À côté de la santé, la fouille d'opinions est utile dans le domaine des affaires.

Dans le domaine des affaires (finance et commerce), la fouille d'opinions est considérée comme un outil d'e-marketing et de veille. Étant outil d'e-marketing, la fouille d'opinions, sur des données des clients, permet à l'entreprise d'identifier des tendances, des demandes du marché et des clients leaders. En plus, elle peut aussi l'aider à obtenir une vue précise de la perception de ses clients sur son image. Dans le même sillage, la fouille d'opinions est aussi bien adaptée au développement de tâches de veille. La veille permet d'anticiper sur des menaces comme une évolution de réglementation ou l'arrivée d'un nouveau produit ou concurrent. Elle permet aussi de capter des opportunités, comme un nouveau procédé technologique permettant, par exemple, de baisser les coûts d'une gamme de produits. Les intérêts de la fouille d'opinions ne se limitent pas seulement au domaine des affaires. Nous les rencontrons aussi dans le domaine social et politique.

Dans le domaine social et politique, les acteurs sociaux et politiques se servent de la fouille d'opinions pour prévoir les menaces qui pèsent sur la société. Il s'agit de détecter des risques de violence, des discours d'incitation à la haine, au meurtre, des risques d'attentat, etc. Ils peuvent l'utiliser pour jauger le niveau de satisfaction des populations, faire une prévision des résultats des élections, attirer l'attention du public ou contrôler et suivre l'opinion des étudiants dans l'éducation.

2.5 - Conclusion

En définitive, nous pouvons retenir que la fouille de textes est un processus d'analyse de données spécialement dédié au traitement automatique de contenus textuels. La constitution de corpus permet de regrouper de manière synthétique et cohérente des données issues de différentes sources dans le but de constituer une base de données pour une éventuelle analyse.

Les techniques de prétraitement utilisées ont pour mission de transformer les données textuelles non structurées en un format exploitable par les méthodes reposant sur les statistiques et/ou la linguistique. Aujourd'hui, la fouille de textes est devenue un défi contemporain avec des campagnes internationales qui sont organisées telles que le Défi de Fouille de Textes (DEFT), le Semantic Evaluation (SemEval) ou le Social Book Search (SBS), CLEF.

Une application récente de la fouille de textes est la fouille d'opinions. Cette dernière se distingue du Traitement Automatique du Langage Naturel (TALN) et de la Recherche d'Informations (RI). La plupart de ces disciplines tentent de résoudre des problématiques de classification de textes. De manière générale, les méthodes issues de ces disciplines sont appliquées aux documents textuels pour détecter des régularités, chercher des similarités, identifier des relations de causalité, accumuler des informations sur leurs auteurs et leurs lecteurs [13][91]. Dans une telle situation, le poids des termes implique leur pouvoir de discrimination.

Cependant, lors des tâches de fouille d'opinions, la présence d'un terme n'est pas très discriminante. Ce type de classification concerne un ensemble fermé de classes qui décrivent un attribut de l'opinion. À cet effet, nous pouvons dire à juste raison que c'est un outil d'aide à la décision. Dans le chapitre suivant, nous étudierons les approches de fouille d'opinions.

3 - ETAT DE L'ART SUR LA FOUILLE D'OPINIONS

3.1 - **Introduction**

La fouille d'opinions vise à détecter les opinions exprimées dans un document et plus généralement dans un corpus. L'émergence de la fouille d'opinions coïncide avec l'avènement du web 2.0 qui est à l'origine de médias sociaux tels que Facebook, Twitter, WhatsApp, etc. [92]. L'application des techniques de fouille d'opinions aux discussions issues de ces médias en ligne représente un véritable avantage compétitif pour les décideurs. Les intérêts que l'on peut tirer de la fouille d'opinions vont de pair avec les applications à cette dernière. Durant ces dernières années, la fouille d'opinions a connu un essor remarquable à travers ses domaines d'applications avec des propositions de solutions performantes [93][94]. À cet effet, nous pouvons dire à juste raison que c'est un outil d'aide à la décision. Malgré la performance de ces outils (ressources et méthodes) proposés, la complexité des commentaires issus de la presse sénégalaise en ligne demeure un défi contemporain.

Dans ce chapitre, notre objectif est de faire un état de l'art sur les approches et méthodes de fouille d'opinions actuelles afin de montrer leurs limites face à la complexité des commentaires de la presse sénégalaise en ligne. Pour cela, nous décrivons les approches de fouille d'opinions utilisées en générale pour analyser les commentaires en ligne. Ensuite, nous allons poser la complexité des commentaires issus de la presse sénégalaise en ligne. Enfin nous mènerons une discussion afin d'ouvrir des perspectives de solutions.

3.2 - **Approches de fouille d'opinions**

Le processus de fouille d'opinions est spécifiquement composé de trois grandes étapes à savoir le prétraitement de données, l'étiquetage d'opinions et la classification des documents. Le prétraitement concerne les tâches de segmentation, de lemmatisation (ou racinisation) et de sélection des termes candidats. L'étiquetage d'opinions consiste à sélectionner des vocabulaires susceptibles de contenir l'indice de subjectivité et leur affecter des polarités à l'aide des ressources linguistiques telles qu'un dictionnaire d'opinions, un lexique d'opinions, un corpus d'entrainement (ou corpus d'apprentissage) ou une ontologie. Enfin, la classification permet de catégoriser un commentaire en fonction de « favorable » ou « défavorable » à l'égard d'une entité sélectionnée. Pour expliquer plus clairement ce processus, nous proposons le schéma ci-dessous (voir Figure 5).

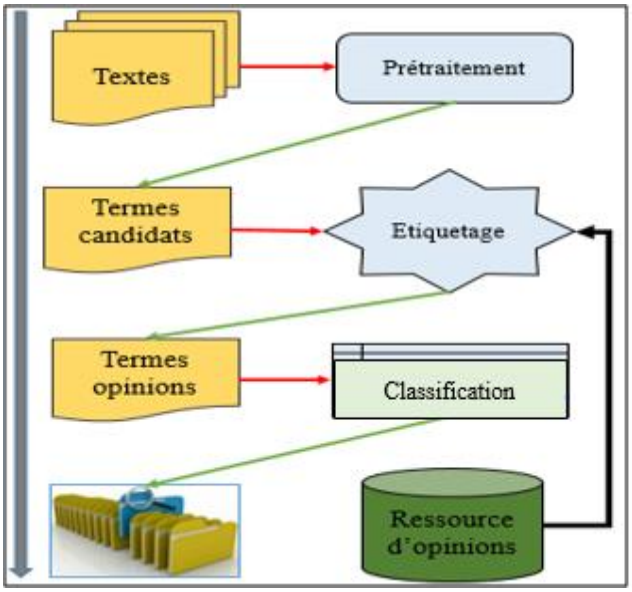


Figure 5 : *Processus de fouille d'opinions*

Dans l'état de l'art, nous nous intéressons aux approches utilisées pour analyser les commentaires en ligne. Selon les ressources utilisées, les approches se déclinent. Il existe trois approches qui sont l'approche lexicale, l'approche par apprentissage automatique et celle hybride comme le montre la Figure 6.

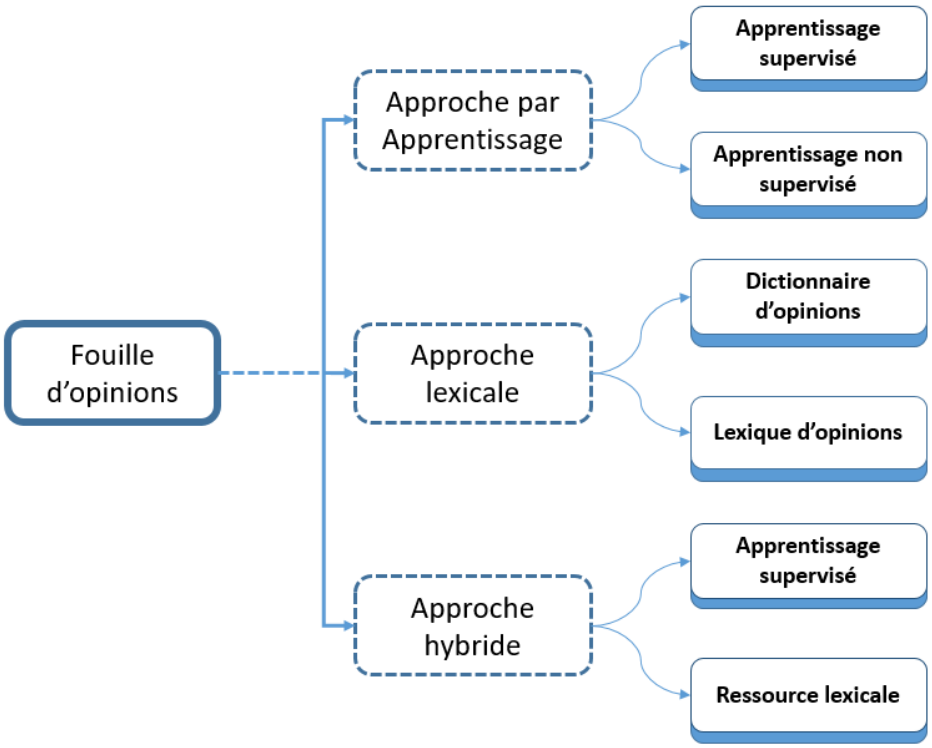


Figure 6 : *Approches de fouille d'opinions*

3.2.1 - Approche lexicale

L'approche lexicale est une démarche fondée sur l'analyse lexicale. Elle consiste d'une part, à déduire l'opinion dégagée par un terme à l'aide d'un dictionnaire ou un lexique d'opinions (de sentiments ou d'émotions) et d'autre part de déterminer l'orientation d'un commentaire à travers le calcul du score.

3.2.1.1 - Étiquetage d'opinions

- **Dictionnaires d'opinions** : En fouille d'opinions, un dictionnaire est un outil composé de mots étiquetés par polarité, intensité ou émotion. La polarité définit l'orientation sémantique d'un terme d'opinion en fonction de valeurs positives ou négatives. Tandis que l'intensité mesure le degré de positivité ou négativité d'un terme ; elle est souvent prise dans un intervalle défini par exemple de -1 à 1 ou de -5 à 5, etc. Quant à l'émotion, elle relève de la joie, de la peur, de la tristesse, de la colère, de la surprise et du dégout [68]. Dans la littérature, il existe plusieurs dictionnaires d'opinions parmi lesquels nous pouvons citer *Inquirer*, *HM*, *WordNet*, *BabelNet*, *HowNet*, *WordNet-Affect*, *SentiWordNet* et *SenticNet*. Parmi ces dictionnaires *WordNet-Affect*, *SentiWordNet* et *SenticNet* sont les plus utilisés pour l'étiquetage d'opinions de textes écrits en français [95][96]. À côté de ces dictionnaires, Python et R ont respectivement proposé *TextBllob*[7] et *Sentiment*[8] qui fonctionnent comme des dictionnaires d'opinions.
- **Lexique d'opinions** : Dans son sens premier, un lexique est une ressource linguistique tout comme le dictionnaire. À la différence de ce dernier, le lexique constitue un ensemble de mots ou de groupes de mots qui partagent une propriété sémantiquement commune d'un domaine. Un lexique d'opinions est une liste de termes annotés par polarité, intensité ou émotions. Dans la littérature, il existe plusieurs lexiques d'opinions conçus par des informaticiens en collaboration avec des linguistes [97][98]. Dans notre contexte, nous nous intéressons aux outils réalisés pour étiqueter les textes écrits en français, notamment FEEL et FWLSA-score. FEEL[9] est un lexique d'opinions open source proposé par Abdaoui et al. [99]. Il contient plus de 14 000 termes (mots et groupes de mots) distincts exprimant la subjectivité. Ce lexique est composé d'attributs suivants : polarité, joie, peur, tristesse, colère, surprise et dégoût. La proposition de ces

[7] https://textblob.readthedocs.io/en/dev/extensions.html
[8] http://edutechwiki.unige.ch/fr/Analyse_de_sentiments_avec_R
[9] http://www.lirmm.fr/~abdaoui/FEEL

auteurs est une réadaptation d'autres travaux [100]. Dans ce même sillage, Kandé et al. [101] ont proposé FWLSA-score qui est un lexique bilingue (français et wolof). Ce lexique est une extension de FEEL. C'est un travail remarquable, mais l'outil n'est pas open source.

3.2.1.2 - Calcul de score d'un document

Ce calcul s'effectue sur la base du score mesuré en fonction de la présence de termes issus de ces ressources linguistiques dans un document. Pour cela, on s'intéresse au résultat issu de ce calcul classique (voir Figure 7) :

- Soit $C = \{t_1, t_2, \dots, t_n\}$, un commentaire composé de n termes $t_1, t_2, \dots, t_n$;
- Soit P (Polarité), la valeur de chaque terme qui peut être -1 ou 1.

Le score d'un commentaire C, noté $Score(C)$ est par définition la somme des polarités de termes qui composent le commentaire

$$Score(C) = \sum_{i=1}^{n} P(t_i)$$

- Si Score(C) > 0 alors C a une orientation positive ;
- Si Score(C) < 0 alors C a une orientation négative ;
- Si Score(C) $= 0$ alors C n'a pas une orientation (neutre).

Figure 7 : *Classification de documents par approche lexicale*

Dans cette approche, les mots (ET, OU, SOIT, CEPENDANT, MAIS, …) ont un rôle prépondérant. Ils facilitent la détermination de l'orientation du terme. Par exemple, la conjonction ET dit que des termes conjoints ont généralement la même polarité ; alors que des expressions telles que, MAIS, CEPENDANT entrainent des changements d'opinion. Cette forme est considérée comme la cohérence de l'opinion qui n'est pas toujours cohérente dans la pratique [102][103].

L'approche lexicale a surtout l'avantage de permettre des calculs rapides sur de grands corpus. Beaucoup de travaux récents s'intéressent à cette approche [104][105][106]. Cependant, il faut signaler que les dictionnaires présentent quelques limites. En effet, ils ne prennent pas en compte le contexte d'utilisation du terme. Or, l'orientation sémantique d'un terme peut fortement dépendre de son contexte d'emploi. En plus, ils ne prennent pas aussi en

compte les groupes de mots c'est-à-dire les mots composés ou locutions qui sont utilisés comme expressions d'opinions. Quant à la mise en place d'une ressource adaptée aux besoins d'applications spécifiques, elle nécessite un travail laborieux.

Parallèlement à cette approche, l'approche par apprentissage automatique a été proposée. À présent, nous allons décrire cette approche.

3.2.2 - Approche par apprentissage automatique

L'approche par apprentissage automatique consiste à entrainer la machine à identifier des vocabulaires d'opinions afin de déduire ou de prédire l'opinion majoritaire d'une entité [107][108]. Dans ce contexte, il existe essentiellement deux étapes : la sélection de termes candidats et la classification d'opinions.

3.2.2.1 - Sélection de termes candidats

En fouille d'opinions, un terme candidat est un mot ou groupe de mots susceptible de contenir des opinions. Dans la pratique, la sélection des termes se basent sur plusieurs critères parmi lesquels nous pouvons citer la nature grammaticale des mots, la fréquence d'occurrences de termes (en anglais *Term Frequency* ou TF) et la fréquence des termes et des documents inversés (en anglais *Term Frequency-Inverse Document Frequency* ou TF-IDF).

- ***Critère basé sur la nature grammaticale de mots*** : Dans un texte, certaines structures grammaticales ont un sens précis et c'est grâce à elles que l'on décrit les éléments qui nous entourent avec exactitude. Il s'agit de noms, d'adjectifs, d'adverbes et de verbes. Les critères basés sur la nature grammaticale de mots s'appuient sur ces structures pour sélectionner les indices d'opinions dans la phrase. C'est le cas des auteurs qui s'intéressent aux adjectifs comme des indicateurs d'opinion [109][110]. Dans la plupart des cas, les méthodes basées sur ce critère sont considérées comme de l'apprentissage non supervisé.
- ***Critère basé sur le TF*** : La fréquence des occurrences de termes d'un document est déterminée par le comptage de chaque terme du document déjà prétraité. Ce critère permet de limiter la dimension du document. Il est très utilisé par les méthodes de classification automatique supervisée.

- ***Critère basé sur TF-IDF*** : Le critère de TF-IDF [111] permet de sélectionner seulement les termes jugés pertinents, c'est-à-dire les termes dont le score dépasse un seuil défini. Ce critère est fréquemment utilisé dans les méthodes de classification par similarité.

Par ailleurs, certains auteurs s'intéressent à la fréquence de mots corrélés aux expressions nominales pour sélectionner les vocabulaires d'opinions dans les phrases [102].

3.2.2.2 - Classification d'opinions

« Classifier consiste à définir des classes et classer est l'opération permettant de mettre un objet dans une classe définie au préalable »[112]. L'analyse d'opinion par classification met en jeu un processus automatisé dans lequel une polarité est affectée à chaque terme candidat. En fouille d'opinions, les systèmes de classification de textes les plus efficaces sont basés sur des algorithmes d'apprentissage supervisé [41][42] que nous avons décrit dans le chapitre précèdent. Ces algorithmes utilisent des corpus d'entrainement afin de prédire l'étiquète d'un terme candidat. Un corpus d'entrainement est un ensemble de données (paragraphe, phrase, mots, etc.) qui peuvent être annotées ou non. Dans notre contexte, les corpus sont annotés suivant les polarités ou émotions. Il existe des corpus d'entrainement open source conçus lors des campagnes d'évaluation internationales. Malheureusement, ils ne sont adaptés qu'aux textes écrits dans des langues officielles.

3.2.3 - Approche hybride

L'approche hybride consiste à combiner des méthodes issues des deux approches décrites précédemment. Dans la plupart des cas, il s'agit de combiner l'usage de ressources linguistiques et d'algorithmes d'apprentissage afin d'exploiter les avantages de chacune et d'obtenir le résultat le plus proche possible de la réalité [113][114]. À ce niveau, Lo et al. [115] ont proposé une solution d'analyse de sentiments.

D'autres auteurs combinent plusieurs algorithmes afin d'améliorer la qualité de la précision dans la classification de documents [116][117]. Dans cet ordre d'idées, il y a des auteurs qui utilisent simultanément leurs algorithmes de classification sur un dictionnaire d'opinions d'une part et sur un corpus annoté d'autre part afin de prouver la performance et la fiabilité des solutions proposées [118][119].

Ces propositions sont souvent validées par l'analyse linguistique qui se distingue par son caractère manuel. C'est un travail dévolu aux experts des domaines, notamment les linguistiques qui associent les vraies étiquettes aux termes.

En réalité, la plupart des outils (ressources et méthodes) de fouille d'opinions sont adaptés pour des textes en Français, Anglais ou Portugais. Ces langues officielles sont bien formalisées et ont suffisamment d'outils pour l'apprentissage automatique de manière générale. Cependant, la complexité de commentaires issus de la presse sénégalaise en linge remet en cause la performance de ces outils proposés.

3.3 - Complexité de commentaires de la presse sénégalaise en ligne

En voulant appliquer les outils de fouille d'opinions existants sur les données soumises à l'analyse, nous faisons face à de nombreux obstacles [120] comme le montre la Figure 8 dont les détails seront donnés dans cette section.

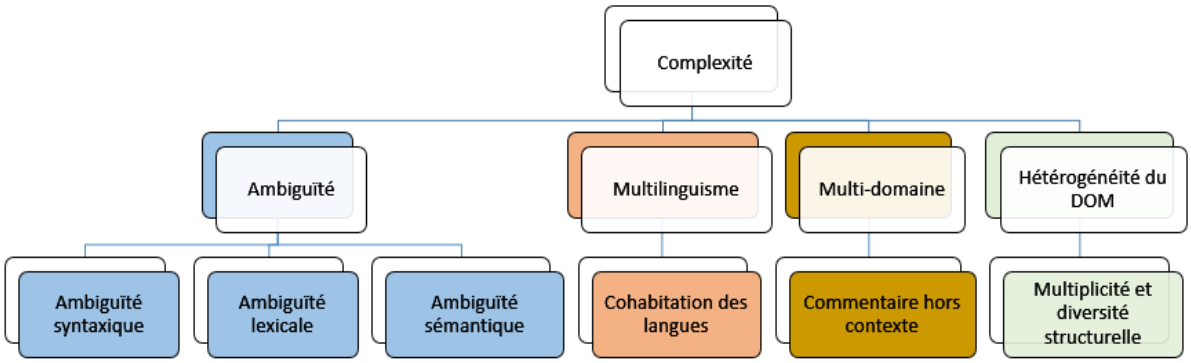


Figure 8 : *Complexité de commentaires issus de la presse sénégalaise en ligne*

3.3.1 - Obstacles liés aux ambiguïtés

Si l'on veut comprendre l'ambiguïté, il est indispensable, préalablement à toute tentative de traitement technique de s'interroger sur le statut linguistique de ce phénomène. L'ambigüité dans la langue et dans l'exercice de la langue est le reflet sans doute des interrogations en matière de syntaxe et de sémantique. Elle peut être considérée comme des propriétés intrinsèquement associées à une phrase prise hors contexte et qui provoque diverses interprétations. Elle met en jeu la question du statut de la phrase ainsi que celle de son mode de correspondance à l'énoncé réel.

En réalité, c'est une problématique liée à la structuration syntaxico-sémantique de l'énoncé qui peut être regroupée en trois (03) types : à savoir les ambigüités syntaxique, sémantique et lexicale.

3.3.1.1 - Ambiguïté syntaxique

La syntaxe est l'étude scientifique de la construction des phrases. Elle tente de répondre à un besoin de cohésion c'est-à-dire l'organisation du texte d'un document. La cohésion repose sur des mots et sur des procédés grammaticaux qui relient les phrases entre elles et donnent au texte son unité. Ainsi, la syntaxe vise à déterminer les principes qui gouvernent les relations de combinaison et de dépendance entre les mots et les groupes de mots au sein de la phrase. Ces principes sont décrits par la grammaire de la langue concernée.

En fouille de texte, on parle d'ambiguïté syntaxique quand un texte est susceptible de contenir des fautes d'orthographe, grammaticales et de mise en forme.

- ***Fautes d'orthographe*** : Selon le dictionnaire Larousse, « L'orthographe est l'ensemble de règles et d'usages qui régissent la manière d'écrire correctement les mots d'une langue ». En langue française, si l'écriture d'un mot est reconnue par un dictionnaire comme telle alors il est correct, sinon le mot est considéré comme incorrect. Dans des commentaires en ligne, nous avons souvent des erreurs diacritiques (fautes d'accents, de tréma et de cédille) et des erreurs de ligature c'est-à-dire la combinaison de deux lettres pour former un seul graphème et un seul phonème (par exemple ae, œ, ph ou ch).
- ***Fautes grammaticales*** : La grammaire est la description des contraintes caractéristiques d'une langue donnée. Elle est donc constituée de principes universels et spécifiques à une langue. En d'autres termes, la grammaire est un formalisme permettant de définir un langage formel. Dans notre contexte, nous nous intéressons aux fautes d'accord en genre et en nombre et aux fautes de conjugaison.
- ***Fautes de mise en forme*** : Les fautes de mise en forme regroupent la catégorie d'erreurs d'espace, de ponctuation, de casse, de nombre ou d'unités de mesure et d'élision.

Pour illustrer ces propos, nous allons donner en exemple un extrait de commentaires écrit entièrement en français contenant des fautes d'orthographe et syntaxiques (voir Illustration 10).

Illustration 10 : Extrait de commentaire ayant une ambiguïté syntaxique

« bravo félicitation et bonne continuation vous êtes le future mais votre maturité technique et mentale doivent être prisensenten exemple che l'equipe senior et vous meritez une selection che eux… »

Dans cet exemple, nous voulons souligner la mauvaise structuration syntaxique des phrases contenues dans les commentaires. Dans ce même sillage, nous avons aussi l'ambiguïté sémantique.

3.3.1.2 - Ambiguïté sémantique

Appartenant à la linguistique, la sémantique étudie les différents sens du mot et du langage. Elle veille sur la cohérence d'un texte à travers les mots contenus dans les phrases et qui peuvent permettre aux lecteurs de le comprendre et de l'interpréter. L'ambiguïté sémantique est une ambigüité profonde qui est causée par l'utilisation de figures de style, d'anaphores grammaticales, de polysémie, etc.

- ***Figures de style*** : une figure de style est d'abord une manière de s'exprimer. Il s'agit de modifier le langage ordinaire pour le rendre plus expressif. Il existe des figures d'analogie, d'animation, de substitution, de pensée, d'opposition, de construction, de sonorités, d'insistance et d'atténuation. L'utilisation de figures de style dans les commentaires pose d'énormes difficultés même chez l'humain pour la compréhension du sens de la phrase.
- ***Anaphores grammaticales*** : En grammaire, une anaphore est un mot ou un syntagme qui dans un énoncé, assure une reprise sémantique d'un précèdent segment appelé antécédent. Elle permet de donner une structure hiérarchique au discours tout en gardant un lien avec tous les éléments présents. C'est par cette continuité qu'on est en mesure de bien comprendre le sens d'une phrase. L'usage de l'anaphore grammaticale permet d'éviter la répétition lexicale. L'anaphore est un procédé fondamental qui participe à la cohérence d'un texte.
- ***Polysémie*** : La polysémie est la caractéristique d'un mot ou d'une expression qui a plusieurs sens ou significations différentes. Dans cette situation, le sens d'un mot dépend du contexte de la phrase dans lequel on le trouve. Certains mots peuvent être utilisés au sens propre ou figuré. Le sens propre est l'utilisation principale du mot ; il est utilisé dans un cadre concret. Tandis que le sens figuré du mot est utilisé dans un autre cadre que son cadre habituel, souvent de façon imagée.

- ***Cas particuliers*** : nous avons d'autres cas d'ambiguïtés sémantiques qui mettent en comparaison deux mots ou deux expressions ; il s'agit, de l'homonymie, de la synonymie et de l'antonymie. L'homonymie désigne deux ou plusieurs mots qui ont la même forme phonique (ou graphique), mais qui n'ont pas le même sens. Tandis que deux mots sont synonymes quand ils ont des écritures différentes, mais ont des significations très semblables. À l'inverse de la synonymie, deux mots sont antonymes quand l'un signifie le contraire de l'autre. En outre, le sens de certains mots peut dépendre des préférences et de l'idéologie de l'auteur. Ces cas sont souvent rencontrés dans les textes politiques.

En fouille d'opinions, l'ambigüité sémantique entraine souvent un détournement du sens originel de termes qui entraine des résultats erronés.

3.3.1.3 - Ambiguïté lexicale

Le lexique d'une langue constitue le vocabulaire de la langue c'est-à-dire l'ensemble des lemmes d'une langue. Dans le cas des commentaires en ligne, l'ambiguïté lexicale est causée par des abréviations personnalisées. En effet, une abréviation conventionnée est considérée comme un raccourcissement de mot ou de groupe de mots, représentés alors par un caractère ou un groupe de caractères issus de ce mot. L'abréviation personnalisée consiste donc à condenser l'écriture d'un mot selon la seule convenance du locuteur. Dans les commentaires, ces types d'abréviations sont nombreux et causent beaucoup d'entraves aux outils de traitement de ces textes. À côté d'obstacles causés par l'ambiguïté, les commentaires sénégalais contiennent des informations que nous qualifions d'obstacles liés au multilinguisme.

3.3.2 - Obstacles liés au multilinguisme

3.3.2.1 - Notion de multilinguisme

Il convient de partir d'une définition préliminaire du concept. Le mot multilinguisme (ou plurilinguisme) décrit le fait qu'une communauté (ou personne) soit capable de s'exprimer dans plusieurs langues. Le multilinguisme est un phénomène complexe qui résulte de la cohabitation de langues [121]. Sous cet aspect, le multilinguisme est l'analyse des différentes formes de coexistence des langues à l'intérieur d'une communauté linguistique et de la compétence d'un rédacteur. Il est, en particulier, la croissante composition multi-ethnique et multiculturelle de notre pays. Pour nous, il s'agit de mettre en exergue les obstacles causés par

la cohabitation de langues étrangères et nationales dans les commentaires issus de la presse en ligne.

3.3.2.2 - Cohabitation de langues étrangères et nationales dans les commentaires en ligne

Au Sénégal, vingt-cinq (25) langues nationales cohabitent avec le français, l'arabe et d'autres langues étrangères[10]. Ces langues nationales s'imposent de plus en plus dans les débats télévisés, les émissions radio, les panneaux publicitaires et même dans les commentaires en ligne. C'est la raison pour laquelle, nous rencontrons dans ces commentaires des vocabulaires empruntés aux langues nationales telles que le wolof, le pulaar, le mandingue, etc. Pour appuyer ce propos, nous donnons en exemple deux extraits de commentaires : le premier est un mélange de texte français en noir et wolof en bleu (voir Illustration 11) et le second est entièrement composé de mots empruntés du wolof (voir Illustration 12).

Illustration 11 : Extrait de commentaire bilingue (français-wolof)

« Sénégal vraiment dafa diote gnou holate sougnou bopp finalement gnoune sah hamougnou lougnou beugu. Ngirr yalla rék pour cette fois faisons un vote objectif… »

Illustration 12 : Extrait de commentaire entièrement wolof

« gno tey! rew mi macky moko môme koumou nekhoul rek kharoul gno teyyyy.macky rek …»

C'est un phénomène très récurrent qui entrave la compréhension du sens réel des phrases même au niveau de l'humain. À l'heure où nous parlons les langues nationales sont peu dotées d'outils de fouille de textes notamment en termes de ressources. À côté de ces obstacles, ceux liés au multi-domaine sont aussi constatés.

3.3.3 - Obstacles liés au multi-domaine

Nous définissons le multi-domaine dans les commentaires journalistiques comme un commentaire qui ne porte pas sur la même thématique que l'article auquel il est associé ou des réponses incohérentes. Dans tous les cas, il s'agit de commentaires qui sont sensés abonder dans le même sens que l'article auquel ils sont associés en confirmant ou en infirmant. Malheureusement, ces commentaires ouvrent le débat sur d'autres sujets. Nous qualifions ces

[10] Selon la Direction de l'Alphabétisation et des Langues Nationales au Sénégal

types de commentaires comme des commentaires hors contexte ou hors sujet c'est-à-dire en dehors du contexte abordé. Le « hors-contexte » ou « hors-sujet », dans sa forme la plus pure, revendique explicitement le droit de parler d'autre chose que du sujet traité pour prendre position par rapport à une préoccupation.

Dans le cas de commentaires journalistiques, les critiques peuvent être considérés comme autant de prises de position dans des débats publics sur le destin collectif des sénégalais ou sur la difficile transition pour telle ou telle catégorie sociale. À ce niveau, l'article devient le prétexte d'un débat existant, mais occulté. C'est une autre manière plus évidente et plus sûre pour les lecteurs d'émettre leurs points de vue sur les questions qui les préoccupent. Dans la presse en ligne, certains lecteurs visent à légitimer leurs points de vue exprimés dans l'espace de discours pour guider les autres lecteurs dans la compréhension de l'actualité [122]. Ces types de commentaires sont exprimés avec des arguments pour ou contre l'exactitude d'un article qui traite un sujet précis. Ce faisant, ces lecteurs se glissent subrepticement dans la peau de gens simples [123]. Pour eux, il ne s'agit pas de parler de scénario ou de réalisation, c'est une question de vie dont il s'agit.

Pour illustrer ces propos, nous donnons en exemple un extrait de commentaire politique retrouvé dans la rubrique sport. L'extrait parle du soutien des candidats perdants de l'élection présidentielle 2019 à M. Idrissa Seck alors que l'article auquel il est associé traite du racisme dont le joueur français Kylian Mbappé est victime (voir Illustration 13).

Illustration 13 : Texte sur la politique dans la rubrique sport

> *« Merci à la grande coalition Pr Amsatou Sow, Abdoul Mbaye, Hadjibou Soumare, Cheikh Bamba, Kalifa Sall, vous avez fait le choix de l'intellect, le de la patrie… »*

En effet, l'analyse d'opinions sur des données contenant plusieurs entités nécessite une attention particulière, car il existe des mots ou groupes de mots dont les sens dépendent de leur domaine. Certains domaines ont des vocabulaires intrinsèquement positifs ou négatifs.

3.3.4 - Hétérogénéité des structures DOM (Document Object Model)

Le DOM est un modèle d'objets de document (page web) qui est utilisé par W3C[11] comme le modèle standard de structuration d'un document HTML[12] ou XML[13]. Ce modèle définit la composition d'une page web sous forme d'arborescence à travers des sections de découpage reliées entre elles afin de faciliter l'interprétation du document par navigateur d'une part et de rendre le site attrayant d'autre part. Dans cette logique, chaque concepteur définit l'architecture de ses pages en se fondant sur le modèle DOM.

Au niveau des portails sénégalais d'informations web, nous avons constaté que la structure des pages web diffère d'un site à l'autre. Par ailleurs, l'acquisition de données se fait principalement en utilisant des outils de web scraping. Le web scraping est une technique qui consiste à extraire des contenus du web en utilisant un script ou un programme. Il se base sur la structure DOM d'un page web pour en extraire les contenus.

Dans la littérature, nous avons trouvé peu de travaux destinés à extraire des données à partir des sites d'informations [23][24]. Ces travaux n'ont pris en compte l'acquisition de commentaires en vue de la fouille d'opinions à cause de l'hétérogénéité des structures DOM. Dès lors, l'acquisition des articles et des commentaires associés devient l'une des principales difficultés qui nécessite une solution adaptée.

3.4 - **Discussion**

Dans un élan de discussion, nous menons des critiques à l'encontre d'outils existants en fouille d'opinions afin de justifier notre choix.

3.4.1 - Limites des outils existants

Les ressources et méthodes proposées dans la littérature sont destinées à analyser des textes écrits dans des langues officielles (françaises, anglaises, portugaises, etc.). L'avantage des langues officielles réside dans leur caractère formel et structuré. Dans une telle circonstance, il y a des mots à valeur intrinsèquement positive ("généreux, délicieux") et d'autres à valeur intrinsèquement négative ("avare, mauvais") [124]. Ces langues officielles ont une syntaxe et des vocabulaires standards. Ces vocabulaires sont considérés comme des expressions d'évaluation universelles.

[11] http://www.standard-du-web.com/world_wide_web_consortium.php
[12] [12] https://fr.wikipedia.org/wiki/Hypertext_Markup_Language
[13] https://fr.wikipedia.org/wiki/Extensible_Markup_Language

En somme, les solutions de fouille d'opinions portent sur des textes bien structurés écrits dans des langues officielles. En comparant ces langues officielles au langage urbain, nous constatons que nos langues sont peu dotées de ressources et méthodes. En plus, l'efficacité des algorithmes de fouille de textes est fortement dépendante des outils utilisés et de la langue traitée. En outre, les spécificités propres à chaque type de données nécessitent un traitement particulier. Ainsi, utiliser un système de fouille d'opinions entraîné pour la langue anglaise ou française sur une collection de commentaires sénégalais créerait beaucoup de bruits. En guise d'illustration, nous avons extrait un échantillon de ces commentaires et nous avons utilisé TreeTagger [25] dont les statistiques sont fournies dans le Tableau 3.

Tableau 3 : Statistiques de POST

Nombre d'articles	Nombre de commentaires	Termes en français	Termes en anglais	Termes inconnus
1	264	49%	5%	46%

TreeTagger est un outil open source, performant, robuste et efficace pour l'étiquetage morphosyntaxique et la lemmatisation de textes écrits en français [125]. Nous constatons qu'il y a beaucoup de tokens inconnus (mots, ponctuation, etc.). Il faut rappeler que les tokens inconnus peuvent être des tokens invalides induits notamment par des erreurs de segmentation (*tokenisation*), des tokens dont les orthographes sont inconnus, des tokens dont les mots sont des emprunts, des inconnus lexicaux ou typographiques [126].

Un problème proche de l'adaptation de la langue est l'adaptation au niveau du domaine. L'influence du domaine sur l'opinion est un enjeu crucial. Il existe des mots dont l'orientation peut changer selon le contexte dans lequel ils sont employés [127]. Il peut s'agir de mots polysémiques ou bien d'homonymes ayant des orientations différentes. Dans une telle situation, la désambiguïsation sémantique s'appuie justement sur les mots du contexte [128]. L'orientation d'un mot non polysémique peut également changer à l'intérieur d'un même domaine, selon l'objet qu'il évalue. En outre, l'orientation des mots peut aussi dépendre des préférences et de l'idéologie de l'auteur et c'est alors bien plus difficile à détecter. Les textes politiques sont notamment très sensibles à cela. C'est dans ce contexte que nous envisageons de proposer des solutions idoines.

3.4.2 - Nos choix

À titre de synthèse, nous faisons dans ce tableau une récapitulation des trois (03) approches précédemment présentées (voir Tableau 4).

Tableau 4 : Récapitulation d'approches de fouille d'opinions

Approches	Identification d'opinions	Classification
Approche lexicale	Dictionnaires ou Lexiques d'opinions	Score mesuré en fonction de la présence de termes issus de ces dictionnaires dans le document
Approche par apprentissage	Corpus d'entrainement ou Corpus d'apprentissage	Algorithmes d'apprentissage supervisé
Approche hybride	Utilisation simultanée des deux approches	

Ici, nous allons d'abord présenter dans ce tableau synoptique un résumé des trois approches en insistant sur les avantages et inconvénients de chaque approche afin de nous permettre de mieux choisir (voir Tableau 5).

Tableau 5 : Comparaison d'approches de fouille d'opinions

Approches	Avantages	Inconvénients
Approche basée sur le lexique	○ Facilitation dans l'identification des vocabulaires subjectifs ○ Rapidité dans la détermination de l'orientation des documents (somme des valeurs : positives et négatives) ○ Précision dans la prédiction de la polarité d'un terme	○ Affectation des mêmes polarités aux mêmes termes quel que soit le domaine (dictionnaire de sentiments) ○ Non prise en compte de groupes de mots (dictionnaire de sentiments) ○ Difficulté dans la constitution de lexique d'opinions : tâche dévolue aux experts des domaines ○ Complexité dans la gestion des négations
Approche par apprentissage	○ Moins d'effort humain ○ Automatisation de la classification d'opinions	○ Difficulté dans la constitution de corpus d'entrainement : processus manuel

Approche hybride	○ Processus sûr et fiable	○ Difficulté dans l'implémentation

À la lecture du tableau 3, l'approche hybride peut s'avérer performante dans un contexte où les langues sont peu dotées d'outils pour le traitement automatique de ces langages naturels. Dans pareille situation, les données présentent beaucoup de non-conformité par rapport à la grammaire des langues normalisées, ces étapes requièrent des tâches difficiles et fastidieuses. Pour ces raisons, nous avons opté pour l'utilisation des deux approches de manière hybride.

Notre motivation est de mettre en place des ressources et méthodes pour analyser les commentaires issus de la presse sénégalaise en ligne dans le but de déterminer l'opinion majoritaire des lecteurs. En termes de ressources, nous envisageons de créer un lexique d'opinions et une ontologie d'évènements composés de termes et de concepts issus du langage urbain sénégalais. Quant aux méthodes, nous pensons à réadapter les algorithmes de l'apprentissage supervisé pour enrichir les ressources.

3.5 - **Conclusion**

En somme, faire participer le lecteur au débat, c'est lui permettre de discuter des thématiques d'utilité publique ou privée. Le lecteur prend ainsi conscience et ne se réduit plus au simple consommateur. Mais il se considère comme un grand producteur d'informations en exprimant ses opinions par rapport aux interpellations quotidiennes. À travers cette collaboration, les commentaires en ligne deviennent de plus en plus abondants sur internet et restent des éléments précieux et utiles parmi les types de données.

La fouille d'opinions sur ces types de commentaires permet de manière dynamique de déterminer l'opinion majoritaire des internautes. Cependant, les solutions proposées en fouille d'opinions sont adaptées aux données dont les textes présentent peu de complexités linguistiques. L'application directe de ces solutions existantes sur les commentaires sénégalais entrainerait beaucoup de bruits. C'est dans cette optique que nous envisageons une solution idoine afin d'analyser les commentaires de la presse sénégalaise en ligne. Dans le chapitre suivant, nous allons proposer une architecture d'un système de fouille d'opinions.

4 -MODÉLISATION DE COMMENTAIRES JOURNALISTIQUES POUR LA FOUILLE D'OPINIONS

4.1 - **Introduction**

La fouille d'opinions sur les aspects est souvent très utilisée dans les applications pratiques, car elle fournit des informations détaillées sur différents aspects de l'entité et permet une analyse plus fine, plus efficace et fiable. Avec cette approche, Liu et al. [77] ont proposé une représentation de l'opinion autour de l'entité, de la source, des aspects de l'entité, du temps et de la valeur. Cependant, l'implémentation de cette approche dans la pratique est un processus complexe et difficile à réaliser. À cela s'ajoute toute la complexité des commentaires issus de la presse sénégalaise en ligne. En plus, la vitesse de production de l'information exige un traitement en temps réel de ces données. Dans l'esprit d'innovation, nous avons tenté de repenser la problématique afin de proposer une solution idoine permettant de rendre les commentaires journalistiques en ligne intelligibles et facilement exploitables.

Comme dit **Le Moigne** « *Un système compliqué, on peut le simplifier pour découvrir son intelligibilité. Un système complexe, on doit le modéliser pour construire son intelligibilité.* » [129]. C'est la raison pour laquelle nous proposons la modélisation d'un commentaire et d'un réseau de commentaires journalistiques [130]. Nous définissons la modélisation par la conception de l'information contenue dans le système afin de structurer le stockage [131][132]. En d'autres termes, il s'agit essentiellement de proposer une représentation formelle d'un commentaire journalistique et une description des relations entre commentaires. Dans une telle situation, nous adoptons la solution « clé-valeur » qui propose des formats de données légers comme le JSON pour le stockage. Cette modélisation est conçue à l'origine pour répondre aux besoins de performance de bases de données face à des données volumineuses et hétérogènes. La pérennité de l'information est un autre élément justifiant la décision de modéliser le système. Ce type de représentation des commentaires journalistiques peut satisfaire autant que possible nos préoccupations.

L'intérêt de la modélisation de commentaires est multiple : (i) résolution de la complexité d'extraction des entités et leurs aspects; (ii) résolution des obstacles liés au multi-domaine et (iii) facilitation de la collecte.

Ce chapitre est structuré autour de trois (03) grandes sections, à savoir la modélisation dans la presse en ligne, la modélisation d'un commentaire journalistique et la modélisation d'un réseau de commentaires.

4.2 - **Modélisation dans la presse en ligne**

4.2.1 - Presse en ligne

L'avancée des technologies de l'information et de la communication (TIC) notamment le web 2.0 [133] a entrainé d'énormes potentialités interactionnelles. Cette émergence permet à la fois un accès direct aux données et une meilleure appropriation de ces données via de nouveaux modes de traitements et de visualisation. Elle permet aussi l'interactivité en donnant aux internautes la possibilité de trouver, d'organiser, de partager et de créer de l'information d'une manière à la fois personnelle et globalement accessible. C'est dans ce contexte qu'est née la presse en ligne.

La presse en ligne s'inscrit pleinement dans la tradition journalistique consistant à aller chercher de l'information brute pour la présenter de manière adéquate au public. A la différence de la presse traditionnelle, la presse en ligne définit un processus qui met en relation le journaliste et le lecteur dans un cadre de perpétuel échange. A ce niveau, les lecteurs ne sont plus des consommateurs, mais participent de façon dynamique à la génération d'informations qui pourraient intéresser le public. Dès lors, l'attention de plus en plus grande portée aux lecteurs et à ses attentes, peut être interprétée comme une contribution au renouvellement du débat social [134]. Cette nouvelle tendance plus générale crée une promotion de participation et d'autonomie des lecteurs.

Dans cette situation, nous nous sommes penchés sur les opportunités que pourraient offrir les commentaires issus de la presse en ligne de manière générale. Nous avons constaté que ces commentaires associés aux articles journalistiques sont particulièrement révélateurs. La richesse de ces commentaires peut alimenter la dimension active de ces sources d'une part et permettre de déterminer les avis de lecteurs d'autre part. Par conséquent, ces portails web peuvent être considérés comme une source privilégiée pour la fouille d'opinions.

4.2.2 - Modélisation

Le concept de modélisation fait l'objet de nombreuses définitions. Chaque discipline tente de donner une définition qui lui est propre en fonction des objectifs visés. La modélisation informatique désigne la conception de l'information contenue dans le système afin de structurer le stockage [131][132]. Ce type de modélisation part d'une représentation abstraite, dans le sens où les valeurs des données individuelles observées sont ignorées au profit de la structure, des relations, des noms et des formats pertinents de stockage. Ainsi, il existe plusieurs types de

modèles de formalisation des données. Dans ce rapport, nous mettons l'accent sur le modèle clé-valeur.

Le modèle clé-valeur utilise une technique plus souple pour formaliser des données non structurées (documents textuels) [135]. Il stocke les données sous forme de paires clé-valeur dans laquelle une clé sert d'identifiant unique [136] [137]. Les clés et les valeurs peuvent se présenter sous toutes les formes, des objets simples aux objets composés complexes. Dans la plupart des cas, la clé et la valeur sont des chaînes de caractères quelconques. En effet, la modélisation clé-valeur vise à formaliser et stocker un ensemble de données non structurées pour maintenir la flexibilité et le passage à l'échelle en évitant l'opération de jointure [139]. Le choix du modèle clé-valeur est la réponse à cette question : quel modèle peut satisfaire les contraintes énoncées sur les commentaires de la presse en ligne en générale et le cas du Sénégal en particulier.

Au regard des difficultés inhérentes à ces commentaires et surtout de la vitesse de production de l'information, le traitement en temps réel de ces données sera rendu possible grâce à un modèle qui peut structurer les commentaires en utilisant une représentation exclusivement tabulaire de l'information journalistique. Dans une telle situation, nous adoptons le formatage clé-valeur qui propose souvent des formats de données légers comme le XML ou le JSON pour le stockage. Cette modélisation est conçue à l'origine pour répondre aux besoins de performance des bases de données face à des données volumineuses et hétérogènes. La pérennité de l'information est un autre élément justifiant la décision de modéliser le système. Ce type de représentation des commentaires journalistiques peut satisfaire autant que possible nos préoccupations.

4.3 - **Modélisation d'un commentaire journalistique**

4.3.1 - Principe

Pour modéliser un commentaire journalistique, il faut juste préciser des objets et leurs propriétés et les relations entre ces objets. Cela dépend de la problématique à résoudre qui relève de plusieurs niveaux d'abstractions :

- **Premier niveau** : Il s'agit de manipuler les commentaires journalistiques afin qu'ils répondent aux traitements spécifiques avec un besoin d'efficacité des algorithmes.
- **Deuxième niveau** : Ce niveau repose sur l'utilisation d'un modèle qui sert à décomposer l'information et à l'organiser suivant certains types de base pour former des catégories.

- **Troisième niveau** : La troisième considération consiste à proposer un modèle qui offre une très forte structuration. Cette structuration permettra de faciliter le stockage et le traitement de ces masses de données non structurées avec autant de perspectives.

4.3.2 - Représentation formelle de commentaires journalistiques

Considérons un commentaire journalistique comme un avis qui porte sur un article publié sur un portail web dédié à l'information. Le commentaire est émis par un lecteur à un moment précis dans le temps. Ce commentaire lui-même peut faire l'objet de réponse, ainsi de suite.

Partant de cette considération, nous pouvons identifier les concepts suivants : le commentaire, le lecteur, l'article et les réponses. Cette façon de décrire les objets est souvent la représentation que l'on rencontre dans des bases de données traditionnelles. Dans ces types de représentation, les valeurs sont regroupées en catégories et le nom de l'attribut décrit le rôle de la catégorie dans la description de l'objet. On peut étendre cette description à travers une représentation « attribut-valeur » en considérant chaque paire comme élément atomique.

Les attributs doivent permettre de regrouper plus clairement les éléments par concepts ou rubriques dans le but d'obtenir des jeux de données corrects et cohérents. Les concepts ou rubriques doivent capturer les principales idées, connaissances ou attitudes exprimées dans le texte. Ces attributs peuvent être numériques ou catégorielles dont les valeurs peuvent être prises sous forme de distributions ou d'intervalles [142].

En tenant compte de tous ces paramètres, nous modélisons un commentaire journalistique comme un objet à huit dimensions comme le montre la figure ci-dessous (Figure 15).

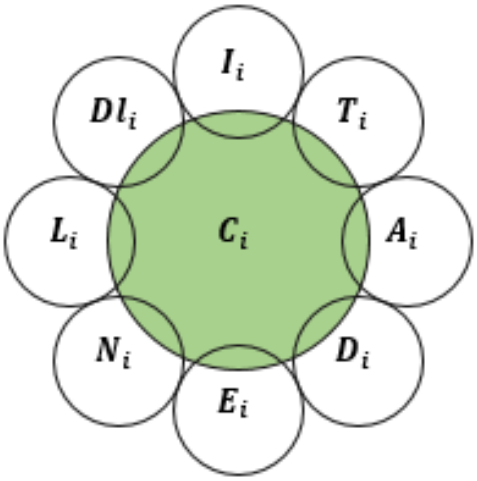


Figure 9 : *Les 08 dimensions d'un commentaire*

Formellement, cette représentation graphique peut être traduite comme suit :

$$C = \{I, T, A, D, E, N, L, Dl\}$$

- I : Identifiant ; comme le nom l'indique, il sert à identifier le commentaire C dans un ensemble de commentaires. Il est de type chaîne de caractère.
- T : Texte ; il représente le contenu du commentaire C. Il est composé de texte brut qui peut être une ou plusieurs phrase(s).
- A : Auteur ; l'auteur est la personne qui exprime le commentaire C. Les lecteurs commentent pour la plupart sous anonymat des articles de la presse sénégalaise en ligne.
- D : Date ; c'est la date à laquelle le commentaire a été publié. Elle est souvent au format jj/mm/aaaa à hh :mm.
- E : Entité ; l'entité est l'objet sur lequel porte le commentaire. Elle constitue la cible du commentaire. L'entité peut être un article publié ou un commentaire dans le cas d'une réponse.
- N : Nombre ; c'est le nombre de commentaires associés à l'entité.
- L : « Like »; "Like" signifie "aimé". Il s'agit du nombre de fois que l'entité a été appréciée positivement par les autres lecteurs.
- Dl : « Dislike » ; Dislike veut dire le nombre de fois qu'une entité a été refusée, rejetée ou simplement appréciée négativement par les lecteurs.

Figure 10 : *Représentation formelle d'un commentaire journalistique*

Un modèle se rapporte toujours à ce qu'on espère en déduire. De même, un modèle n'est jamais parfait, ni totalement représentatif de la réalité : le choix de paramètres et de relations qui les lient éclaire la finalité. Ainsi, nous pouvons sans difficulté établir les différentes relations qui peuvent exister entre les commentaires. L'ensemble de ces relations nous permettront de représenter un réseau de commentaires.

4.4 - **Modélisation d'un réseau de commentaires journalistiques**

L'analyse de réseau peut être définie comme l'étude d'un phénomène relationnel. En d'autres termes, il s'agit de manipuler des objets et d'établir des liens entre ces objets. La

modélisation d'un réseau de commentaires s'inscrit dans cette logique. Il s'agit d'une ou de plusieurs méthodes de description de relations entre les commentaires (voir Figure 17).

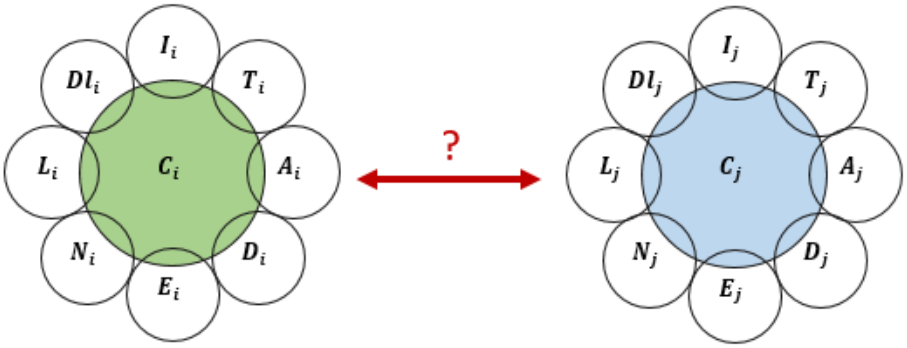


Figure 11 : *Relation entre deux commentaires*

À présent, nous proposons ci-dessous la description formelle de cette définition.

Soit C_i^j le commentaire principal $j \in N$ de l'article i. D'autres internautes peuvent aussi réagir soit sur l'article principal i c'est-à-dire produire d'autres commentaires principaux C_i^1 C_i^2 ..... C_i^N ou réagir sur un commentaire principal et dans ce cas, nous parlons de sous commentaire S_n^k avec $n \in N$ étant le commentaire principal (le niveau) et k le sous commentaire. En d'autres termes :

- C_2^3 : Est le commentaire principal de numéro 3 de l'article de presse 2.
- S_3^6 : Est le sous commentaire de numéro 6 du sous commentaire 3.
- S_7^2 : Est le sous commentaire de numéro 2 du sous commentaire 7.

Figure 12 : *Description formelle des relations entre les commentaires journalistiques*

4.4.1 - Relations entre commentaires

Nous décrivons les relations de manière récursive sur le mode : (i) un commentaire principal (C) peut avoir un ou plusieurs sous-commentaires, (ii) Un sous commentaire (S) peut lui aussi avoir un ou plusieurs sous commentaires (iii), ainsi de suite (voir Figure 19).

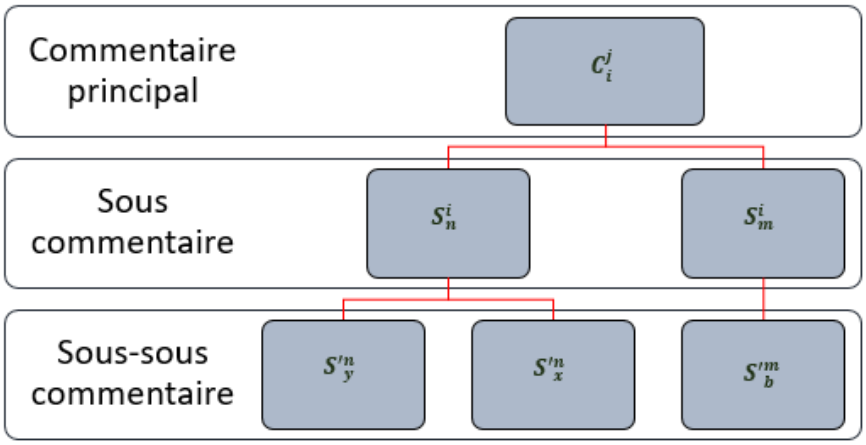


Figure 13 : *Arborescence de commentaires journalistiques*

Nous identifions les étroites relations « père-fils » et « frères » qui lient les commentaires d'un article de presse. Pour déterminer ces relations, nous nous intéressons aux entités (E) de commentaires (C). Pour rappel, l'entité est l'objet sur lequel porte le commentaire, c'est-à-dire la cible du commentaire. L'entité peut être un article ou un commentaire dans le cas d'une réponse. Nous complétons ces deux relations par des mesures de similarités afin de regrouper les commentaires par thématiques.

4.4.1.1 - Relation « père-fils »

Dans notre situation, nous nous plaçons dans le contexte d'arborescence de commentaires. En effet, un arbre possède une seule et unique racine. Il est relié à d'autres nœuds qu'on considère comme ses fils par des branches ou arêtes. Tous les nœuds peuvent posséder ou non un ou plusieurs fils. En revanche, chaque fils possède un seul et unique père, à l'exception de la racine qui n'en a pas. Ainsi, l'arborescence d'une telle représentation est définie en choisissant un sommet appelé racine et en orientant les arêtes de sorte qu'il existe un chemin de la racine vers tous les autres sommets.

- **Principe** : Si un commentaire quelconque est l'entité d'un autre commentaire alors le premier représente le père et le second, le fils.
- **Formalisation** : Formellement, cette relation se traduit comme le montre la Figure 20 :

Soit $\mathcal{M}$, un ensemble de commentaires noté

$\mathcal{M} = \{ C_1, C_2, \dots, C_n \}$ et E_i une entité

$$Si\ E_i(C_2) == C_1\ alors$$

$$C_1\ est\ le\ père\ et\ C_2\ est\ le\ fils$$

Figure 14 : *Relation « père-fils »*

Nous utilisons la relation « père-fils » pour décrire la hiérarchisation. À sa suite, nous allons parler aussi de la relation de « frère ».

4.4.1.2 - Relation « frère »

Dans la représentation structurée de liens familiaux entre les personnes, un frère est celui avec qui on est uni par des liens quasi fraternels. Par abus de langage, cette représentation est utilisée dans la théorie des graphes avec l'arbre généalogique. Ce type de modélisation met en exergue des objets qui appartiennent au même groupe que l'on considère comme une famille. Dans notre contexte, il s'agit de décrire cette relation à partir d'un ensemble de commentaires.

- **Principe** : Deux commentaires qui portent sur la même entité E_i sont considérées comme « frère ».
- **Formalisation** : Formellement, nous modélisons cette relation comme suite (voir Figure 21) :

Soit $\mathcal{M}$, une famille de commentaires notée

$\mathcal{M} = \{ C_1, C_2, \dots, C_n \}$ et E_i une entité

$$Si\ E_i(C_1), E_i(C_2)\ alors$$

$$C_1\ et\ C_2\ sont\ des\ frères$$

Figure 15 : *Relation « frère »*

La relation « frères » nous permet de décrire deux (2) commentaires de même niveau. À côté de ces deux relations qui parlent des relations familiales, nous approfondissons l'étude sur des mesures de similarité pour étendre ces relations.

4.4.2 - Similarité de documents

L’étude de similarité de documents (commentaires) consiste à regrouper les commentaires ayant le même profil dans une même classe à travers une mesure de distance. L’objectif visé à travers cette étude est de faciliter la catégorisation de commentaires journalistiques en vue de les organiser par évènement. L’évènement n’est rien d’autre qu’un ensemble de commentaires qui traitent des mêmes thématiques à une période donnée. En d’autres termes, l’évènement est une collection de commentaires portant sur les mêmes entités dans un intervalle de temps précis. Pour atteindre cet objectif, nous allons d’abord définir le principe et après établir une mesure de similarité.

4.4.2.1 - Principe de similarité

En fouille de textes, la similarité est la mesure du degré de ressemblance entre des documents à l’aide d’une mesure de distance entre termes de documents. Elle cherche à déterminer les similitudes entre de documents à travers les termes qu’ils contiennent. Dans cette logique, les documents doivent être représentés par des termes pertinents appelés aussi index [53].

À travers le modèle proposé, la similarité de documents est basée sur un croisement de plusieurs attributs permettant de catégoriser les données par classes. Ces attributs doivent contenir des valeurs qui peuvent être prises sous forme de distributions ou d’intervalles. Une telle démarche fait partie de l'apprentissage non supervisé parce que les concepts ne sont pas prédéterminés et les instances utilisées pour l'apprentissage ne sont pas pré-classifiées.

4.4.2.2 - Mesure de similarité

Dans le cas des commentaires journalistiques, la ressemblance ou dissemblance entre les commentaires étant mesurées sur un ensemble d’attributs descriptifs notamment les entités(E), les contenus de textes(T) et les dates de commentaires(D). Formellement, cela se traduit par la formule décrite à la Figure 22.

Pour regrouper les commentaires en k groupes disjoints dont les classes sont inconnues à priori, nous nous appuierons sur la métrique Cosinus qui est très populaire en fouille de textes [38][39]. La métrique Cosinus se base sur les co-occurrences des documents pour déterminer leurs distances.

Soit β, ensemble de commentaires à classer

$\beta = \{ C_1, C_2, \dots, C_n\}$ caractérisés par p descripteurs (E, T, D) ;

Une dissimilarité d est une application telle que

- $d: \beta \times \beta \rightarrow \mathbb{R}^+$
- Qui vérifie les conditions suivantes :

$$d(C_i, C_i) = 0, \forall\, C_i \in \beta$$

$$d(C_i, C_j) = d(C_j, C_i), \forall\, C_i, C_j \in \beta \times \beta$$

Figure 16 : *Distance de similarité*

Cette pratique peut faciliter la recherche ou l'extraction d'informations pertinentes lors de la sélection de la base d'analyse.

4.5 - **Conclusion**

En définitive, la modélisation cherche à rendre un phénomène intelligible afin de réduire sa complexité. Modéliser la connaissance nécessite de disposer de structures sémantiques et de formalismes de représentation permettant de traduire la complexité de données. Lors d'une modélisation, il peut y avoir plusieurs modèles possibles dont chacun présente des avantages spécifiques. Dans notre contexte, le choix du modèle est lié à la nature des problèmes à résoudre. C'est ainsi que nous avons proposé un modèle de commentaire journalistique à huit (08) dimensions. La modélisation de commentaire journalistique est nécessaire pour l'acquisition de données et l'analyse de ces données. Dans le premier cas, la modélisation facilite la collecte et le stockage des données journalistiques. Dans le second cas, cette représentation résout la complexité d'implémentation de la fouille d'opinions basée sur les aspects.

Par la même occasion, nous avons aussi cherché à déterminer un réseau de commentaires à travers une relation hiérarchique et une similarité de documents. La description des relations éclaire la finalité du modèle proposé. L'implémentation du réseau de commentaires permettra, d'une part, de rapprocher des documents en vue de les organiser par thèmes et d'autre part de faciliter la visualisation.

Notre proposition peut facilement être représentée dans un langage de description de données notamment en JSON. Dans le chapitre suivant, nous allons proposer une méthode d'acquisition de données journalistiques sur la base du modèle de commentaire proposé.

5 - ARCHITECTURE D'UN SYSTÈME DE FOUILLE D'OPINIONS DANS LA PRESSE SENEGALAISE EN LIGNE

5.1 - **Introduction**

La presse sénégalaise en ligne (sites web dédiés à l'information, radios et télévisions numériques) diffuse des informations de manière structurée selon une procédure garantissant son libre accès et sa réutilisation par tous, sans restriction technique, juridique ou financière. Elle permet à la fois un accès direct aux informations et une meilleure appropriation de ces informations via de nouveaux modes de traitement et de visualisation. Ce type de communication s'inscrit pleinement dans la tradition journalistique consistant à aller chercher de l'information brute pour la présenter de manière adéquate au public. Seulement, la presse en ligne diffuse l'information à travers des portails web dédiés en mode streaming ou podcasting. Cette mutation innovante crée une promotion de participation et d'autonomie des lecteurs. À cet effet, chaque lecteur peut réagir selon ses opinions, goûts ou attentes par rapport aux articles publiés à travers des commentaires. Ainsi, les données issues de ces portails constituent un ensemble de données libres, disponibles en quantité et éparpillées à travers le web. Toutefois, leur valorisation peut se révéler plus efficace que les formes narratives non seulement pour capter l'attention des populations.

La valorisation consiste à présenter les commentaires journalistiques issus de la presse sénégalaise en ligne, puis à les analyser et expliquer afin de mettre en lumière des informations stratégiques inaccessibles par la simple lecture. En d'autres termes, la valorisation consiste à collecter, analyser et visualiser les résultats dans le but d'aider les acteurs à la compréhension et l'interprétation de ces commentaires. En raison de la complexité des commentaires issus de la presse sénégalaise en ligne [120], il urge de penser à développer des outils performants et efficaces. C'est dans ce contexte que nous envisageons un système de fouille d'opinions qui sera adapté aux types de données dont nous disposons afin de les rendre accessibles, intelligibles et renforcer la participation de lecteurs aux débats citoyens enrichissant la démocratie au Sénégal. Notre système a pour vocation d'ajouter de l'intelligence aux commentaires en provenance de la presse sénégalaise en ligne à travers un processus plus complexe qui va des données à l'information, de l'information à la connaissance et de la connaissance à la décision.

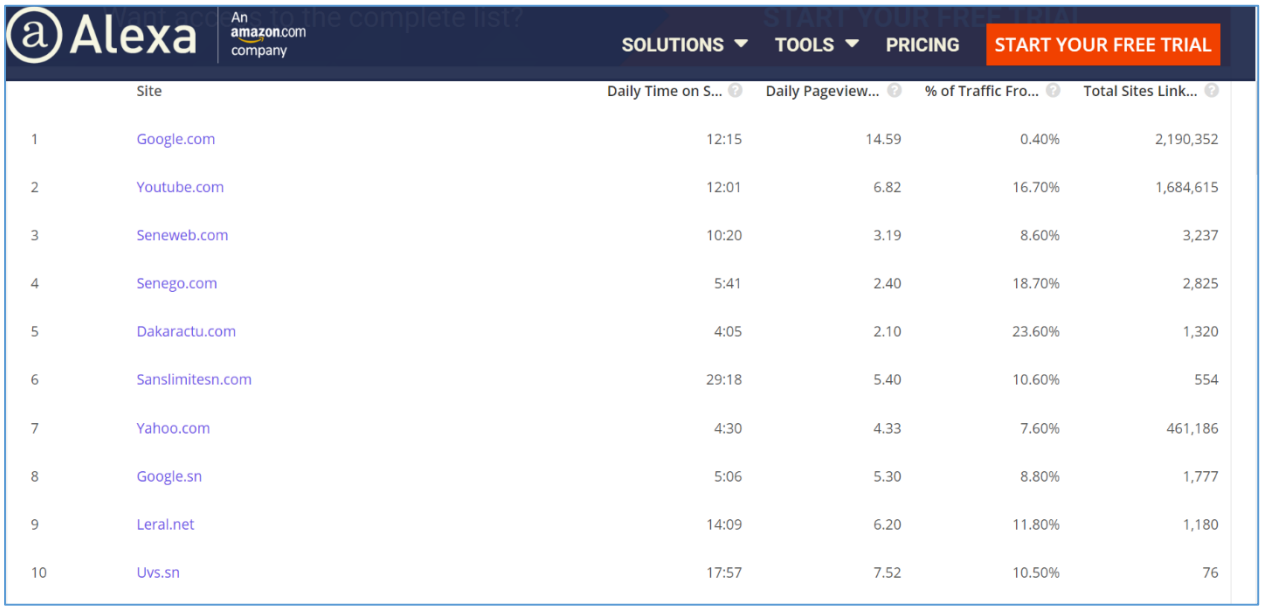
Ce chapitre a pour objectif de présenter l'architecture générale de ce système global qui a pour rôle de collecter des données, de les transformer en informations puis en connaissances afin de les présenter de façon attractive aux acteurs. Par la suite, nous allons faire la cartographie

de la presse sénégalaise en ligne d'abord, ensuite nous présenterons l'architecture et enfin nous proposons une discussion en guise de synthèse.

5.2 - Cartographie de la presse en ligne au Sénégal

5.2.1 - Généralités

Dans le domaine des technologies de l'information et de la communication, ainsi que dans le secteur des télécommunications, le Sénégal se distingue par de multiples initiatives prises par le gouvernement, en collaboration avec les acteurs, pour aménager et adapter l'environnement médiatique[14]. Ce pays a fait des avancées majeures dans le sens de promouvoir la liberté de la presse ces dernières années. Ainsi, caractérisé par la diversité et l'indépendance, le paysage médiatique sénégalais est décrit comme un milieu où la liberté d'expression est valorisée. C'est un pilier fondamental et indissociable de la marche de la démocratie d'une nation. En 2015, le gouvernement et les acteurs de médias ont trouvé un compromis pour le passage de l'analogie au numérique [143]. Cette stratégie nationale vise à faciliter l'accès aux technologies, aux services et à l'information, notamment en faveur des populations. Dès lors, nous assistions à un foisonnement de sites d'informations. En guise d'illustration, Alexa[15] propose un classement de sites d'informations comme le montre la Figure 9.



	Site	Daily Time on S...	Daily Pageview...	% of Traffic Fro...	Total Sites Link...
1	Google.com	12:15	14.59	0.40%	2,190,352
2	Youtube.com	12:01	6.82	16.70%	1,684,615
3	Seneweb.com	10:20	3.19	8.60%	3,237
4	Senego.com	5:41	2.40	18.70%	2,825
5	Dakaractu.com	4:05	2.10	23.60%	1,320
6	Sanslimitesn.com	29:18	5.40	10.60%	554
7	Yahoo.com	4:30	4.33	7.60%	461,186
8	Google.sn	5:06	5.30	8.80%	1,777
9	Leral.net	14:09	6.20	11.80%	1,180
10	Uvs.sn	17:57	7.52	10.50%	76

Figure 17 : *Classement proposé par Alexa [28/01/2020]*

[14] https://www.seneplus.com/article/un-environnement-m%C3%A9diatique-s%C3%A9n%C3%A9galais-%E2%80%9Cdivers-ind%C3%A9pendant-et-durable%E2%80%9D
[15] https://www.alexa.com/topsites/countries/SN

Alexa est un portail web d'Amazon, qui est dédié au classement de sites d'informations à travers des statistiques portant sur le nombre de pages consultées par jour, les trafics provenant de recherches et le nombre de liens partagés par les autres sites. Ce classement du 28/01/2020 fournit les résultats suivants (voir Tableau 6).

Tableau 6 : Les 8 sites sénégalais les plus populaires parmi les 50

Ordre	Nom du site
3	Seneweb.com
4	Senego.com
5	Dakaractu.com
9	Leral.net
11	Senegal7.com
16	Metrodakar.net
22	Senenews.com
50	Xalimasn.com

Avec le passage au numérique, des outils et méthodes de collecte, de traitement et diffusion de l'information ont considérablement évolué. Ces instruments concourent à promouvoir la gratuité de l'accès à l'information. Dans notre recherche scientifique, nous avons tenté d'étendre le classement d'Alexa sur la presse sénégalaise uniquement. Cela nous permet de juger la popularité de ces sites d'informations. Le classement de sites d'informations sénégalais entre eux trouve sa pertinence dans l'identification des sources lors de l'acquisition des données. Dans ce cas précis, nous avons utilisé les données collectées par Sarr et al. [24] afin de mener notre étude. Ces auteurs ont développé un scraper permettant d'extraire des articles journalistiques à partir de sites d'informations sénégalais. À partir de leurs données massives, nous avons pu identifier plusieurs sites d'informations sénégalais représentés à travers la Figure 10.

La Figure 10 montre deux informations majeures : d'une part, la multiplicité de sites dédiés à l'information au Sénégal et d'autre part, la représentation par ordre décroissant de la popularité de ces sites sénégalais. Le point culminant de la figure correspondant au site le plus populaire. À ce niveau, nous retrouvons Seneweb.com au premier plan. De la gauche vers la droite, la figure met en exergue un classement de la plupart des sites d'informations sénégalais.

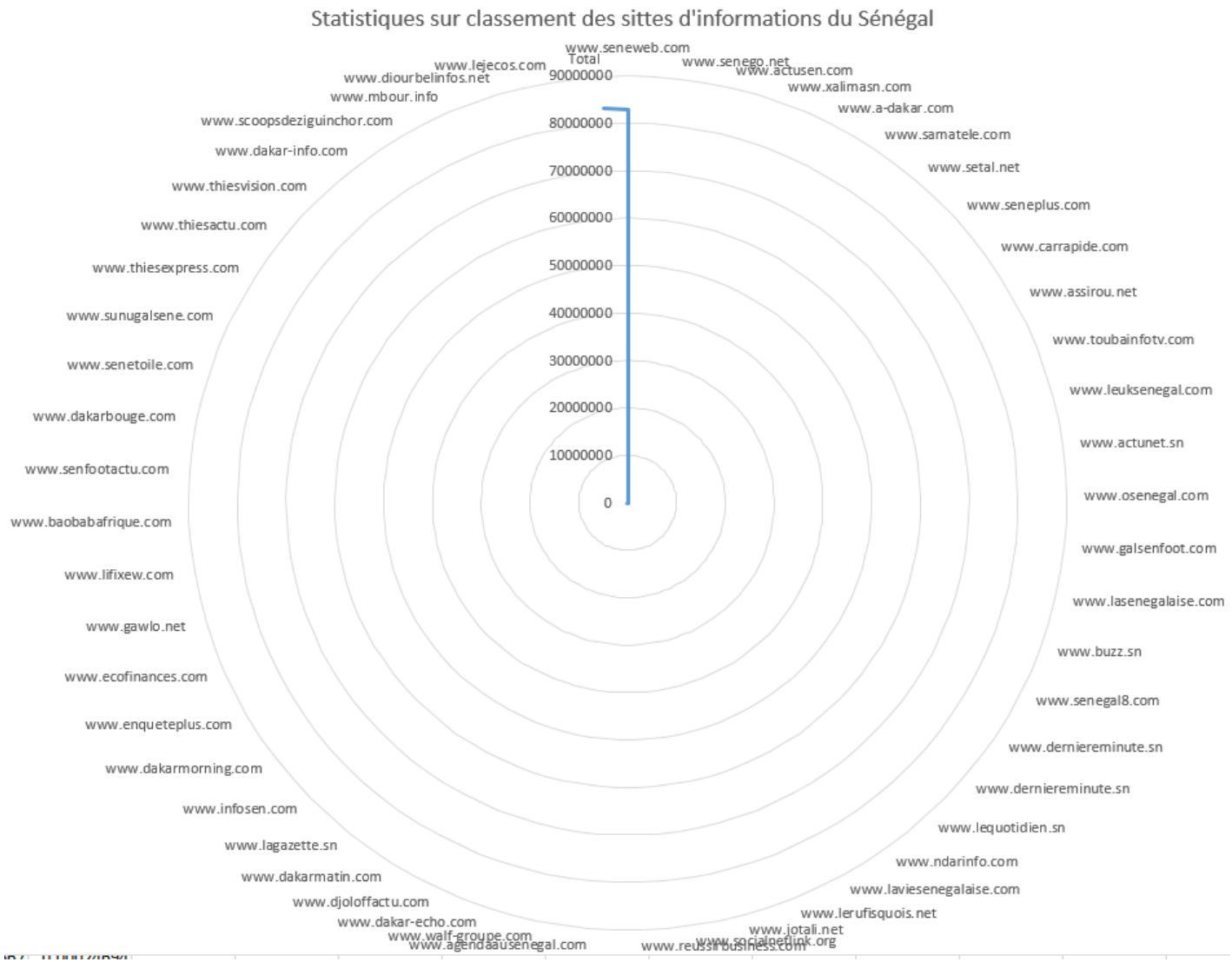


Figure 18 : *Panaroma de sites d'informations sénégalais*

Le Tableau 7 donne une illustration d'éléments statistiques sur lesquels nous nous sommes basés pour effectuer ce classement.

Tableau 7 : Statistiques sur la popularité de sites d'informations sénégalais

Libelle	Publications	Audiences	Partage	Popularité
www.seneweb.com	10246	82979014	3154	99,6194178
www.dakaractu.com	2007	200700	1241	0,24480769
www.senego.net	385	38500	3110	0,05040843
www.leral.net	308	30800	1135	0,03870268
Autres sites d'informations	12642	20200	6033	0,04666336
Total	**25588**	**83269214**	**14673**	**100**

À travers le tableau, nous pouvons lire aisément la dominance de Seneweb.com suivi de Dakaractu.com, Senego.net et Leral.net. Par ailleurs, les sites d'informations s'intéressent à tous les domaines d'activités des sénégalais. Nous abordons les domaines dans la section suivante.

5.2.2 - Domaines d'intérêt

Le paysage médiatique sénégalais est particulièrement riche et diversifié[16]. Ainsi, tous les sites d'informations reflètent les voix des populations dans leur diversité économique, politique, culturelle, religieuse, linguistique, etc. (voir Tableau 8). Autrement dit, ces sources publient dans tous les domaines d'activités des sénégalais et de la diaspora. Elles diffusent des informations sur des problématiques propres aux groupes sociaux et professionnels de la population cible.

Tableau 8 : Statistiques sur les domaines d'activité des sénégalais (extrait sur Seneweb.com)

[16] https://www.seneplus.com/article/un-environnement-m%C3%A9diatique-s%C3%A9n%C3%A9galais-%E2%80%9Cdivers-ind%C3%A9pendant-et-durable%E2%80%9D

Domaine	Publication	Audiences	Commentaire
Politique	2704	27532967	128721
Video	1542	12021020	34591
Audio	1373	4933861	4683
Societe	871	9069284	28472
Afrique	583	2934085	5904
Religion	581	5892662	10740
Justice	566	4879227	16609
International	544	3064847	7272
Sport	533	3780633	10992
Revue de presse	118	137340	149
Necrologie	113	1663703	4194
People	97	1296409	2527
Contribution	92	637215	4397
Economie	80	545781	2464
Buzz	73	1383339	2412
Diplomatie	66	606864	2460
Sante	57	224555	647
Faits-Divers	54	687344	1347
Education	53	489875	2111
Culture	34	212975	635
Media	33	282819	507
Communique	31	184150	526
Image	18	276253	550
Chronique	7	85103	353
Opinion	4	69433	176
En direct	3	52099	47
Publi-Reportage	3	7228	13
Insolite	2	20716	83
Reportage	2	33278	55
Dossier de redaction	2	5865	18
Environnement	1	52519	97
Science	1	42336	95
Telecommunication	1	11392	86
Entretien	1	16353	70
Exclusif	1	30365	42
Technologie	1	4522	2
Immigration	1	3016	1

Ces statistiques sont basées sur les mêmes données collectées [24]. Aujourd'hui, la presse sénégalaise en ligne renferme un ensemble d'informations riches et variées. À présent, nous allons faire une typologie de ces données.

5.2.3 - Typologie de données journalistiques

De nos jours, nous distinguons aisément deux types de données journalistiques à travers les portails web dédiés à savoir les informations fournies par les spécialistes qu'on appelle les articles et celles postées par les lecteurs ou les commentaires comme le montre la Figure 11.

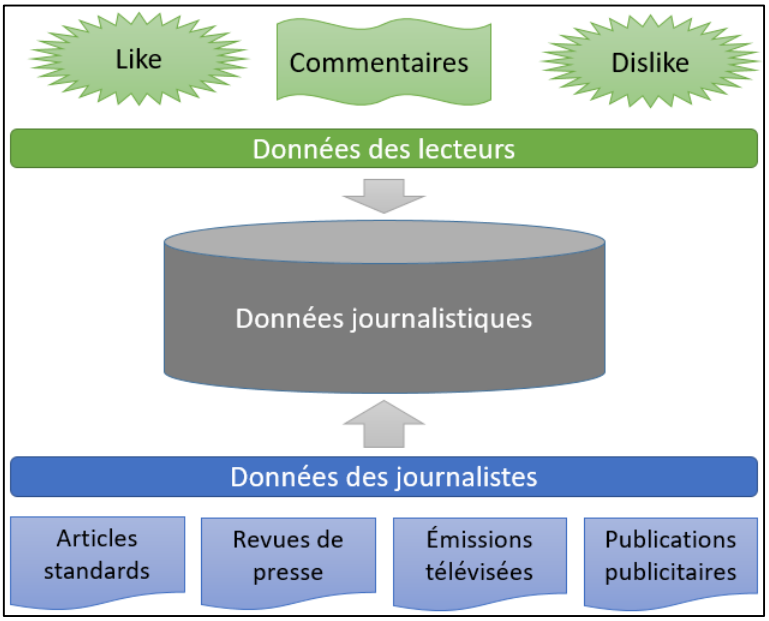


Figure 19 : *Typologie des données journalistiques*

- **Articles** : Un article journalistique présente une description objective d'un fait. C'est une narration basée sur différentes sources d'informations orales ou écrites et qui relatent un évènement. Il peut aussi présenter la position de la rédaction sur un thème d'actualité ou mettre en valeur un dossier publié dans un journal. Un article peut être un reportage, c'est-à-dire un texte dans lequel le journaliste rend compte d'un événement particulier en se rendant sur le lieu. En outre, il peut être une interview, une enquête menée sur des recherches, témoignages ou analyses pour rendre compte d'un phénomène ou d'un événement. En général, un article est composé d'un titre qui donne l'idée du sujet, d'un résumé, de mots clés et d'un contenu assimilé aux détails. En plus, des métadonnées telles que la source (auteur ou autre source), le domaine ou la rubrique (politique, économie, étranger, société, culture, sports, etc.), la date de publication, etc. sont aussi associées à l'article.
- **Commentaires** : Les commentaires post-articles constituent la forme la plus visible de la participation de lecteurs. Ces interventions sont d'autant plus profitables, car il arrive que les lecteurs soient des experts des sujets traités. Les commentaires journalistiques

sénégalais prennent des formes aussi variées que les critiques formulées. Un commentaire peut être un texte dans lequel le lecteur donne son opinion, sa position, son sentiment par rapport à un article publié ou une question qui le préoccupe. Il peut aussi être un « like » ou un « dislike » qui se traduisent respectivement par « j'aime » ou « je n'aime pas ». Ces deux concepts sont souvent des images symboliques à la forme du pouce de la main de l'humain. Si la position du pouce est orientée vers le haut alors il s'agit de « like » sinon c'est le « dislike ». Tout compte fait, les commentaires contiennent beaucoup d'informations utiles.

Pour illustrer notre argumentation, nous montrons la disposition d'informations qu'on rencontre sur les pages de Seneweb.com (Figure 12).

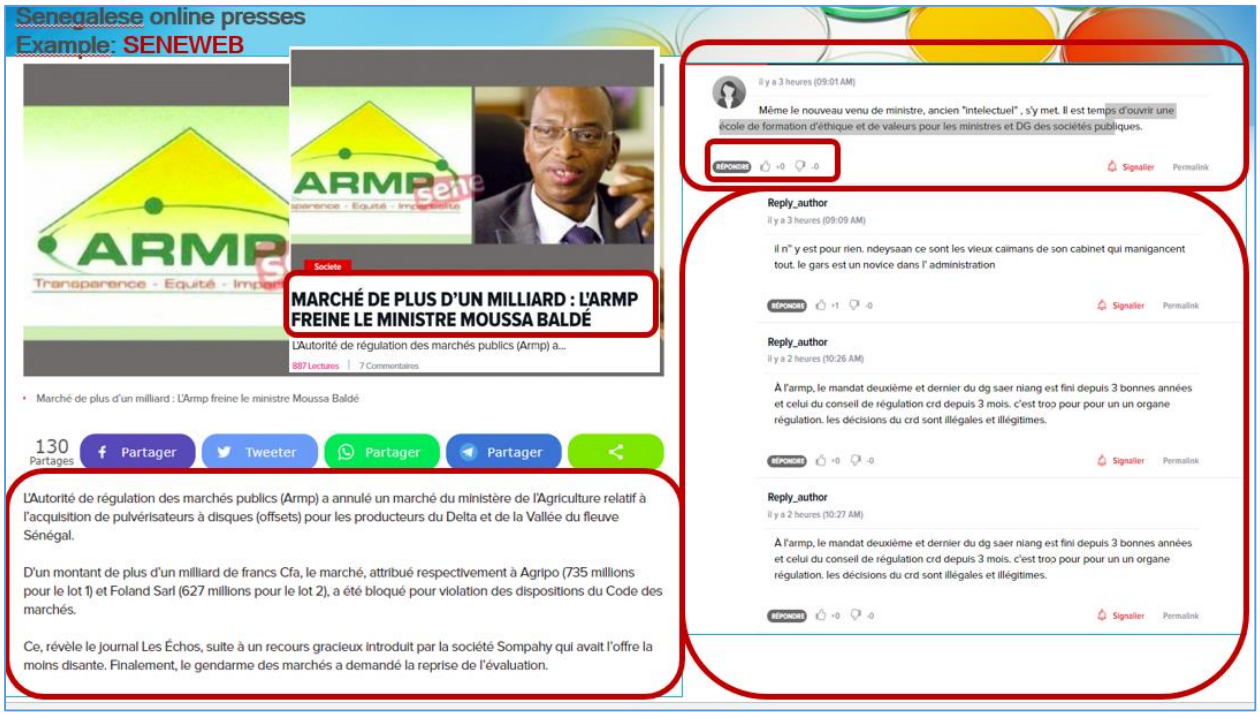


Figure 20 : *Informations sur Seneweb*

Chaque site propose une structuration qui lui est propre. La diversité de structures impacte sur la disposition des informations et rend leur valorisation complexe. Néanmoins, certains auteurs tels que Sarr et Al. [144][145] se sont intéressés à la vérification de faits dans la presse sénégalaise en ligne. Leurs travaux portent uniquement sur les articles. Bien que les articles permettent aux internautes d'avoir accès aux informations nationales et internationales, les commentaires eux restent un élément important porteur de la voix des populations. À présent, nous allons décrire les opportunités offertes par ces commentaires.

5.2.4 - Opportunités offertes par les commentaires

Notons que les commentaires issus de la presse sénégalaise en ligne peuvent être considérés comme autant de prises de position dans les débats publics. Ces types de commentaires ont un enjeu stratégique et contiennent des informations utiles pouvant aider les décideurs dans l'orientation de leurs choix. Nous pouvons y découvrir par exemple : (i) l'influence potentielle de commentaires en ligne dans les prises de décision ; (ii) l'accord ou le désaccord des internautes concernant une proposition particulière ; (iii) l'aperçu de l'opinion publique sur la situation économique, politique et sociale du pays ; (iv) et les attentes des internautes sur les politiques publiques.

Par ailleurs, les commentaires journalistiques ont une plus grande influence et une valeur plus subjective que les articles qui peuvent être laissées à l'interprétation. Ils permettent d'influencer l'opinion publique en orientant indirectement les pensées des populations. Cette technique se traduit par la mise en valeur d'idées et de représentations (factuelles ou fictionnelles) qui inévitablement façonnent notre vision de la réalité. Dans cette même logique, ils peuvent être un moyen de subversion politique, voire d'incitation à la révolte ou à la rébellion.

En réalité, les opportunités d'extraction de connaissances utiles à partir de commentaires issus de la presse sénégalaise en ligne sont innombrables. La richesse de ces commentaires peut alimenter la dimension active de ces sources d'une part et permettre de déterminer les avis de lecteurs d'autre part. Par conséquent, ces portails web peuvent être considérés comme une source privilégiée pour la fouille d'opinions. À la suite de cette partie, nous allons présenter l'architecture du système de fouille d'opinions.

5.3 - Architecture générale et fonctionnelle

L'architecture proposée est, à cet effet, sur la base d'une nouvelle approche pour résoudre la problématique relative à la complexité des commentaires issus de la presse sénégalaise en ligne. La solution décrite dans cette architecture est une plateforme web sémantique destinée à répondre à l'ensemble du processus de fouille d'opinions tel que le web scraping, l'indexation, l'étiquetage et la classification d'opinions, la mise en place d'une base de connaissances et la visualisation des résultats. Ensuite, nous procédons à la présentation de l'architecture d'une part et la description des interactions entre les modules d'autre part.

5.3.1 - Présentation de l'architecture

Notre architecture présente une suite de modules indépendants qui permettent de mettre en œuvre les principales phases du processus de la plateforme que nous comptons mettre en place (voir Figure 13).

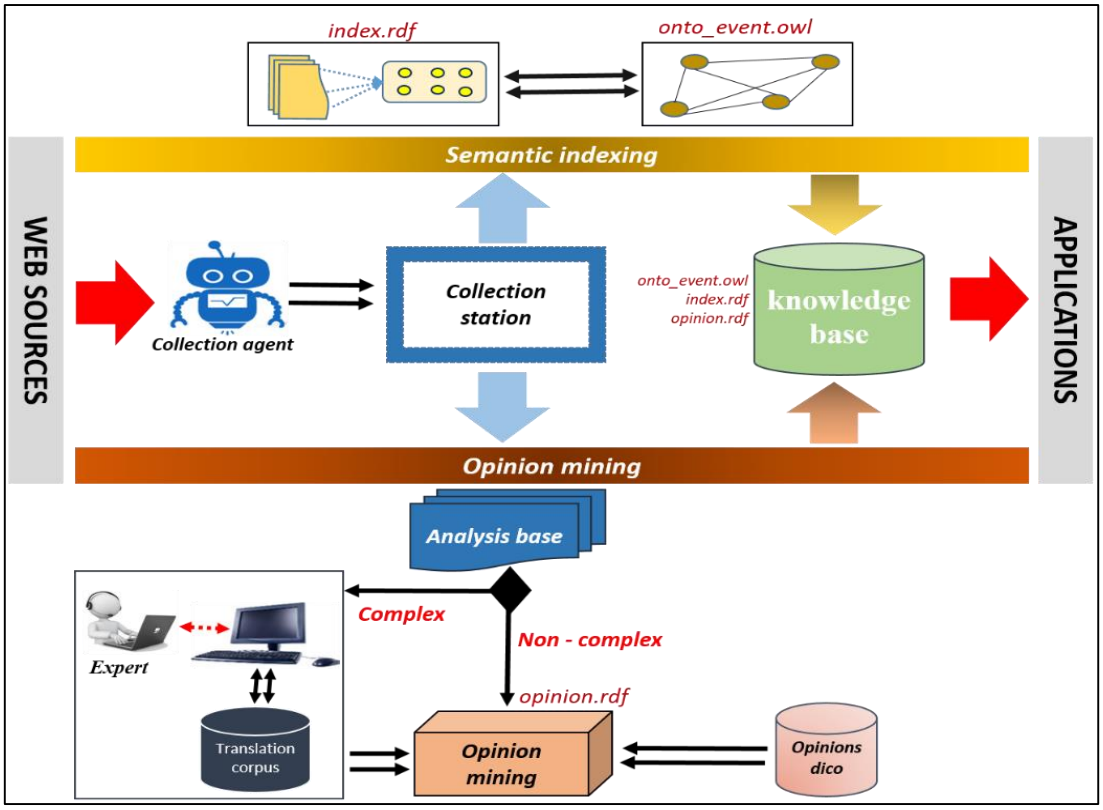


Figure 21 : *Architecture d'un système de fouille d'opinions*

Ces modules sont composés de l'acquisition de données, l'indexation sémantique, la fouille d'opinions et enfin la construction et l'alimentation de la base de connaissances [146].

- ***Module « Acquisition de données »*** : La phase d'acquisition de données est confiée à un agent de collecte appelé scrapeur qui se charge d'interroger les sources web identifiées afin d'en extraire les informations utiles et pertinentes. L'outil regroupe de manière synthétique et cohérente ces informations collectées dans une base de données. Lors de la fusion, il élimine des éléments qui ne respectent pas les règles établies afin d'obtenir un corpus propre et exploitable. En outre, ce module impose une représentation formelle des données pour faciliter le stockage au format json. D'une part, il sépare les articles et les commentaires, d'autre part il représente chaque commentaire selon ses caractéristiques à travers le modèle formel. Notre scrapeur fonctionnera comme un automate de web scraping pour des extractions ponctuelles. Il mènera des opérations récurrentes à des intervalles réduits.

- ***Module « Indexation sémantique »*** : De manière générale, l'indexation consiste à représenter des documents par des mots ou groupes de mots jugés représentatifs. Notre module « Indexation sémantique » utilise des concepts de l'ontologie d'évènements qui sera mise en place pour indexer les articles. Cette ontologie sera réalisée dans le contexte journalistique sénégalais avec comme label le langage urbain. En effet, le but de ce module est de créer un rapprochement entre les articles en se basant sur la similarité sémantique de concepts. Au-delà des synonymies ou antonymies, nous accordons une place privilégiée aux entités nommées. Dans ce contexte, les entités nommées faciliteront la description et la compréhension de documents afin d'optimiser la recherche d'informations dans la base de connaissances. Le résultat de l'indexation sémantique obtenu sera ainsi stocké dans un fichier rdf (Resource Description Framework) appelé *index.rdf*.
- ***Module « Fouille d'opinions » :*** De manière générale, la fouille d'opinions repose sur l'identification et la classification d'opinions d'un corpus. Notre module « Fouille d'opinions » a pour rôle d'analyser des commentaires relatifs à une entité sélectionnée afin de déterminer le point de vue de la majorité de lecteurs sur cette entité. Il s'agit de faire la synthèse de toutes les opinions sur cette entité en proposant la moyenne des notes attribuées à l'entité. Le résultat est stocké dans un fichier nommé *opinion.rdf*.
- ***Module « Construction de base de connaissances » :*** Une base de connaissances est une synthèse de l'expertise d'un domaine généralement formalisé à l'aide d'une ontologie. L'ontologie mise en place est stockée dans un fichier owl (Web Ontology Language) appelé *onto_event.owl*. Notre base de connaissances est construite autour des fichiers d'indexation sémantique (index.rdf), de fouille d'opinions (opinion.rdf) et d'ontologie (onto_event.owl). Cette base de connaissances permet de capitaliser les données stockées sous une forme organisée et maîtrisée. Elle permet de créer une relation sémantique entre les différentes informations intégrées afin d'optimiser la recherche d'informations.

Les modules dans l'architecture entretiennent des communications. Dans ce qui suit, nous revenons sur ces communications entre les modules.

5.3.2 - Inter-action entre les modules du système de fouille d'opinions dans la presse sénégalaise en ligne

La fouille d'opinions sur les données journalistiques est un long processus qui part de l'identification des sources à la visualisation des résultats de l'analyse en passant par la collecte et le traitement des données. Cette partie est consacrée à décrire techniquement le fonctionnement du système de fouille à travers les interactions entre les modules (voir Figure 14).

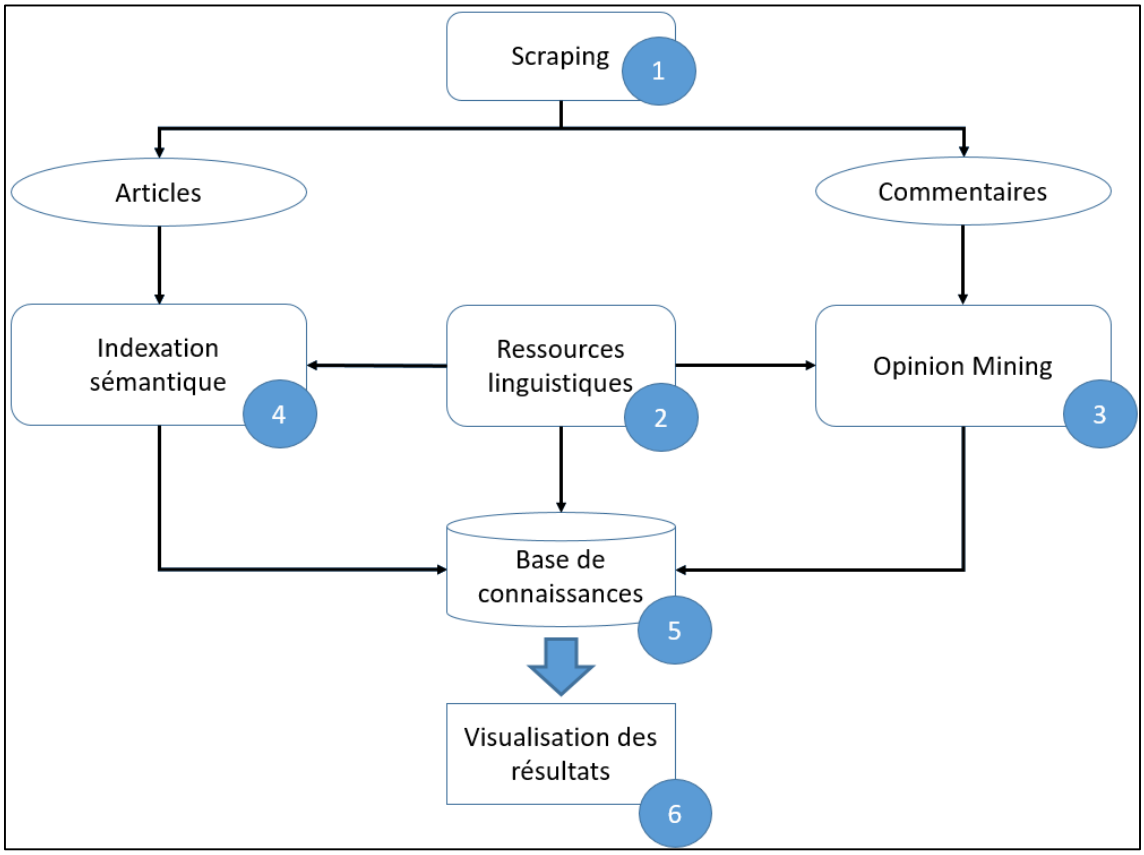


Figure 22 : *Le fonctionnement de la plateforme*

Dans la phase d'acquisition, nous avons utilisé la technique du web scraping qui consiste à interroger des sources identifiées pour extraire des contenus des pages web de manière précise. Cette technique permet de structurer les données extraites pour faciliter leur utilisation ultérieure. Ce travail est motivé par le besoin de disposer d'un corpus propre et facilement exploitable. C'est un module vital pour notre système.

Dans le même sillage, la mise en place de ressources linguistiques telles que le lexique d'opinions et l'ontologie d'évènements est nécessaire pour le fonctionnement du système. Elle a pour objectif de doter le système de ressources pour le traitement automatique de textes écrits en langage urbain sénégalais exprimés dans les commentaires en ligne.

Les modules de fouille d'opinions et d'indexation sémantique ont nécessairement besoin de ces ressources. La fouille d'opinions utilisera le lexique pour l'étiquetage d'opinions des commentaires. Tandis que l'indexation sémantique s'appuiera sur les concepts de l'ontologie envisagée afin d'indexer les articles journalistiques.

À partir des résultats issus des modules précédents, nous allons constituer une base de connaissances. Cette base de connaissances permettra la recherche et le partage des informations organisées par thématiques, par entités nommées et par tendance. Elle facilitera aussi la visualisation qui n'est rien d'autre que la représentation des données de manière simple, didactique et pédagogique.

5.4 - **Discussion**

En fin de compte, l'architecture mise en place répond à la problématique liée aux obstacles que posent les commentaires issus de la presse sénégalaise en ligne [147]. En plus, la formalisation des commentaires est un atout pour faciliter le stockage et éviter la complexité de la fouille d'opinions basées sur les aspects. Le modèle défini permet aussi de réorganiser les commentaires afin de résoudre les obstacles liés au multi-domaine. L'approche hybride dans ce contexte permet de répondre aux besoins de performance des méthodes et de fiabilité des résultats. La présence de l'ontologie donne à notre plateforme tout le caractère web sémantique.

L'adoption de l'approche web sémantique confère à notre solution un système d'extraction et de fouille d'opinions innovant et intelligent. Ce système peut offrir les services suivants : (1) des données ouvertes liées d'un journaliste web; (2) recherche sémantique sur les mots clés, les entités nommées, les évènements, les domaines ou rubriques ou sur une thématique ; (3) partage d'informations et de connaissances ; (4) statistiques pour donner des tendances.

En somme, notre architecture propose une solution particulièrement intéressante pour la fouille d'opinions de données textuelles récentes, rédigées dans des langues peu dotées. Notre système de fouille d'opinions devra être capable de fournir une synthèse d'opinions par rapport à une entité cible donnée. L'idée est de pouvoir guider le processus décisionnel d'un utilisateur en lui proposant un résumé des avis d'autres lecteurs.

5.5 - **Conclusion**

En conclusion, nous pouvons retenir que le rôle primordial de la presse sénégalaise en ligne est d'informer rapidement et largement les populations sur des faits importants, des événements nationaux et internationaux. En effet, la dimension participative observée dans ces sites d'informations est une évolution majeure et centrale qui caractérise ces sources. Cette participation des lecteurs peut s'avérer bénéfique à un autre niveau, puisque les commentaires présentent des opportunités immenses. Désormais, ces commentaires posent un enjeu stratégique pour les décideurs.

La mise en place d'un système de fouille d'opinions pour la valorisation de ces types de données peut permettre de connaître l'opinion publique des lecteurs. Ainsi, les résultats issus de cette plateforme peuvent aider les populations dans la compréhension des faits et leurs interprétations. En sus, ils peuvent donc permettre aux hommes d'affaires et aux consommateurs de défendre des intérêts financiers et commerciaux. Ils peuvent également être considérés comme des régulateurs et animateurs de la vie en société. L'architecture proposée a amplement présenté les modules et leurs interactions dans ce système. Le chapitre suivant sera consacré à la formalisation des commentaires journalistiques en vue de la fouille d'opinions.

6 -ACQUISITION DE DONNEES JOURNALISTIQUES EN VUE DE LA FOUILLE D'OPINIONS

6.1 - **Introduction**

L'acquisition de données constitue une phase primordiale dans le processus de fouille d'opinions. Cette phase consiste à recueillir des données et à les regrouper de manière synthétique et cohérente sous forme de bases de données en vue d'une analyse d'opinions. Dans la littérature, il existe beaucoup de travaux proposant des méthodes pour extraire des informations contenues dans les pages web. Aujourd'hui, peu de travaux sont orientés dans l'acquisition de données journalistiques disponibles en ligne [23][24]. De plus, les solutions proposées dans ce sens n'intègrent pas la collecte de commentaires. À cela vient s'ajouter l'hétérogénéité des structures de données et la vélocité de la production d'informations qui entravent l'automatisation du web scraping. L'hétérogénéité des structures de données entraine l'usage et le développement de programmes spécifiques tandis que la vélocité de la production de données requiert la collecte en temps réel.

Face à ces contraintes, nous avons proposé *OpinionScraper* [148] qui est un outil de collecte, de fusion et de catégorisation de données journalistiques afin de les stocker au format json. *OpinionScraper* permet d'extraire les informations à partir de pages Web de manière optimale. Il représente aussi ces informations en fonction du modèle défini en vue de la fouille d'opinions. L'intérêt de la mise en place d'un outil de scraping est de constituer une base de données facilement exploitable à partir de commentaires journalistiques afin de répondre à la complexité algorithmique de la fouille d'opinions.

L'objectif de ce chapitre est centré sur la présentation de notre solution. Dans un premier temps, nous allons présenter les deux (02) principales tâches de l'outil, à savoir la collecte de commentaires et leur catégorisation par similarité. En second lieu, nous présenterons l'architecture qui implémente *OpinionScraper* et son utilisation.

6.2 - **Collecte de commentaires journalistiques**

La collecte est une opération qui consiste à interroger des sources d'informations pertinentes afin d'acquérir des données susceptibles de répondre à nos besoins. L'identification des sources (sourcing) est l'ensemble des opérations préalables à la collecte de données. Elle vise à sélectionner des sources (sites web, blogs, forums, etc.) contenant ou susceptibles de contenir de l'information désirée. Les sources que nous avons identifiées correspondent aux sites dédiés à l'information de manière générale et à la presse sénégalaise en ligne en particulier. L'enrichissement et la prolifération de ces sources d'informations ont fait des données

journalistiques un ensemble de données utiles et attrayantes. Une fois les sources connues, nous pouvons aisément définir notre stratégie d'extraction de données. La phase de collecte de commentaires journalistiques est un processus composé de tâches d'extraction, de formatage et de dédoublonnage.

6.2.1 - Extraction de commentaires

Généralement, les sites d'informations sont constitués d'un ensemble de pages web où chacune est identifiée par une URL (Uniform Resource Locator) affichée dans les navigateurs. L'URL dite principale (page d'accueil) contient souvent tous les liens des publications récentes. Chaque lien est redirigé vers une autre page qui contient aussi des données intéressantes.

Notre solution s'adapte à la structure de ces sites. Il s'agit de repérer tous les liens contenus sur la page d'accueil, ensuite de parcourir de manière récursive lien par lien afin de récolter des informations spécifiées. En pratique, il existe plusieurs librairies pour cette opération ; entre autres, nous avons Scraperwiki[17], Scrapy[18], Newspaper[19], RCrawler[20] ou Rvest[21]. Nous avons utilisé Rvest qui est un package R [149], open source, spécialement conçu pour récolter de données sur le web. Rvest propose des fonctionnalités nécessaires à l'identification et l'extraction de données dans une page web. Parmi lesquelles, nous avons utilisé des fonctions comme :

- *read_html()* : la fonction *read_html()* permet d'importer le contenu d'une page web à l'aide de l'URL dite principale.
- *html_nodes()* : cette fonction permet d'extraire des informations d'intérêt à partir d'une page importée par la fonction *read_html()* à l'aide de la syntaxe XPath. XPath permet d'accéder aux informations contenues dans des balises
- *html_text(), html_attrs()* : toutes ces fonctions servent à aspirer et nettoyer les éléments d'intérêt que nous avons isolés à travers la fonction html_nodes().

Ensuite, nous donnons en exemple ce petit script pour étayer l'utilisation des fonctionnalités de Rvest dans le web scraping (voir figure 23).

[17] https://scraperwiki.com/
[18] https://scrapy.org/
[19] https://newspaper.readthedocs.io/en/latest/
[20] https://cran.r-project.org/web/packages/Rcrawler/index.html
[21] https://rpubs.com/ryanthomas/webscraping-with-rvest

```
Fonction : Extraction ;
Def ← fonction (url, noeudParent, noeudFils){
        Liens ← read_html(url) %>% html_nodes(noeudParent) %>%
        html_attr('href');
        Pour tout lien de la liste des Liens Faire
                DonneesBruite ← read_html(lien) %>% html_nodes(noeudFils) %>%
                html_text();
        Fin Pour
        }
Fin
```

Figure 23 : *Fonction d'extraction de données en ligne avec Rvest*

Nous avons paramétré ces fonctions pour automatiser l'extraction de commentaires journalistiques. Ainsi, les commentaires extraits se présentent dans un format non structuré comme le montre la Figure 24. Pour ce jeu de données, nous avons trois articles : articles 1865, 1866 et 1867. Les articles constituent les entités principales. Sur chaque entité, nous avons les commentaires associés. Ce jeu de commentaires collectés révèle quelques fois des incohérences. À cet effet, le formatage devient une nécessité pour obtenir un corpus structuré et cohérent.

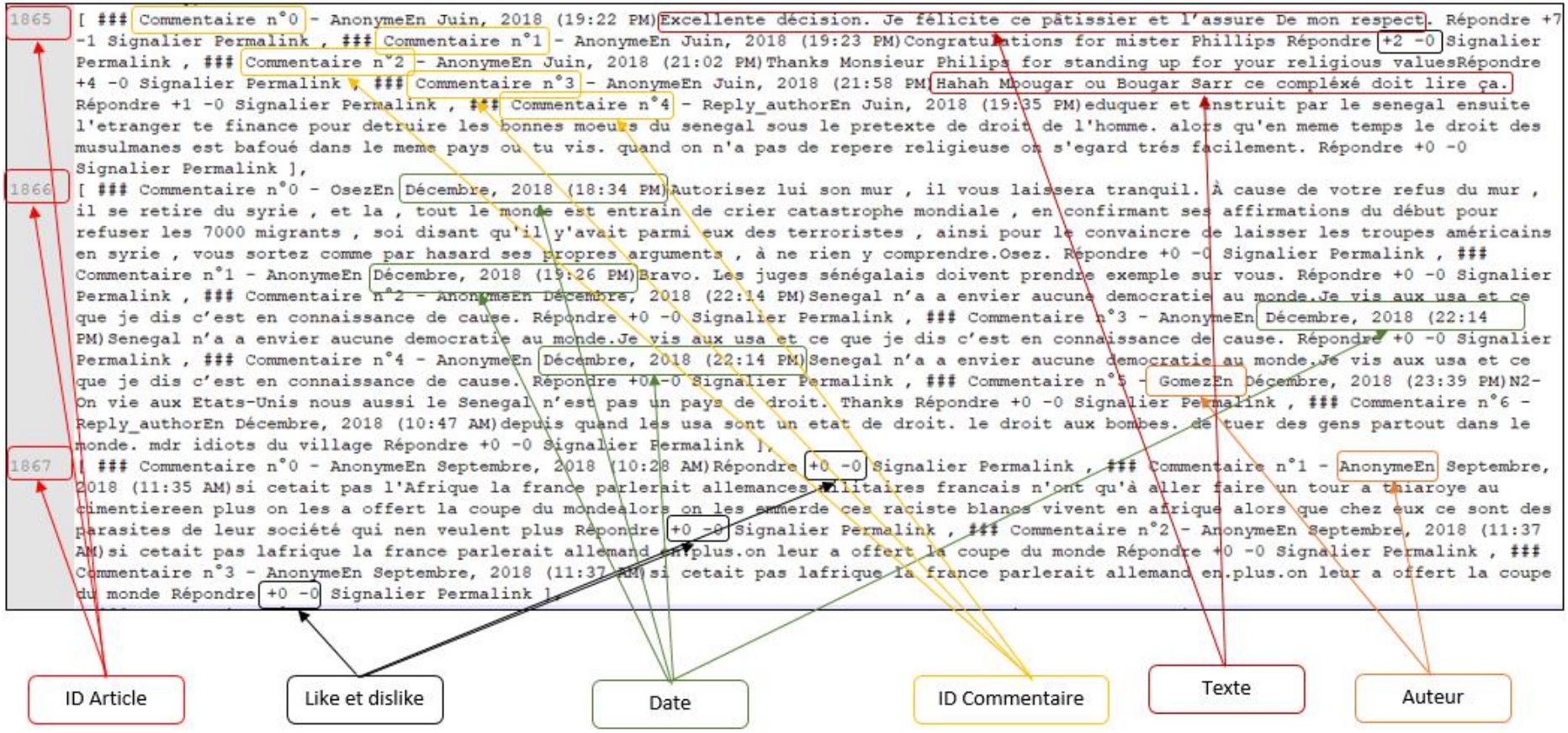


Figure 24 : *Echantillon de commentaires extraits avec Opinion Scraper*

6.2.2 - Formatage de données

Le formatage consiste à structurer les données brutes selon les attributs définis dans notre modèle de commentaire. Il est basé sur les unités textuelles qui composent ces phrases, telles que des règles de catégorie permettant de regrouper les données par classes. Notre objectif est de représenter les commentaires sous format tabulaire. Pour cela, nous allons d'abord définir certains concepts notamment catégories, des motifs et règles de catégories avant de procéder à l'extraction des motifs

6.2.2.1 - Définition des concepts de base

- **Définition 1** : Une **catégorie** est un mot ou groupe de mots permettant de regrouper plus clairement les commentaires journalistiques par rubriques ou classe dans le but de constituer une base de données. La définition de catégories consiste à choisir des variables à l'image des concepts considérés comme unités statistiques de plus haut niveau. Dans notre contexte, les catégories ne sont rien d'autres que les attributs définis précédemment (voir section 5.2.2 du chapitre sur la modélisation de commentaires journalistiques pour la fouille d'opinions).
- **Définition 2** : Un **motif** est un segment de texte identifié à partir d'un commentaire contenant une information exprimée dans le texte.
- **Définition 3** : Une **règle de catégories** est un ensemble de critères permettant d'extraire un motif et de l'affecter dans une catégorie. Autrement dit, elle permet d'identifier un motif, l'extrait et l'associe à une catégorie.

6.2.2.2 - Extraction de motifs

Pour extraire les motifs, nous avons utilisé les expressions régulières comme le montre la Figure 25.

```
#extraction de date
dat=c("\\d\\d(/)\\d\\d(/)\\d\\d\\d\\d à \\d\\d(:)\\d\\d",
      "- (.*?)PM\\)","- (.*?)AM\\)")
#extraction des likes
lik=c("\\+\\d","\\+\\d\\d")
#extraction des dislike
dlik=c("\\-\\d","\\-\\d\\d")
```

Figure 25 : *Exemple d'extraction de motifs*

Les expressions régulières permettent de définir plusieurs critères de recherche en même temps afin d'identifier des motifs (patterns) à l'intérieur de contenu textuel. Sa puissance réside dans le fait qu'il s'applique indépendamment de la structure du document à analyser. En

pratique, nous segmentons les commentaires en mots et groupe de mots. Les règles de catégories sont produites automatiquement ou non à l'aide de techniques de regroupement telles que :

- **Règle de co-occurrences**

La première règle est basée sur les co-occurrences. Une co-occurrence peut être définie comme la présence simultanée de deux ou plusieurs motifs dans la même collection de documents sélectionnés. Les termes considérés comme co-occurrents sont souvent liés par des relations formelles ou syntaxiques. Ils sont aussi liés par synonymie, antonymie ou contenance (hyperonymie ou métonymie). La méthode de co-occurrence permet de créer un lexique par répétition de formes présentes dans un texte. À côté de cette technique, nous avons défini d'autres critères basés sur les règles de dérivation des racines de concepts.

- **Règle 2 : Règle basée sur la dérivation de racines**

La dérivation de racines est une règle de catégorie basée sur la partie invariable d'un concept. C'est une méthode qui s'appuie sur la détection de chaines constituées de morceaux existants plusieurs fois dans le même texte. On symbolise les morceaux par des lettres. On fixe une fréquence minimale d'apparition dans le texte. Nous utilisons cette règle de segments répétés pour extraire l'identifiant des commentaires qui commence toujours par « Commentaire ».

- **Règle 3 : Règle manuelle**

D'autres règles sont créées manuellement sur la base de notre compréhension de données et du contexte. Cette méthode s'appuie sur une ressource externe qui consigne les mots et expressions figées voire semi-figées susceptibles d'être rencontrés dans un texte du domaine.

Pour réaliser cette tâche, nous avons proposé l'algorithme à travers la fonction suivante (voir Figure 26) pour formater ces données. La fonction de formatage prend en entrée les commentaires bruits, les variables de règles. Chaque règle est rattachée à une seule catégorie afin que chaque motif extrait puisse être affecté à une catégorie. Si une correspondance est détectée entre le descripteur et le motif alors ce motif est attribué à cette catégorie. Cette fonction fusionne ou agrège des données dans un format unique en éliminant certains segments qui ne respectent pas les règles établies.

```
Def ← fonction (DonneesBruite, idCom, auteurCom, dateCom, entiteCom, likeCom, dislikeCom, texteCom){
        Commentaires ← unlist (strsplit (DonneesBruite, "###") ;
        Pour tout motif de la liste des Commentaires  Faire
                Identifiant ← str_extract (commentaire, idCom) ;
                Entite ← str_extract (commentaire, entiteCom) ;
                Date_publication ← str_extract (commentaire, dateCom) ;
                Auteur ← str_extract (commentaire, auteurCom) ;
                Like ← str_extract (commentaire, likeCom) ;
                Dislike ← str_extract (commentaire, dislikeCom) ;
                Texte ← str_extract (commentaire, texteCom) ;
                DfCommentaire ← data.frame (Identifiant, Entite, Date_publication,
                Auteur, Like, Dislike, Texte)
        Fin Pour
        }
Fin
```

Figure 26 : *Fonction de formatage*

Le formatage a pour rôle de séparer des différentes parties d'un commentaire. Lors de la collecte de commentaires, il n'est pas rare de voir apparaître des doublons. Le dédoublonnage permet de peaufiner la représentation de données.

6.2.3 - Dédoublonnage

Un doublon se caractérise par la présence répétée des mêmes informations dans une collection. En d'autres termes, c'est un ensemble de données souvent erronées qui se présentent inutilement de manière répétitive dans un corpus. De plus, ces types de données échappent facilement aux détections superficielles.

La présence de doublons risque d'altérer la réalité des résultats et affecte directement les décisions prises par les acteurs. C'est une menace pour la qualité et la fiabilité des données. Son impact peut influencer sur l'intégrité des données. Au regard de ces dérives, le dédoublonnage est devenu une technique importante pour affiner les résultats. La technique du dédoublonnage a pour objectif d'identifier et de supprimer les doublons présents au sein d'une collection.

Pour détecter les doublons, nous nous basons sur le modèle de commentaire défini, à savoir l'entité E, le contenu T et l'auteur A. Considérons deux commentaires étant des doublons

s'ils ont les mêmes textes et les mêmes auteurs et que ces deux portent sur la même entité. Formellement, on peut détecter des doublons comme suite (voir Figure 27) :

Soit $\mathcal{M}$, famille de commentaires noté $\mathcal{M} = \{ C_1, C_2, \dots, C_n \}$ caractérisés par p descripteurs (E, T, A) ;

$$Si\ C_1(p) == C_2(p)\ alors$$
$$C_1\ et\ C_2\ sont\ des\ doublons$$

Figure 27 : *Règle de détection des doublons*

Les données collectées et bien formatées peuvent faire l'objet de catégorisation afin de les regrouper par thème.

6.3 - Catégorisation de commentaires par similarité

Rappelons que la catégorisation consiste à regrouper les commentaires journalistiques par évènements à travers l'étude de similarité. Notre approche de similarité est basé sur le Cosinus qui est une métrique très populaire en fouille de textes pour mesurer la similarité entre des documents [38][39]. Nous considérons que chaque document est représenté par un vecteur de termes. La formule est la suivante (voir Figure 28) :

Soient d_1 et d_2 deux documents correspondant à des vecteurs de termes :

$$\cos(d_1, d_2) = \frac{\langle d_1 | d_2 \rangle}{\|d_1\| . \|d_2\|}$$
$$Dist(d_1, d_2) = 1 - \cos(d_1, d_2)$$

Figure 28 : *Formule de Cosinus*

Plus la distance entre ces deux documents est proche de 0, plus ces documents sont sémantiquement similaires. Autrement dit, le calcul de similarité avec le Cosinus nous permet de déduire que deux documents sont proches s'ils ont de nombreux termes en commun. La classification de commentaires journalistiques nécessite le prétraitement des documents, la sélection de termes candidats et la proposition de l'algorithme.

6.3.1 - Prétraitement

Le prétraitement consiste à transformer le document du format texte en format matriciel pour en appliquer les méthodes d'analyse de données. Cette étape est composée de l'étiquetage grammatical, la suppression de termes vides et la représentation de documents dans une matrice.

L'étiquetage grammatical consiste à associer aux mots du texte des informations grammaticales, telles que leur nature (nom, adjectif, verbe, article, etc.) et éventuellement leur forme canonique. Pour étiqueter notre base d'analyse, nous avons utilisé TreeTagger [25] qui est l'un des étiqueteurs les plus couramment utilisés pour réaliser cette tâche [150]. Il repose sur une méthode d'étiquetage qui utilise les arbres de décision pour déterminer ces informations. TreeTagger supporte plusieurs langues (français, anglais, allemand, etc.) et est adaptable à d'autres langues si un lexique et/ou un corpus d'entrainement annoté manuellement sont disponibles pour ces dernières. Le caractère open source et la performance ont favorisé le choix de TreeTagger. Les résultats de l'étiquetage servent de support à des tâches plus complexes ou de plus haut niveau linguistique : l'extraction terminologique, la recherche d'informations , la fouille d'opinions, etc.[151].

À la suite de l'étiquetage grammatical, nous avons cherché à éliminer certains éléments qui sont considérés comme des termes vides. Il s'agit des articles, des prépositions, des déterminants qui ont souvent un sens moins précis. La suppression de ces éléments nous conduit vers un corpus dont les dimensions sont réduites.

Enfin, nous associons une pondération à chaque terme retenu de la liste pour mesurer le degré d'importance de ces termes dans les documents. Pour cela, nous utilisons la technique TF*IDF qui est issue de l'approche statistique basée sur la pondération de termes afin de déterminer le degré de pertinence d'un terme dans le document [31]. L'objectif de l'approche statistique est de trouver les termes candidats dont le comportement dans le document varie positivement comparé à leur comportement global dans la collection [111]. Ainsi, plus le score TF-IDF d'un terme est élevé, plus celui-ci est important dans le document analysé.

Le résultat obtenu à l'issu du prétraitement est une matrice contenant les documents en ligne et les termes en colonne (ou inverse). Chaque cellule $\boldsymbol{C_{ij}}$ de cette matrice est le poids du terme $\boldsymbol{j}$ dans le document $\boldsymbol{i}$.

6.3.2 - Extraction de termes candidats

Pour l'extraction de termes candidats, plusieurs techniques peuvent être utilisées. Dans cette section, nous définissons deux critères afin de considérer un terme comme étant candidat pour indexer le document. Ces deux critères sont d'ordre statistique et morphosyntaxique.

- Le critère statistique est basé sur le résultat de pondération. À cet effet, la valeur de discrimination d'un terme à travers son poids tf-idf est un facteur déterminant. Nous supposons que plus un terme candidat a une valeur élevée, plus il est discriminant. Ainsi, nous avons sélectionné les termes candidats les plus représentatifs, c'est-à-dire, ayant un poids supérieur à un seuil déterminé.
- Le critère morphosyntaxique est basé sur le résultat de TreeTagger pour qualifier un terme de candidat. Parmi les résultats fournis par cet outil, nous avons sélectionné uniquement les noms et syntagmes nominaux du réseau, considérés comme les termes candidats. L'idée est que généralement les termes sont représentés par des noms communs et groupes nominaux. Par conséquent, cibler les mots et groupes de mots ayant ces structures syntaxiques permet d'extraire les termes essentiels du corpus mais aussi de limiter le bruit que peuvent engendrer les non-termes.

Par ailleurs, les termes constitués exclusivement de chiffres sont élagués. Ceux ayant moins de quatre caractères et ceux contenant des chiffres ou des caractères non alphanumériques sont isolés. Une fois les termes candidats extraits et filtrés, il est ensuite nécessaire de les organiser en classe et de relier ces derniers entre eux.

6.3.3 - Algorithme de calcul de similarité

Cette approche est implémentée à travers l'algorithme suivant (voir Figure 29) :

Algorithme : Classification par Similarité ;
Entrée :
N ← Nombre Total de Documents à classer ;
Seuil ← 0,5 ;
Sortie : Classes ;
Début
Pour i ← 1 à N par pas de 1 **Faire**
Pour j ← N à 1 par pas de -1 **Faire**
$Dist(d_i, d_j) = 1 - \cos(d_i, d_j)$
Si $Dist(d_i, d_j) \geq Seuil$ **Alors**
d_i et d_j appartiennent à la même classe ;
Fin Si
Fin Pour
Fin Pour
Fin

Figure 29 : *Algorithme de calcul de similarité*

Concernant la recherche de documents similaires, nous nous sommes basés sur la mesure du cosinus qui est largement utilisée en RI. Il serait intéressant d'investiguer la combinaison de cette dernière avec des ressources de connaissances du domaine afin d'améliorer la méthode de calcul de similarité basée seulement sur les mots communs. Quant à l'influence du seuil, nous avons choisi un seuil minimal fixé à 0,5. Plus ce seuil est élevé, plus la classification est précise. L'inconvénient de cette élévation est que peu de documents correspondent.

Notre réelle motivation est de regrouper les textes similaires, c'est à dire thématiquement proches. L'intérêt d'une telle démarche est d'organiser les documents de façon à pouvoir effectuer, par la suite, une recherche ou une extraction d'informations efficace.

6.4 - **Implémentation d'OpinionScraper**

6.4.1 - Présentation de l'architecture

OpinionScraper implémente une nouvelle approche de collecte, de fusion et de catégorisation de données journalistiques pour la fouille d'opinions. La Figure 30 présente l'architecture d'OpinionScraper.

Nous avons deux modules dont chacun est autonome dans son fonctionnement : il s'agit de la collecte de données et la catégorisation par similarité. Le premier module se charge

d'interroger les sources web identifiées pour en extraire des données, les formater et les nettoyer afin de les représenter sous le format du modèle défini. Quant à la catégorisation, elle se fonde sur la similarité de commentaires à travers le titre de l'article qui constitue l'entité, le texte du commentaire et les dates de publications de ces commentaires afin de regrouper ces données journalistiques par évènement car les informations journalistiques sont souvent organisées par évènements. Le choix portant sur le titre de l'article est le fait qu'il est constitué de mots clefs qui donnent une idée générale du sujet traité.

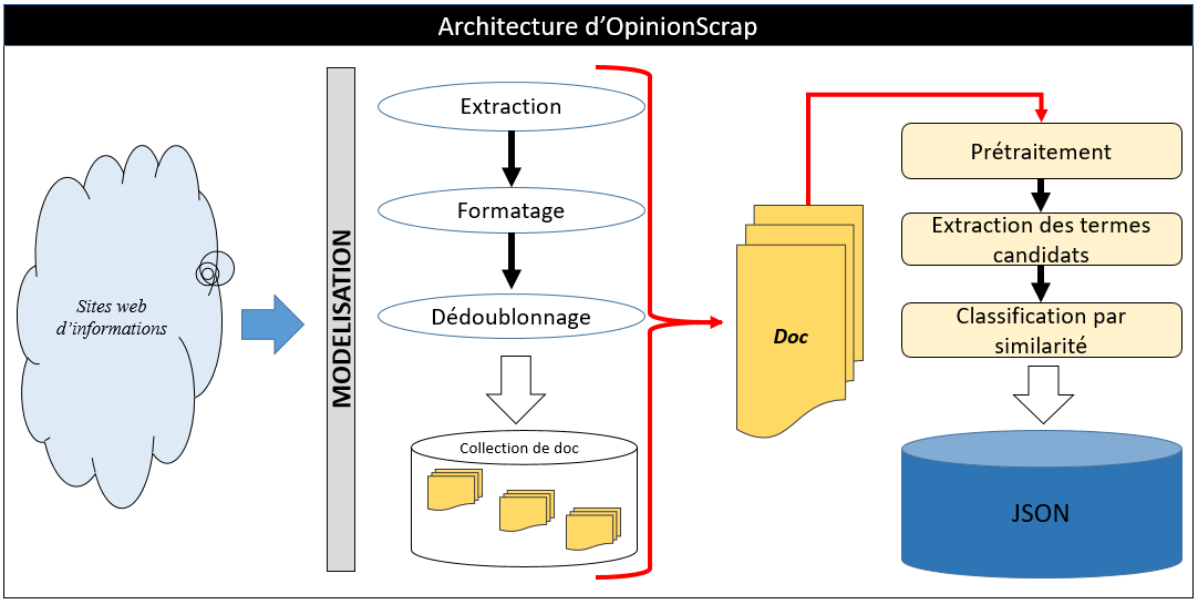


Figure 30 : *Architecture d'acquisition des commentaires journalistiques*

Malgré l'autonomie dans les processus de fonctionnement, ces modules entretiennent une forte relation. Le module catégorisation utilise les résultats issus de la phase de collecte pour son fonctionnement.

En résumé, nous pouvons dire que l'intérêt de notre outil est de collecter des données non structurées disponibles dans la presse sénégalaise en ligne dans le but de les transformer afin d'obtenir des données qui peuvent être traitées successivement. À cet effet, le stockage est une représentation d'informations du monde réel sur un ou plusieurs supports accessibles qui peuvent être interrogés et mis à jour par une communauté d'utilisateurs. Dans notre cas, nous avons choisi le format JSON qui est adapté à nos besoins. JSON permet de stocker des données de différents types à l'image de commentaires journalistiques de manière structurée. Sa structure en arborescence et sa syntaxe simple lui permettent de rester léger et efficace pour l'échange de données. Il facilite aussi l'interopérabilité entre les applications. Cette étude est motivée par la nécessité de constituer un corpus faisant office de base de données afin de

permettre des analyses qualitatives et/ou quantitatives. Dans la section suivante, nous allons expérimenter notre outil sur des sites d'informations sénégalais.

6.4.2 - Application d'OpinionScraper

OpinionScraper peut interroger plusieurs sites d'informations de manière générale et plus spécifiquement la presse sénégalaise en ligne en vue de la fouille d'opinions [148]. Dans l'expérimentation, nous avons mis en exergue les tâches décrites sur la Figure 31.

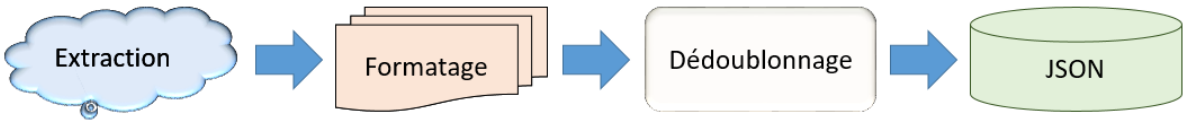


Figure 31 : *Processus d'OpinionScraper*

La figure 31 décrit le fonctionnement ou l'enchainement des étapes successives permettant de collecter les données désirées, de les fusionner et nettoyer afin de les stocker. Dans ce qui suit, nous allons présenter les résultats du test de notre outil sur des cas réels afin d'évaluer son comportement.

6.4.2.1 - Test de l'outil

Nous avons testé notre solution sur cent trente-deux (132) sites d'informations sénégalais qui sont préconfigurés. Les données collectées se présentent sous la forme d'un bloc construit pour faciliter la catégorisation. Nous présentons ici un d'extrait de commentaires bruts collectés à la Figure 32.

```
[### Commentaire n°0 - Djibril Camara 23/07/2019 à 23:52 Yaw mom seytanee gua yak pape
alez niang Répondre,### Commentaire n°1 - Diambar 23/07/2019 à 23:55 Ce gars la nous
amerdes.et pour un lutteur qui lutte juste pour 5mns et empoche des centaines de
millions qu'allez tu dire. Trop de critique pour rien. Ya wara wah tahoul ngay wah wahou
dof Répondre,### Commentaire n°2 - Jules 23/07/2019 à 23:56 Un lutteur pour 100millions
a moins de 5mns wahal ci lolou ba pareRépondre,### Commentaire n°3 - René Diagne
24/07/2019 à 00:18 Vraiment la malhonnêteté intellectuelle est bien de mise chez
certains. Si le président ne l avais pas fait somme toutes allez savoir ce que vous
alliez nous servir.Quant aux stades dignes de ce nom vous en tant qu acteur de média qu
avez vous fait pour alerter ,conscientiser les populations pour un bon entretien..car
tout réside chez nous dans l entretien de nos ouvrages publics…..Répondre,###
Commentaire n°4 - Satar 24/07/2019 à 00:22 En tout cas cela n'encourage pas à faire des
études universitaires pour être médecin ou professeur ou architecte etc. Mieux vaut se
concentrer sur la lutte ou le foot ball.... Hé VOILA COMME ON CRETINISE SON PEUPLE. ON
NE DEVELOPPE PAS UN PAYS OU UNE NATION PAR LE FOOT BALL ET LA LUTTERépondre,###
Commentaire n°5 - Lemou 24/07/2019 à 00:39 N'diaye Doss, cet escroc des passe-ports…Dans
la vir chacun doit tenfre vers ce qu'il peut faire le mieux.Il n'est pas donné à tout le
monde d'être médecin ou prof..Comme aussi tout le monde ne peit être un footballeur ou
lutteur…L'essentiel c'est de réussir sa vie en utilisant ses capacités
naturelles…Répondre,### Commentaire n°6 - Cheikh Cheikhouna 24/07/2019 à 01:49 Pourquoi
vous êtes méchant kou déf dara am dara Répondre,### Commentaire n°7 - Boucar Sene
```


Figure 32 : *Extrait de commentaires bruts collectés*

articles	titres	commentaires	idCommentaires	auteurs	dateComs	likes	dislikes
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	tous des voleurs.....vivement du sang neuf dans la sper...	Commentaire n°0	Anonyme	Mai, 2018 (18:36 PM)	14	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Pas surprenant de la part d'un traitre	Commentaire n°1	Anonyme	Mai, 2018 (18:37 PM)	83	-76
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	sunugaal::sunugaalsunugaal::sunugaalsunugaal::sunu...	Commentaire n°2	[ema	Mai, 2018 (21:04 PM)	0	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	si j'ai bien compris, président wade dit que président m...	Commentaire n°14	Anonyme	Mai, 2018 (11:50 AM)	0	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	C'est du vous dans vous. Si vous aviez géré le pays dans ...	Commentaire n°15	Anonyme	Mai, 2018 (18:37 PM)	5	-5
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	on s'en mords les doigts et nos enfant le feront aussi	Commentaire n°16	Anonyme	Mai, 2018 (19:12 PM)	1	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Thiey Macky même pas honte, j'espère qu'il nous dira d'...	Commentaire n°17	Anonyme	Mai, 2018 (18:38 PM)	19	-10
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	-installation camera de surveillance dans vos apartemen...	Commentaire n°18	Anonyme	Mai, 2018 (19:17 PM)	0	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	7 milliards comme les 7 milliards de taiwan. ça rime bien	Commentaire n°21	Anonyme	Mai, 2018 (10:02 AM)	6	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	laye wade, 0 sur 20	Commentaire n°23	Anonyme	Mai, 2018 (18:38 PM)	14	-23
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	laye wade moma gueuneul fouf matioudo sall	Commentaire n°24	Anonyme	Mai, 2018 (18:42 PM)	15	-10
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Comedien de salon.Bientot il dira que les juges lui ont d...	Commentaire n°26	Anonyme	Mai, 2018 (18:41 PM)	15	-73
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	batard. fils de chien, comment peux-tu parler ainsi.	Commentaire n°27	Anonyme	Mai, 2018 (18:56 PM)	63	-56
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	tu est vraiment ingrat plus qu'un ingrat tu est vraiment ...	Commentaire n°28	Anonyme	Mai, 2018 (23:59 PM)	0	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Wade vient a la rescousse de Idy. Il veut deplacer le deba...	Commentaire n°29	Anonyme	Mai, 2018 (18:43 PM)	5	-31
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Saisir la maison de Maitre Wade au point E.Je n'y crois p...	Commentaire n°30	Anonyme	Mai, 2018 (18:44 PM)	79	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	kii day lèep.	Commentaire n°31	Anonyme	Mai, 2018 (18:49 PM)	1	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	mr. relisez l"article. macky pretend que wade doit 550 mil...	Commentaire n°32	Anonyme	Mai, 2018 (19:07 PM)	0	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Triste. Le vieux manipulateur... Va te reposer , its high tim...	Commentaire n°34	Anonyme	Mai, 2018 (18:45 PM)	0	-49
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Wade dans ses œuvres Manipulateur de circonstances t...	Commentaire n°35	Anonyme	Mai, 2018 (18:50 PM)	1	-67
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	cette maison quand il y habitait ton mentor louait à der...	Commentaire n°36	Anonyme	Mai, 2018 (22:13 PM)	66	-1
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	baises ta maman toi, sale repondeur automatique	Commentaire n°38	Anonyme	Mai, 2018 (23:52 PM)	0	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Toujours Karim.Wade est devenu complètement fou	Commentaire n°39	Anonyme	Mai, 2018 (18:52 PM)	0	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	C DEMANDER A MACKY SALL OU il veut ALLER / TOUJOUR...	Commentaire n°40	Anonyme	Mai, 2018 (18:53 PM)	50	0
article n° 5368	Wade en colère : 'Macky Sall a ordonné la saisie de ma ...	Rien que de la com pour entre dans la danse de l'affaire...	Commentaire n°41	Sunugaal	Mai, 2018 (18:56 PM)	0	-65

Figure 33 : *Extrait de données formatées et nettoyées*

Le formatage permet d'améliorer la représentation du résultat obtenu lors de l'extraction. En plus, la technique de nettoyage à travers le dédoublonnage permet de peaufiner le résultat obtenu. Le résultat issu de cette fusion de données peut servir d'analyses quantitatives comme le montre la Figure 33.

Le but du stockage de données est d'obtenir un schéma de l'organisation de données stable et invariant permettant de construire une solution physique, concrètement une base de données. Notre proposition représente aisément les données collectées et fusionnées dans le format JSON. En guise d'illustration, nous donnons un extrait de commentaires formatés en JSON à la Figure 34.


```
{"titres":"Baaba Maal : ' la musique sénégalaise à l'échelle international",
 "commentaires":" Toujours notre fierte nationale et continentale",
 "idCommentaires":"article n° 328#Commentaire n°0",
 "auteurs":"Anonyme",
 "dateComs":"Tue Jul 23 21:07:27 UTC 2019  ",
 "likes":"+0",
 "dislikes":"-0"
},
{"titreArticle":"Débat : Sadio Mané a t-il perdu le Ballon d'Or.",
 "commentaires":" même si on gagnait la CAN il ne merite pas le ballon d or mondial. Il n as pas encore
 atteint le niveau pour l être ",
 "idCommentaires":"article n° 330#Commentaire n°3",
 "auteurs":"Anonyme",
 "dateComs":"Sat Jul 20 00:07:13 UTC 2019  ",
 "likes":"+12",
 "dislikes":"-1"
},
{"titreArticle":"Seydi Gassama : 'Le tweet de Macky Sall fait la fierté de l'Afrique",
 "commentaires":" c totament vrai. le peuple africain n'est pas une merde car elle reste hereux dans sa
 pauvreté mais ce sont nos chefs d'etat comme macky.. et d'autres qui sont là depuis 30 ans qui sont de la
 merde. depuis l'independance à nos jour il ya beaucoup de chose qui ont bouger car ce ne sont pas les
 aeroports, autoroutes... qui changera la vie quotidien et sa nempeche pas les jeunes de quitter leur pays
 pour un monde meilleur ",
 "idCommentaires":"article n° 2069#Commentaire n°30",
 "auteurs":"Anonyme",
 "dateComs":" Janvier  2018 (18:26 PM)",
 "likes":"+30",
 "dislikes":"-0"
},
```


Figure 34 : *Extrait de commentaires formatés*

6.4.2.2 - Optimisation

Pour optimiser le temps et extraire uniquement les informations d'intérêt de manière précise, nous avons utilisé les sélecteurs CSS. En développement web, le CSS est utilisé d'une part pour les mises en forme de pages et de textes ; d'autre part pour regrouper les données par classe. La tâche d'extraction fournit un texte comme résultat avec moins de bruits et le présente sous la forme d'un bloc construit pour le formatage. Le fait de définir une représentation formelle de commentaires dès le départ garantit qu'à chaque nouvelle extraction, nous obtenons exactement les mêmes résultats. De cette manière, l'extraction et l'attribution de motifs extraits aux catégories sont plus précises et plus répétables.

6.5 - **Conclusion**

En définitive, la constitution de corpus à partir de données disponibles sur la presse en ligne est un travail de construction de modèle adapté à nos besoins. À cet effet, nous avons proposé une méthode de scraping qui extrait les bonnes informations à partir de pages Web, les agrège en supprimant les doublons, les catégorise et les stocke dans un fichier au format json. La solution mise en place effectue un scraping continu en un intervalle de temps réduit pour suivre l'évolution des commentaires.

L'originalité de notre travail réside dans le suivi de l'évolution des commentaires sur les sites web dédiés à l'information, le dédoublonnage et la catégorisation par similarité. Cette approche innovante résout la complexité de données d'une part et facilite l'acquisition de données d'autre part. En plus, elle a l'avantage d'éviter la complexité de la fouille d'opinions basée sur les aspects à partir de la puissance de représentation des commentaires. En traitement automatique de textes, si les données sont moins complexes, alors l'analyse devient plus pertinente. En guise de perspectives, nous comptons étendre la solution aux autres sources d'informations notamment les réseaux sociaux (Facebook, WhatsApp). Dans le chapitre suivant, nous allons parler de la fouille d'opinions proprement dite.

7 -VERS UN LEXIQUE (BILINGUE) FRANÇAIS-WOLOF POUR L'ETIQUETAGE D'OPINIONS

7.1 - **Introduction**

L'étiquetage d'opinions est une étape cruciale dans le processus d'analyse d'opinions. Elle consiste à identifier dans un document le vocabulaire porteur d'indices d'opinions. Autrement dit, il s'agit de distinguer des textes qui relatent des faits (description objective) de ceux qui présentent des opinions (description subjective) au sein d'un ensemble de documents. Cela revient à sélectionner dans le texte les termes susceptibles de contenir de la subjectivité et de leur affecter une polarité (positive ou négative). Pour qualifier un terme (mot ou groupe de mots) d'indice d'opinion de manière automatique, il est nécessaire d'utiliser une ressource linguistique telle qu'un corpus d'entrainement annoté, un dictionnaire ou un lexique d'opinions en fonction de l'approche adoptée. Ces ressources de fouille d'opinions sont très rares pour les langues nationales du Sénégal comme nous l'avons souligné dans l'état de l'art.

En raison de ces limites, la construction d'un lexique d'opinions pour l'analyse de commentaires issus de la presse en ligne devient une impérieuse nécessité. C'est dans ce contexte que nous avons mis en place ***SenOpinion*** [152]. ***SenOpinion*** est un lexique d'opinions composé de termes en français et wolof, destiné à étiqueter les commentaires écrits en langage urbain sénégalais. Ce langage combine les langues étrangères comme le français (qui est la langue majoritairement utilisée et les langues nationales, notamment le wolof). La particularité de la langue wolof dans notre étude réside dans sa forte présence dans les communications [153][154][101]. Nous trouvons cette langue dans les outils tels que Wikipédia, Windows et Google. Elle s'impose de plus en plus dans les commentaires en ligne des sénégalais.

Une fois que les termes sont étiquetés, nous pouvons déterminer des statistiques avec des données à travers les différents niveaux de texte, tels que le niveau du corpus (tendance globale), le niveau d'articles et le niveau de commentaires. À cet effet, nous proposons un modèle mathématique de calcule d'opinions d'un commentaire basé sur les termes étiquetés, les likes (j'aime) et dislikes (je n'aime pas).

Ce chapitre a un double objectif : d'une part, de mettre en place un lexique pour l'étiquetage d'opinions ; d'autre part, de proposer un modèle mathématique pour la classification non supervisée des commentaires selon les positions favorables, défavorables ou neutres. Il est structuré en quatre (04) sections hormis l'introduction et la conclusion : d'abord, nous allons décrire le processus de construction de SenOpinion ; ensuite, nous proposerons une méthode d'utilisation du lexique pour l'étiquetage et un modèle de calcul d'opinions. Enfin, nous allons l'expérimenter sur des données réelles.

7.2 - **Construction de SenOpinion**

La construction du lexique SenOpinion a suivi un processus composé de deux étapes à savoir la collecte et l'annotation de données comme le montre la Figure 35.

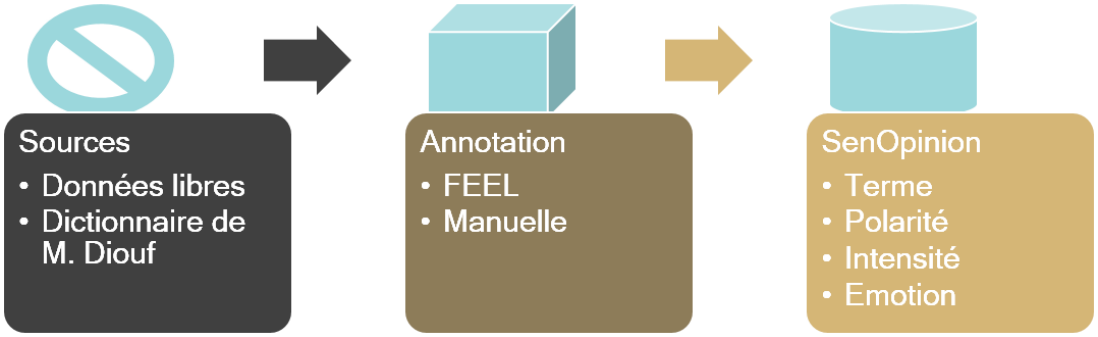


Figure 35 : *Processus de conception de SenOpinion*

7.2.1 - Collecte de données

Nous avons collecté des données à partir de plusieurs sources notamment le dictionnaire bilingue de Jean Léopold Diouf [155] et d'autres documentations en ligne comme les sites salysenegal.net[22], Afroweb.chez.com[23]. Ces sources proposent des données libres composées de mots et d'expressions en français et Wolof dans un langage contemporain. Cet ensemble de termes est étendu en utilisant d'une part les relations de synonymie et d'antonymie [156][157], les définitions [158] et d'autre part en s'appuyant sur plusieurs indices notamment les conjonctions (et, mais) [103], la co-occurrence entre mots [159] et la proximité de contextes d'évaluation [160]. Le résultat des données collectées est présenté dans un tableau en deux colonnes. La première colonne correspond aux termes en français et la seconde désigne leurs traductions en wolof. Ces données collectées et présentées sous format tabulaire ont fait l'objet d'annotation d'opinions.

7.2.2 - Annotation

L'annotation est une tâche plus spécifiquement linguistique. Elle consiste à associer des étiquettes aux termes sélectionnés [181]. Dans un premier temps, nous avons utilisé l'annotation semi-automatique à l'aide de FEEL qui est un lexique open source en français [99]. Pour cela, nous cherchons d'abord les polarités des termes français dans FEEL pour ensuite les affecter directement à leurs correspondants en wolof. Cette tâche a été complétée par un travail manuel.

[22] http://www.salysenegal.net/dicowolof/dicowolof.htm
[23] http://afroweb.chez.com/frm_wofr.htm

Celui-ci consiste d'une part à vérifier, à corriger si c'est nécessaire et à valider l'annotation proposée par FEEL ; d'autre part à proposer des étiquettes pour les termes non identifiés par FEEL. Notre lexique se présente comme le montre la figure 36.

word	polarity	joy	fear	sadness	anger	surprise	disgust	intensity
daan	positive	1	0	0	0	0	0	0.5
daanu	negative	0	1	1	0	0	0	-0.5
daara	positive	0	0	0	0	0	0	0.5
dab	negative	0	0	1	0	0	0	-0.5
dadd sa banneex	positive	1	0	0	0	0	0	0.5
dagg	negative	0	1	0	1	0	0	-0.5
dajaloo	positive	0	0	0	0	0	0	0.5
daje	positive	0	0	0	0	1	0	0.5
dall	positive	0	0	0	0	0	0	0.5
damm	negative	0	1	0	0	1	0	-0.5
damp	positive	1	0	0	0	0	0	0.5
daw	positive	0	0	0	0	0	0	0.5
dee	negative	0	0	1	0	0	1	-0.5
dee na	negative	0	0	1	0	0	1	-0.5
def	positive	0	0	0	0	0	0	0.5
def xew	positive	0	0	0	0	1	0	0.5
defa lakk	negative	0	1	0	1	0	0	-0.5
defar	positive	0	0	0	0	0	0	0.5
deg	negative	0	1	0	0	0	1	-0.5
dëgër	negative	0	1	1	1	0	1	-0.5
degg	positive	0	0	0	0	0	0	0.5
dëgg	positive	0	0	0	0	0	0	0.5
deggal	positive	0	1	0	0	0	0	0.5

Figure 36 : *Extrait de SenOpinion*

- Terme (*Word*) : cet attribut représente les mots et expressions en français ou wolof.
- Polarité (*Polarity*) : L'attribut polarité permet d'indiquer l'orientation de termes (positive ou négative). Cet attribut de sentiments permet de définir si l'opinion est favorable ou défavorable.
- Intensité (*Intensity*) : L'attribut intensité mesure le degré du sentiment exprimé en se basant sur des adverbes tels que très, particulièrement, beaucoup. Il fournit des scores d'opinions dans une fourchette entre -1 et +1 (voir Tableau 9).
- Émotion (*Emotion*) : Nous avons aussi l'attribut émotion qui a pour rôle d'indiquer l'état d'esprit du commentateur. Il se manifeste par l'utilisation de vocabulaire à connotation

de joie (*joy*), peur (*fear*), tristesse (*sadness*), colère (*anger*), surprise (*surprise*) et dégout (*disgut*).

Tableau 9 : Proposition de pondération des termes

Termes	Intensité
Bon	0.5
Mauvais	-0.5
Très bien	1
Meilleur	1
Pire	-1

La mise en place d'un lexique en fouille d'opinions a pour objectif d'alimenter les systèmes de fouille d'opinions notamment les moteurs de fouille d'opinions [162][163]. Au total, nous avons pu étiqueter 697 termes wolofs et 100 nouveaux en français contemporain (langage urbain) que nous avons fusionnés avec FEEL pour donner SenOpinion. Dans la section suivante, nous proposons une méthode d'utilisation de ce lexique.

7.3 - Utilisation de SenOpinion pour l'étiquetage d'opinion

Ici, nous nous plaçons dans un contexte pratique. Pour appliquer notre solution à des données brutes, nous procédons d'abord à l'étiquetage morphosyntaxique et après à l'étude de polarité.

7.3.1 - Étiquetage morphosyntaxique

Les termes susceptibles de contenir des indices d'opinions, sont principalement des noms, adjectifs, verbes et adverbes. À cet effet, l'étiquetage morphosyntaxique (*POST, Part Of Speech Tagging*) aide à identifier les structures grammaticales de chaque token dans une phrase. C'est une étape qui peut être considérée comme préliminaire à tout traitement linguistique plus poussé sur un texte, notamment l'analyse syntaxique. Elle consiste à affecter des étiquettes morphosyntaxiques propres à chaque mot d'une phrase d'un texte (catégorie grammaticale, informations morphologiques comme le genre, le nombre...) [150]. Le Tableau 10 montre et définit les concepts de base en étiquetage morphosyntaxique [164].

Tableau 10 : Concepts de base de l'étiquetage morphosyntaxique

Concepts	**Définitions**

Token	Encore appelé Chuck ou lemme est un constituant syntaxique de la phrase.
Tag	Il indique un descripteur d'éléments de la phrase. Un tag peut être un nom, un verbe, un adverbe, un adjectif, une interjection, un pronom, une conjonction et leurs sous-catégories
Tagging	C'est l'assignation automatique de descripteurs à des tokens donnés, c'est-à-dire, affecter une "étiquette" ("tag", ou catégorie) à chaque "mot" d'un texte.
Tag set	C'est un ensemble de tags à partir duquel le Tagger (étiqueteur) choisit un tag auquel il doit rattacher un mot (token).

La principale difficulté de l'étiquetage morphosyntaxique vient du fait que les mots de la langue sont ambigus. Pour effectuer une phase de désambiguïsation afin de sélectionner la séquence correcte, il est nécessaire de considérer tout d'abord le token de manière hors contexte, ensuite par rapport au contexte proche et, le cas échéant, par rapport à un contexte plus lointain. Pour étiqueter nos corpus, nous avons utilisé TreeTagger [25] qui est un outil open source permettant l'étiquetage de textes en français. Cet outil est un analyseur syntaxique robuste qui supporte le traitement de corpus français et anglais. Nous avons aussi utilisé les fonctions d'extraction des *n-grammes* avec *n* variant entre 1 et 3 mots pour compléter cette segmentation. Enfin, nous avons sélectionné des termes candidats qui sont composés de lemmes suivants : noms, verbes, adjectifs, adverbes et de groupes nominaux dans le but de réduire la dimension des données. Le résultat obtenu est une liste de termes sélectionnés pour l'étude de polarité.

7.3.2 - Étude de polarité

L'étude de polarité consiste à trouver les orientations positives ou négatives des termes candidats. Dans notre contexte, nous cherchons à déterminer la similitude entre les mots issus de notre base d'analyse (liste de termes à l'entrée) et ceux de notre lexique afin de déterminer la polarité des mots de la base d'analyse. Pour trouver des correspondances entre des paires de listes sans identificateurs uniques, nous préférons comparer des chaînes de lettres afin d'associer des étiquettes à chaque terme à travers la procédure suivante :

Le lexique est implémenté par un tableau $T\ [0, \dots, m-1]$

On suppose que :

- Chaque terme a une valeur tirée d'un univers $U = \{0,1, \dots, m-1\}$ où m n'est pas trop large ;
- Il ne peut pas y avoir deux termes avec la même clé ;
- Chaque position dans le tableau correspond à une clé dans U

Soit $X = \{x_1, x_2, \dots, x_n\}$ ensemble d'élément à chercher dans le tableau ; s'il y a un élément x avec la clé k, alors $T[k]$ contient un pointeur vers x. De là, on déduit sa polarité, sinon, x est inconnue.

Cette description peut être traduite en langage machine afin de permettre à l'ordinateur de procéder à l'étiquetage automatique d'opinions. L'algorithme ci-dessous décrit les étapes nécessaires pour l'étude de polarité (voir 37).

Algorithme : Étiquetage d'Opinion ;
Entrée : SenOpinion et Liste des Termes Sélectionnés ;
Sortie : Termes Étiquetés ;
Début
Résultat ← [] ;
Pour tout mot de la liste des termes sélectionnés **Faire**
    Si le mot est dans SenOpinion **Alors**
        Resultat ← SenOpinion (word, polarity, emotion, intensity);
    Fin Si
Fin Pour
Fin

Figure 37 : *Algorithme d'étiquetage d'opinions*

Après avoir étiqueté les termes séparément, nous cherchons à présent à calculer l'opinion d'un ou de plusieurs commentaires associés à un article.

7.4 - **Modèle de calcul d'opinions d'un commentaire**

Le calcul d'opinion d'un commentaire consiste à quantifier les états affectifs des termes qui composent le commentaire. Dans la pratique, notre démarche consiste à mettre en corrélation les attributs tels que l'intensité, le like et le dislike afin de déterminer le score du commentaire. Il y a plusieurs façons de calculer ces scores. Voici les formules que nous proposons pour effectuer de tels calculs (voir Figures 38, 39, 40).

- Soit $C = \{t_1, t_2, \dots, t_n\}$, un commentaire composé de $\boldsymbol{n}$ termes $t_1, t_2, \dots, t_n$;

- Soit *I* (intensité) la fonction qui, à chaque terme de notre lexique, associe une valeur comprise entre -1 et 1.

Le score d'un commentaire *C*, noté *Score(C)* est par définition la somme des intensités de termes qui composent le commentaire

$$Score(C) = \sum_{i=1}^{n} I(t_i) \quad (\mathbf{1})$$

Figure 38 : *Calcul de score d'un commentaire sans like ni dislike*

Si le commentaire *C* comporte des likes (*L*) et dislikes (*D*) alors son score est calculé comme suit :

$$Score(C) = (L+1)\sum_{\substack{i=1 \\ I(t_i) \geq 0}}^{n} I(t_i) + (D+1)\sum_{\substack{i=1 \\ I(t_i) \leq 0}}^{n} I(t_i) \quad (\mathbf{2})$$

Figure 39 : *Calcul de score d'un commentaire avec like et dislike*

- $Si\ Score(C) > 0\ alors\ le\ commentaire\ C\ est\ favorable;$
- $Si\ Score(C) < 0\ alors\ le\ commentaire\ C\ est\ defavorable;$
- $Si\ Score(C) = 0\ alors\ le\ commentaire\ C\ est\ neute.$

Figure 40 : *Opinion d'un commentaire*

À partir de ce modèle proposé, nous avons tenté de définir le baromètre de satisfaction. Un baromètre de satisfaction est pour nous la mesure du degré d'appréciation de chaque commentaire. Il est obtenu à partir d'un score que nous appelons score de normalisation. Le score de normalisation est calculé à l'aide du rapport entre l'intensité de termes sur le nombre de termes d'opinions du commentaire. Cette normalisation permet de comparer deux (02) commentaires qui ont la même orientation afin de pouvoir les comparer.

Nous proposons une normalisation du score d'un commentaire noté $Score_N(C)$ comme suit :

$$Score_N(C) = \frac{\sum_{i=1}^{n} I(t_i) \atop I(t_i) \geq 0 \; + \; \sum_{i=1}^{n} I(t_i) \atop I(t_i) \leq 0}{Card(C)} * 100$$

Figure 41 : *Calcul de score normalisé*

À présent, nous allons expérimenter notre outil sur des données réelles.

7.5 - **Résultats**

Pour appliquer notre solution sur un ensemble de données réelles, nous avons extrait des commentaires associés à un seul article pour tester et évaluer le comportement de notre outil sur des données réelles. Le Tableau 11 donne des informations sur l'article et ses commentaires.

Tableau 11 : Présentation de notre jeu de données

Titre de l'article	***Nb de commentaires***	***Nb de Like***	***Nb de Dislike***	***Nb de termes candidats***
Wade en colère : 'Macky Sall a ordonné la saisie de ma maison de point E	96	579	792	15850

Pour cela, nous avons apporté des transformations sur les documents pour faciliter le prétraitement. Ces transformations sont notées dans le Tableau 12.

Tableau 12 : Transformation de données

Eléments	Résultats
(? !;)	Point (.)
Or, mais, car, parce que, cependant	Point (.)
Autres signe de ponctuations	Espace

7.5.1 - Présentation des résultats

À l'issu de l'utilisation de notre lexique, nous pouvons voir les termes d'opinions de façon plus distincte à l'aide de la visualisation suivante (voir Figure 42).

À partir de la Figure 42, il est possible de constater les tendances dégagées dans notre jeu de données. Ceci peut être interprété par l'observation que les termes à polarité positive sont plus nombreux. Dans l'analyse détaillée, nous voulons examiner la variance émotionnelle de

chaque terme de façon plus distincte. Ce type de visualisation est basé sur les catégories émotionnelles telles que la colère, le dégoût, la joie, la peur, la surprise et la tristesse. Voici une visualisation d'émotions à la Figure 43.

Une fois les termes étiquetés, nous pouvons déterminer des statistiques avec les données à travers les différents niveaux de texte, tels que le niveau de commentaires et le niveau d'articles. Ainsi, les statistiques au niveau des commentaires sont fournies à travers cette visualisation comme le montre la figure suivante (voir Figure 44).

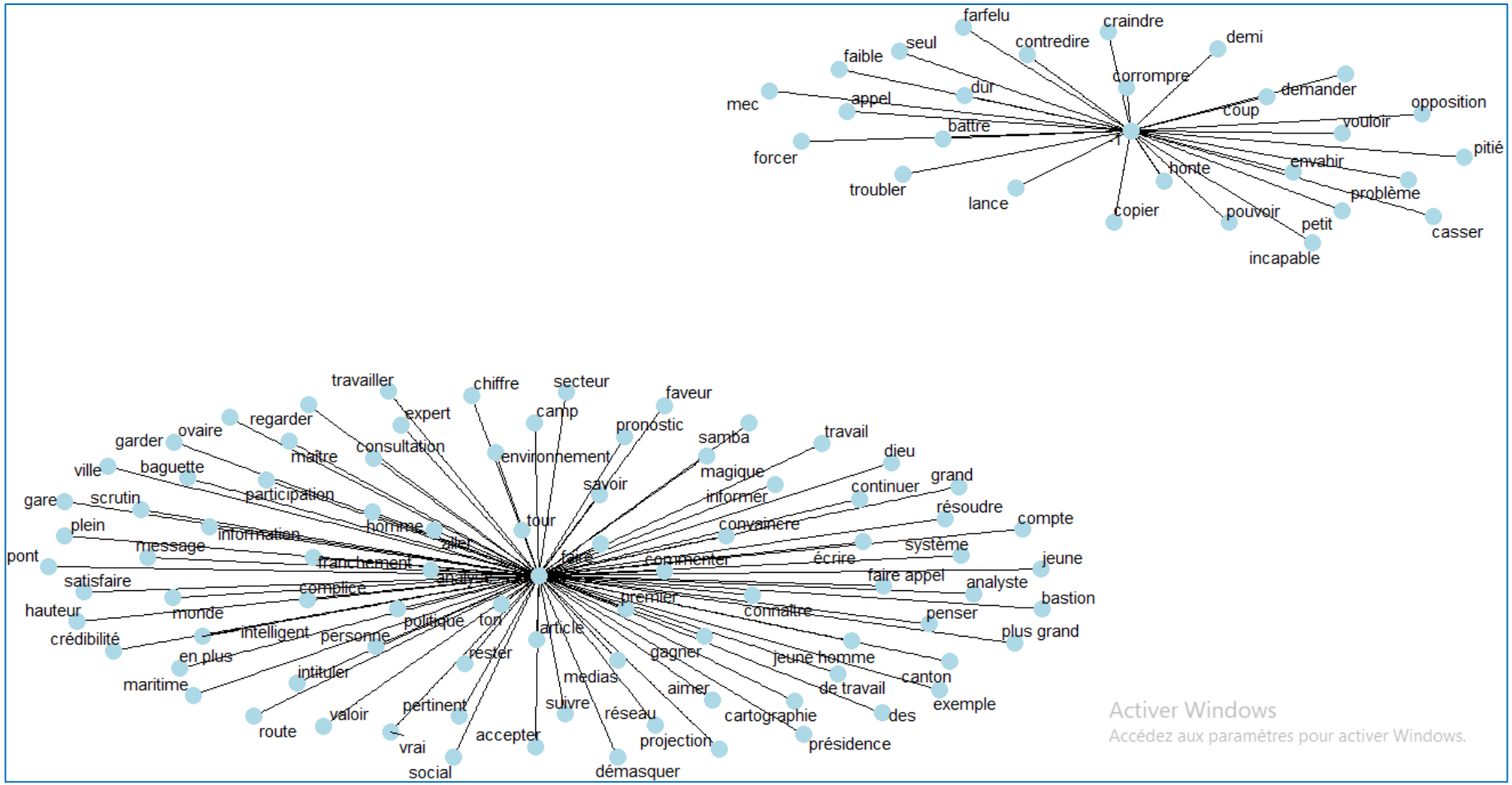


Figure 42 : *Visualisation des opinions extraites de commentaires postés pour un article*

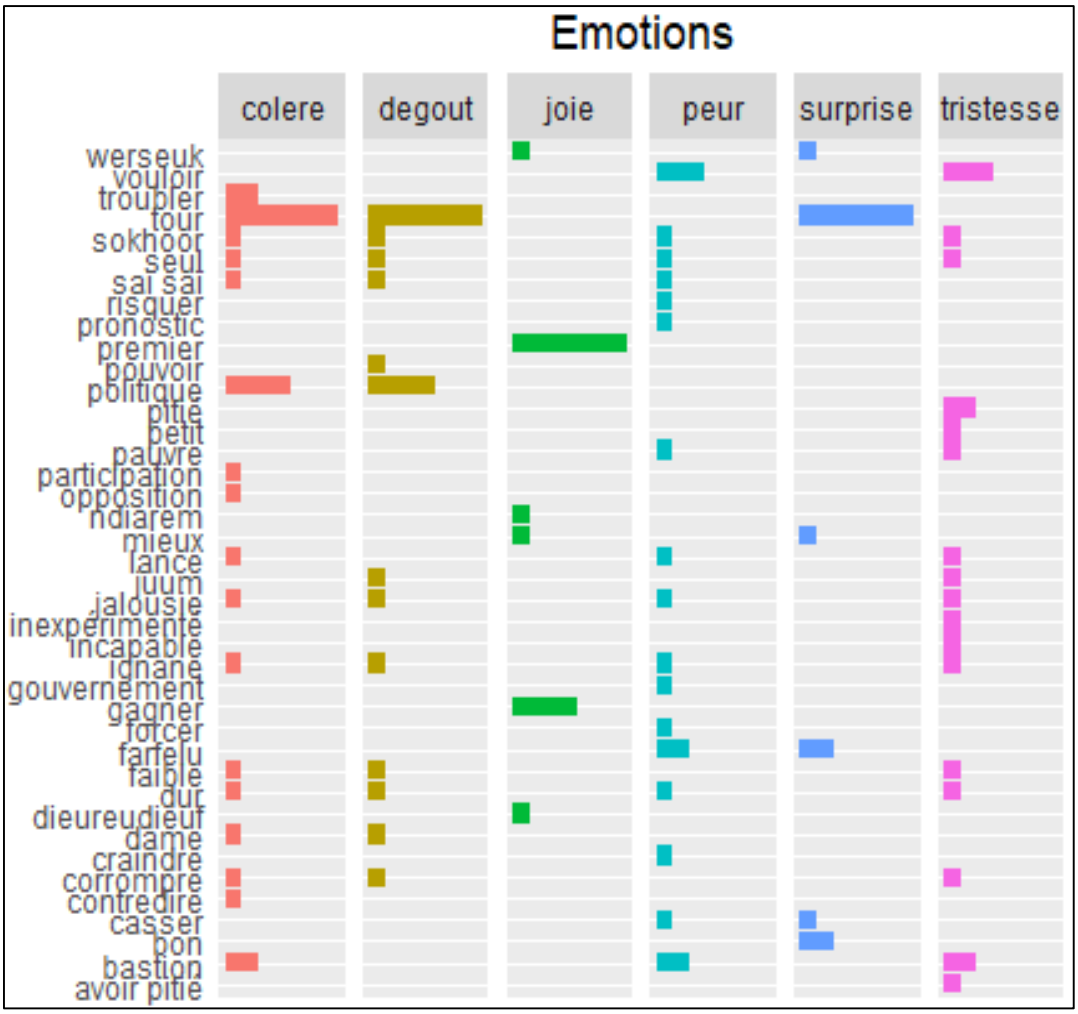


Figure 43 : *Visualisation d'émotions dans nos jeux de données*

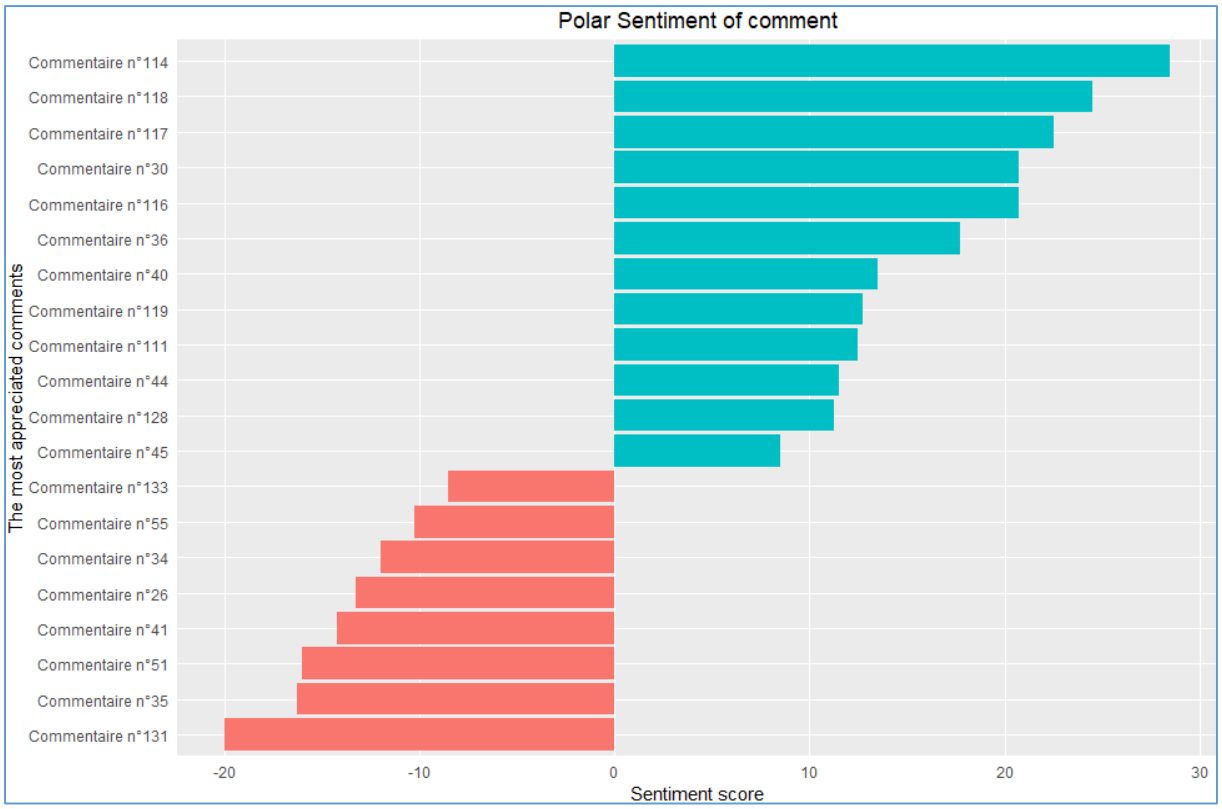


Figure 44 : *Représentation graphique des commentaires les plus subjectifs*

À travers cette visualisation, il est possible de constater l'opinion dégagé par chaque commentaire. Ceci peut être interprété par l'observation que les lecteurs sont favorables à cet article à travers les commentaires. Cela est confirmé à travers la Figure 45.

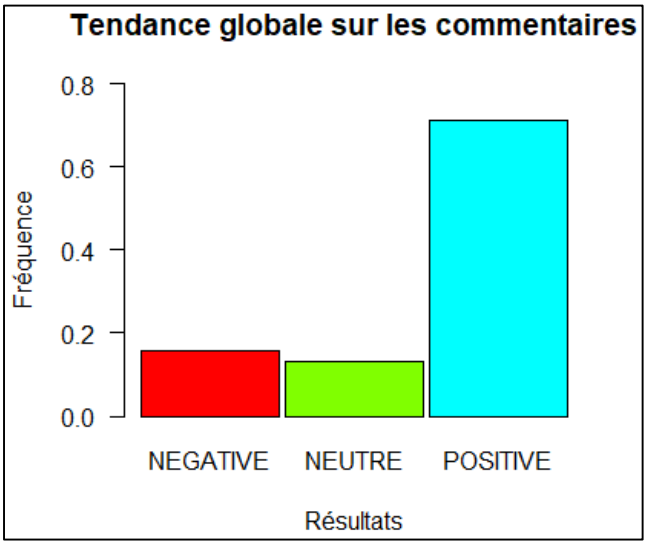


Figure 45 : *Statistiques sur la tendance globale de notre jeu de données*

L'intérêt de la fouille d'opinions est d'analyser des données textuelles afin de déduire les différentes opinions qui y sont exprimées. Les opinions ainsi extraites font ensuite l'objet de

statistiques pour déterminer le ressenti général d'une communauté. Une des façons les plus intéressantes de visualiser les commentaires associés à un article est d'observer l'évolution de l'opinion au fil du temps. Après avoir expérimenté notre solution sur des données réelles, nous pensons l'évaluer en fonction des limites afin de nous projeter dans une perspective d'amélioration.

7.5.2 - Évaluation

L'évaluation de notre lexique sur des données réelles donne les statistiques suivantes (voir tableau 13).

Tableau 13 : Statistiques de l'évaluation

Nb de termes candidats	Proportion de termes étiquetés par SenOpinion	Proportion de termes étiquetés par FEEL
15850	47,45 %	22,57 %

À la lecture du tableau 11, nous constatons que FEEL n'a étiqueté que 22,57% des termes de la basse d'analyse. Cela s'explique par le fait que FEEL ne prend pas en charge le langage urbain sénégalais qui modifie les caractéristiques orthographiques, voire grammaticales, d'une langue afin de réduire sa longueur. Dans cette situation, notre solution présente un score de 47,45% qui est plus élevé que celui de FELL. Cela nous permet de déduire que notre solution est mieux adaptée à analyser les commentaires issus de la presse en ligne.

Cependant force est de constater qu'il reste beaucoup de termes non étiquetés. Ce qui confirme davantage la complexité des commentaires issus de la presse sénégalaise en ligne. Il faut rappeler que le Wolof est une langue qui dispose d'un vocabulaire formalisé. Malheureusement, les internautes ne respectent pas les conventions dans leurs commentaires. Ils écrivent selon leurs convenances ou leurs compréhensions de l'objet. À cela s'ajoute, les abréviations personnalisées.

Pour cela, le lexique doit être amélioré afin de prendre en compte toutes les spécificités de la langue wolof et la question des abréviations personnalisées.

7.6 - **Conclusion**

En définitive, la création d'un lexique d'opinions demande beaucoup d'efforts humains surtout dans un contexte où les outils de traitement automatique des langages naturels sont quasi

inexistants. Cette activité de collecte et d'annotation de données invite les experts notamment les linguistes à vérifier et valider les résultats proposés. Néanmoins, nous avons, d'une part, créé un lexique d'opinions, d'autre part, proposé une méthode d'utilisation tant pour étiqueter des opinions que pour le calcul de score sur les commentaires provenant de la presse sénégalaise en ligne. En effet, la caractérisation automatique des opinions présentes dans les textes est devenue un enjeu majeur pour des applications telles que le système de fouille d'opinions. Notre système fonctionne comme un moteur de fouille d'opinions qui apporte un condensé des opinions rencontrées. Il est réalisé pour résumer des points de vue afin d'extraire l'opinion majoritaire des internautes. Notre réelle motivation est de doter les langues nationales de ressources et méthodes de fouille d'opinions.

A l'avenir, nous comptons étendre cette représentation du lexique pour en faire une représentation ontologique. L'introduction de ces ressources dans un système de fouille d'opinions vise à réduire, voire éliminer, la confusion conceptuelle et terminologique et à tendre vers une compréhension partagée pour améliorer la communication, le partage, l'interopérabilité et le degré de réutilisation possible. En outre, il n'existe pas encore de corpus annoté permettant de valider notre outil. Pour cela, nous allons annoter une collection de commentaires manuellement de concert avec les experts. Ce corpus peut être mis à la disposition de la communauté scientifique pour les besoins de validation des méthodes qui sont expérimentées sur ces types de données.

8 -CONCLUSION GÉNÉRALE

8.1 - **Synthèse**

À l'heure du bilan, nous pouvons retenir que l'avènement des technologies web 2.0 a entrainé une grande mutation dans l'évolution des médias. Nous avons évolué de médias classiques dans lesquels les lecteurs sont considérés comme des consommateurs d'informations aux médias numériques qui proposent une interaction entre les internautes. Cette émergence a provoqué l'accélération de la production et de la circulation de l'information notamment via les réseaux sociaux, les blogs et la presse en ligne. Elle a aussi contribué à l'accès quasi incontrôlable à l'information via l'internet, la démocratisation de la production d'informations avec la participation d'internautes à travers des commentaires. Dans notre pays, le journalisme en ligne a montré ses preuves tant dans la diffusion de l'information que dans la participation des lecteurs. Ainsi, le lecteur donne son point de vue sur les différentes publications (après lecture). Les données issues de ces sources contiennent d'informations capitales. La fouille d'opinions sur les données issues de la presse sénégalaise en ligne notamment les commentaires peut avoir un enjeu stratégique chez les managers lors des prises de décision.

La fouille d'opinions est un processus de collecte et d'analyse de données spécialement dédié au traitement automatique de contenus textuels. Cette technique vise à classer des documents d'un corpus en fonction des positions favorable, défavorable ou neutre. Elle se distingue du Traitement Automatique des Langages Naturels (TALN) et de la Recherche d'Informations. Pendant cette dernière décennie, beaucoup de travaux se sont penchés sur cette problématique, en prenant le problème sous différents angles (principalement statistique et/ou linguistique). Elle est très utilisée par les acteurs socio-économiques et politiques dans l'orientation de leurs politiques.

Cependant, en raison de l'hétérogénéité des sources et de la complexité sémantico-syntaxique et lexicale des commentaires issus de la presse sénégalaise en ligne, les ressources et méthodes existantes en fouille d'opinions sont inappropriées pour analyser tous ces commentaires. Face à cette situation, nous avons proposé des solutions comme contributions à la recherche.

8.2 - **Contributions**

Nos contributions s'articulent autour des points suivants :

- ***Formalisation de la complexité des commentaires issus de la presse sénégalaise en ligne*** [147] : ici nous avons répertorié les difficultés que posent réellement les commentaires en ligne des sénégalais en termes d'interprétations. Nous avons identifié deux phénomènes majeurs à savoir l'utilisation du langage urbain et les commentaires hors sujet. Ainsi, nous avons formalisé cette problématique d'interprétation en trois (03) obstacles qui sont : les obstacles liés à l'ambiguïté, au multi-domaine et au multilinguisme. C'est une contribution à l'état de l'art.
- ***Modélisation d'un commentaire et d'un réseau de commentaires journalistiques*** [165]: cette étude répond simultanément à trois (03) besoins. Le premier besoin s'inscrit dans la facilitation de la collecte, la fusion et le stockage. Le deuxième répond à la complexité dans la mise en œuvre de la fouille d'opinions basée sur les aspects. En dernier temps, le besoin de visualiser les commentaires à travers les relations mises en exergue.
- ***Architecture d'un système de fouille d'opinions dans la presse sénégalaise en ligne*** [146] : l'architecture proposée représente le système de fouille d'opinions envisagée qui sera une plateforme web sémantique. Cette architecture décrit amplement les différents modules et leurs interactions à travers un processus global.
- ***OpinionScraper*** [148]: c'est l'outil d'acquisition, de catégorisation et de stockage de données en provenance de la presse en ligne. Cet outil basé sur une méthode innovante, a résolu la contraignante question de l'hétérogénéité des sources qui est une problématique contemporaine du web scraping.
- ***SenOpinion*** [152]: SenOpinion est un lexique constitué sur la base du langage urbain afin de doter nos langues de ressources de traitement automatique.
- ***Proposition de méthode de fouille d'opinions*** : notre modèle classification non supervisée est une approche innovante permettant de déterminer les tendances sur la base de l'intensité de termes, likes et dislikes qui composent un document (commentaire).

8.3 - **Perspectives**

En guise de perspectives, nous envisageons :

- **Une méthode d'apprentissage automatique pour l'enrichissement du lexique** : La méthode pour l'enrichissement du lexique sera une approche basée sur l'apprentissage supervisée. Ici, le lexique fera office de corpus d'entrainement. Son enrichissement nous

permettra de prendre en compte le maximum de termes afin de fournir des résultats précis, fiables et proches de la réalité.

- **Une ontologie d'évènements pour le journalisme sénégalais en langage urbain comme label** : Une ontologie est un ensemble structuré de concepts organisés dans un graphe, liés par des relations sémantiques et logiques. L'extension de la représentation de notre lexique en représentation ontologique aura un impact en termes de performance et de fiabilité sur la qualité des résultats obtenus. L'introduction de ces ressources dans un système de fouille d'opinions vise à réduire, voire éliminer, la confusion conceptuelle et terminologique et à tendre vers une compréhension partagée pour améliorer la communication, le partage, l'interopérabilité et le degré de réutilisation possible. Par la même occasion, nous mettrons en place un modèle d'indexation sémantique.
- **Un modèle de visualisation** : À l'avenir, nous proposerons un modèle de visualisation des résultats de recherche à travers un portail de consultation. En effet, l'information utile est encore enfouie et difficile à retrouver, ce qui nécessite des techniques de visualisation efficaces et adaptées à ce contexte particulier. La visualisation de termes d'opinions contenus dans un corpus de documents à travers un "nuage de termes" construit à partir de l'ensemble de termes discriminants responsables de la classification. Souvent, ce sont des termes clés qui doivent être fréquents et discriminants vis-à-vis de la classe d'opinion qu'ils caractérisent.
- **Une méthode de correction automatique** : Une méthode de correction automatique peut permettre de corriger certains usages considérés comme des fautes grammaticales en français ou des emprunts.

RÉFÉRENCES

[1] A. Jeyapriya et C. K. Selvi, « Extracting aspects and mining opinions in product reviews using supervised learning algorithm », in *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, 2015, p. 548-552.

[2] B. Liu et L. Zhang, « A survey of opinion mining and sentiment analysis », in *Mining text data*, Springer, 2012, p. 415-463.

[3] L. Sproull et S. Kiesler, « Connections: New ways of working in the networked organization », *Camb. MA*, 1991.

[4] W. Medhat, A. Hassan, et H. Korashy, « Sentiment analysis algorithms and applications: A survey », *Ain Shams Eng. J.*, vol. 5, nº 4, p. 1093-1113, 2014.

[5] S.-M. Kim et E. Hovy, « Determining the sentiment of opinions », in *Proceedings of the 20th international conference on Computational Linguistics*, 2004, p. 1367.

[6] A. Mudinas, D. Zhang, et M. Levene, « Combining lexicon and learning based approaches for concept-level sentiment analysis », in *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, 2012, p. 1-8.

[7] S. Tufféry, *Data mining et statistique décisionnelle : La science des données*, Cinquième édition. Editions Technip, 2017.

[8] S. Tufféry, *Data mining et statistique décisionnelle: l'intelligence dans les bases de données*. Editions Technip, 2005.

[9] R. K. Lomotey et R. Deters, « RSenter: Tool for topics and terms extraction from unstructured data debris », in *Big Data (BigData Congress), 2013 IEEE International Congress on*, 2013, p. 395-402, Consulté le: juill. 09, 2017. [En ligne]. Disponible sur: http://ieeexplore.ieee.org/abstract/document/6597163/.

[10] R. K. Lomotey et R. Deters, « Analytics-as-a-service (aaas) tool for unstructured data mining », in *Cloud Engineering (IC2E), 2014 IEEE International Conference on*, 2014, p. 319-324, Consulté le: sept. 15, 2017. [En ligne]. Disponible sur: http://ieeexplore.ieee.org/abstract/document/6903489/.

[11] J. Singh et V. Gupta, « A systematic review of text stemming techniques », *Artif. Intell. Rev.*, vol. 48, nº 2, p. 157-217, 2017.

[12] R. Kosala et H. Blockeel, « Web mining research: A survey », *ACM Sigkdd Explor. Newsl.*, vol. 2, nº 1, p. 1-15, 2000.

[13] É. Gaussier et F. Yvon, *Modèles statistiques pour l'accès à l'information textuelle*. Lavoisier, 2011.

[14] V. Tunali et T. T. Bilgin, « PRETO: A high-performance text mining tool for preprocessing turkish texts », in *Proceedings of the 13th International Conference on Computer Systems and Technologies*, 2012, p. 134-140.

[15] U. Baskaran et K. Ramanujam, « Automated scraping of structured data records from health discussion forums using semantic analysis », *Inform. Med. Unlocked*, vol. 10, p. 149-158, 2018.

[16] K. Sundaramoorthy, R. Durga, et S. Nagadarshini, « NewsOne—An Aggregation System for News Using Web Scraping Method », in *Technical Advancements in Computers and Communications (ICTACC), 2017 International Conference on*, 2017, p. 136-140.

[17] H. A. Sleiman et R. Corchuelo, « Trinity: on using trinary trees for unsupervised web data extraction », *IEEE Trans. Knowl. Data Eng.*, vol. 26, nº 6, p. 1544-1556, 2014.

[18] S. Khalil et M. Fakir, « RCrawler: An R package for parallel web crawling and scraping », *SoftwareX*, vol. 6, p. 98-106, 2017.

[19] W. Nadee et K. Prutsachainimmit, « Towards data extraction of dynamic content from JavaScript Web applications », in *2018 International Conference on Information Networking (ICOIN)*, janv. 2018, p. 750-754, doi: 10.1109/ICOIN.2018.8343218.

[20] S. Chakrabarti, *Mining the Web: Discovering knowledge from hypertext data*. Elsevier, 2002.

[21] L. Ou-Yang, *News, full-text, and article metadata extraction in Python 3. Advanced docs:: codelucas/newspaper*. 2018.

[22] H. Wickham, « Rvest: Easily harvest (scrape) web pages », *R Package Version 03*, vol. 1, 2015.

[23] K. Sundaramoorthy, R. Durga, et S. Nagadarshini, « NewsOne : An Aggregation System for News Using Web Scraping Method », in *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*, avr. 2017, p. 136-140, doi: 10.1109/ICTACC.2017.43.

[24] E. N. Sarr, S. Ousmane, et A. Diallo, « FactExtract: Automatic Collection and Aggregation of Articles and Journalistic Factual Claims from Online Newspaper », in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018, p. 336-341.

[25] H. Schmid, « Treetagger| a language independent part-of-speech tagger », *Inst. Für Maschinelle Sprachverarbeitung Univ. Stuttg.*, vol. 43, p. 28, 1995.

[26] F. M. Hasan, N. UzZaman, et M. Khan, « Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla », in *Advances and innovations in systems, computing sciences and software engineering*, Springer, 2007, p. 121-126.

[27] A. Urieli, « Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit », PhD Thesis, Université Toulouse le Mirail-Toulouse II, 2013.

[28] S. E. Lee et S. S. Han, « Qtag: introducing the qualitative tagging system », in *Proceedings of the eighteenth conference on Hypertext and hypermedia*, 2007, p. 35-36.

[29] E. Brill, « A simple rule-based part of speech tagger », in *Proceedings of the third conference on Applied natural language processing*, 1992, p. 152-155.

[30] P. Denis et B. Sagot, « Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français », 2010.

[31] S. Tata et J. M. Patel, « Estimating the selectivity of tf-idf based cosine similarity predicates », *ACM Sigmod Rec.*, vol. 36, n° 2, p. 7-12, 2007.

[32] M. Anjaria et R. M. R. Guddeti, « A novel sentiment analysis of social networks using supervised learning », *Soc. Netw. Anal. Min.*, vol. 4, n° 1, p. 181, 2014.

[33] A. De et S. K. Kopparapu, « Unsupervised clustering technique to harness ideas from an Ideas Portal », in *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, 2013, p. 1563-1568.

[34] M. Berry et G. Linoff, « Data Mining: Techniques appliquées au marketing, à la vente et aux services clients », 1997, Consulté le: juill. 14, 2017. [En ligne]. Disponible sur: http://www.citeulike.org/group/14833/article/8969556.

[35] E. G. Talbi, « Fouille de données (Data Mining): Un tour d'horizon », *Lab. D'informatique Lille*.

[36] N. BECK, « Application de méthodes de clustering traditionnelles et extension au cadre multicritère », *Mém. Fin D'études Univ. Libre Brux.*, 2006.

[37] P. Georges, « REDUCTION DE BASE DE DONNEES PAR LA CLASSIFICATION AUTOMATIQUE », 2004, Consulté le: juill. 17, 2017. [En ligne]. Disponible sur: https://pdfs.semanticscholar.org/2d6b/2102ddd8297c9afd30620de18938ba9b866c.pdf.

[38] Y. Matsuo et M. Ishizuka, « Keyword extraction from a single document using word co-occurrence statistical information », *Int. J. Artif. Intell. Tools*, vol. 13, n° 01, p. 157-169, 2004.

[39] Z. Liu, P. Li, Y. Zheng, et M. Sun, « Clustering to find exemplar terms for keyphrase extraction », in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 2009, p. 257-266, Consulté le: juill. 02, 2017. [En ligne]. Disponible sur: http://dl.acm.org/citation.cfm?id=1699544.

[40] R. Mihalcea et P. Tarau, « TextRank: Bringing Order into Text. », in *EMNLP*, 2004, vol. 4, p. 404-411, Consulté le: juill. 02, 2017. [En ligne]. Disponible sur: https://dias.users.greyc.fr/ict/mihalcea-2004.pdf.

[41] W. Jin, H. H. Ho, et R. K. Srihari, « OpinionMiner: a novel machine learning system for web opinion mining and extraction », in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, p. 1195-1204.

[42] W. Medhat, A. Hassan, et H. Korashy, « Sentiment analysis algorithms and applications: A survey », *Ain Shams Eng. J.*, vol. 5, n° 4, p. 1093-1113, 2014.

[43] K.-M. Schneider, « Techniques for improving the performance of naive bayes for text classification », in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2005, p. 682-693.

[44] L. Wikarsa et S. N. Thahir, « A text mining application of emotion classifications of Twitter's users using Naive Bayes method », in *2015 1st International Conference on Wireless and Telematics (ICWT)*, 2015, p. 1-6.

[45] E. Charton et R. Acuna-Agost, « Quel modèle pour détecter une opinion? Trois propositions pour généraliser l'extraction d'une idée dans un corpus », *Actes Trois. DÉfi Fouille Textes*, p. 35, 2007.

[46] M. A. Aizerman, « Theoretical foundations of the potential function method in pattern recognition learning », *Autom. Remote Control*, vol. 25, p. 821-837, 1964.

[47] A. S. Manek, P. D. Shenoy, M. C. Mohan, et K. R. Venugopal, « Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier », *World Wide Web*, vol. 20, n° 2, p. 135-154, 2017.

[48] A. L. Berger, V. J. D. Pietra, et S. A. D. Pietra, « A maximum entropy approach to natural language processing », *Comput. Linguist.*, vol. 22, n° 1, p. 39-71, 1996.

[49] X. Yan et T. Huang, « Tibetan sentence sentiment analysis based on the maximum entropy model », in *Broadband and Wireless Computing, Communication and Applications (BWCCA), 2015 10th International Conference on*, 2015, p. 594-597.

[50] T. Kohonen, *Content-addressable memories*, vol. 1. Springer Science & Business Media, 2012.

[51] S. Poria, E. Cambria, et A. Gelbukh, « Aspect extraction for opinion mining with a deep convolutional neural network », *Knowl.-Based Syst.*, vol. 108, p. 42-49, 2016.

[52] F. Scholer, D. Kelly, et B. Carterette, « Information retrieval evaluation using test collections », *Inf. Retr. J.*, vol. 19, nº 3, p. 225-229, 2016.

[53] A. Singhal, « Modern information retrieval: A brief overview », *IEEE Data Eng Bull*, vol. 24, nº 4, p. 35-43, 2001.

[54] T. Ono, H. Hishigaki, A. Tanigami, et T. Takagi, « Automated extraction of information on protein–protein interactions from the biological literature », *Bioinformatics*, vol. 17, nº 2, p. 155-161, 2001.

[55] I. Tellier, « Introduction `a la fouille de textes universit´e de Paris 3 - Sorbonne Nouvelle Table des mati`eres ». .

[56] T. Poibeau, *Extraction automatique d'information: Du texte brut au web sémantique*. ISBN.

[57] C. Nedellec, M. O. A. Vetah, et P. Bessières, « Sentence filtering for information extraction in genomics, a classification problem », in *European Conference on Principles of Data Mining and Knowledge Discovery*, 2001, p. 326-337.

[58] L. Tanguy, « Traitement automatique de la langue naturelle et interprétation: contribution à l'élaboration d'un modèle informatique de la sémantique interprétative », PhD Thesis, Université de Rennes 1, 1997.

[59] A. Daud, W. Khan, et D. Che, « Urdu language processing: a survey », *Artif. Intell. Rev.*, vol. 47, nº 3, p. 279-311, 2017.

[60] D. Jurasky et J. H. Martin, « Speech and Language Processing: An introduction to natural language Processing », *Comput. Linguist. Speech Recognit. Prentice Hall N. J.*, 2000.

[61] K. Humphreys, G. Demetriou, et R. Gaizauskas, « Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures », in *Biocomputing 2000*, World Scientific, 1999, p. 505-516.

[62] S. Hengchen, S. Van Hooland, R. Verborgh, et M. De Wilde, « L'extraction d'entités nommées: une opportunité pour le secteur culturel? », *I2D Inf. Donnees Doc.*, vol. 52, n° 2, p. 70-79, 2015.

[63] M.-F. Moens, *Information extraction: algorithms and prospects in a retrieval context*, vol. 21. Springer Science & Business Media, 2006.

[64] R. Plutchik, « A general psychoevolutionary theory of emotion », in *Theories of emotion*, Elsevier, 1980, p. 3-33.

[65] R. Plutchik, « Théorie des émotions », *Wikipédia*. mars 21, 2019, Consulté le: oct. 24, 2019. [En ligne]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Robert_Plutchik&oldid=157735068.

[66] P. Ekman, « Basic emotions », *Handb. Cogn. Emot.*, vol. 98, n° 45-60, p. 16, 1999.

[67] E. Cambria, A. Livingstone, et A. Hussain, « The hourglass of emotions », in *Cognitive behavioural systems*, Springer, 2012, p. 144-157.

[68] P. Skórzewski, « Using book dialogs to extract emotions from texts in Polish », *dimension*, vol. 1, p. 1, 2019.

[69] L. Brisson, « Platon, Timée, Critias », *Trad. Inéd. Introd. Notes Paris Flammarion*, 1992.

[70] P. D. Turney, « Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews », in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, p. 417-424.

[71] B. Pang, L. Lee, et S. Vaithyanathan, « Thumbs up?: sentiment classification using machine learning techniques », in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, p. 79-86.

[72] N. Jindal et B. Liu, « Identifying comparative sentences in text documents », in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, p. 244-251.

[73] R. Bhalla, « Designing an Authentic and Opinion Mining Framework in Big Data for Feature Sentiment Mining and Feature Opinion Pairs », PhD Thesis, Lovely Professional University, 2019.

[74] L. Sun, S. Li, J. Li, et J. Lv, « A novel context-based implicit feature extracting method », in *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, 2014, p. 420-424.

[75] A. Mukherjee et B. Liu, « Aspect extraction through semi-supervised modeling », in *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1*, 2012, p. 339-348.

[76] S. P. Karunathilake, J. Shamal, R. G. H. Pemathilake, et G. U. Ganegoda, « Feature Extraction from Online User Reviews », in *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2018, p. 265-272.

[77] B. Liu, « Sentiment Analysis and Subjectivity. », *Handb. Nat. Lang. Process.*, vol. 2, p. 627-666, 2010.

[78] R. Rakotomalala, « WM.A - Opinion mining and sentiment analysis », 2017. https://www.google.com/search?ei=a7KpXqKGFt6C1fAPiLGS4A8&q=WM.A+-+Opinion+mining+and+sentiment+analysis&oq=WM.A+-+Opinion+mining+and+sentiment+analysis&gs_lcp=CgZwc3ktYWIQAzoECAAQR1DqgLwBWOqAvAFgk4i8AWgAcAF4AIABhgWIAYYFkgEDNS0xmAEAoAECoAEBqgEHZ3dzLXdpeg&sclient=psy-ab&ved=0ahUKEwji59r1jY7pAhVeQRUIHYiYBPwQ4dUDCAw&uact=5 (consulté le avr. 30, 2020).

[79] M. Hu et B. Liu, « Mining and summarizing customer reviews », in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, p. 168-177.

[80] B. Liu, M. Hu, et J. Cheng, « Opinion observer: analyzing and comparing opinions on the web », in *Proceedings of the 14th international conference on World Wide Web*, 2005, p. 342-351.

[81] L. Zhang et B. Liu, « Aspect and entity extraction for opinion mining », in *Data mining and knowledge discovery for big data*, Springer, 2014, p. 1-40.

[82] F. Hemmatian et M. K. Sohrabi, « A survey on classification techniques for opinion mining and sentiment analysis », *Artif. Intell. Rev.*, p. 1-51, 2017.

[83] E. Gupta, G. Rathee, P. Kumar, et D. S. Chauhan, « Mood swing analyser: a dynamic sentiment detection approach », *Proc. Natl. Acad. Sci. India Sect. Phys. Sci.*, vol. 85, n° 1, p. 149-157, 2015.

[84] X. Pu, G. Wu, et C. Yuan, « Exploring overall opinions for document level sentiment classification with structural SVM », *Multimed. Syst.*, vol. 25, n° 1, p. 21-33, 2019.

[85] J. Wiebe et E. Riloff, « Creating subjective and objective sentence classifiers from unannotated texts », in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2005, p. 486-497.

[86] T. Wilson *et al.*, « OpinionFinder: A system for subjectivity analysis », in *Proceedings of hlt/emnlp on interactive demonstrations*, 2005, p. 34-35.

[87] O. Appel, F. Chiclana, J. Carter, et H. Fujita, « A hybrid approach to the sentiment analysis problem at the sentence level », *Knowl.-Based Syst.*, vol. 108, p. 110-124, 2016.

[88] V. S. Shirsat, R. S. Jagdale, et S. N. Deshmukh, « Sentence Level Sentiment Identification and Calculation from News Articles Using Machine Learning Techniques », in *Computing, Communication and Signal Processing*, Springer, 2019, p. 371-376.

[89] T. C. Chinsha et S. Joseph, « A syntactic approach for aspect based opinion mining », in *2015 IEEE International Conference on Semantic Computing (ICSC)*, 2015, p. 24-31.

[90] Y. Toussaint, « Fouille de textes: des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances », PhD Thesis, 2011.

[91] K. Dramé, I. Diop, L. Faty, et B. Ndoye, « Indexation et appariement de documents cliniques avec le modèle vectoriel », *DEFT*, p. 91, 2019.

[92] A. M. Kaplan et M. Haenlein, « Users of the world, unite! The challenges and opportunities of Social Media », *Bus. Horiz.*, vol. 53, nº 1, p. 59-68, 2010.

[93] S. Kiritchenko, X. Zhu, et S. M. Mohammad, « Sentiment analysis of short informal texts », *J. Artif. Intell. Res.*, vol. 50, p. 723-762, 2014.

[94] M. R. Wigan et R. Clarke, « Big data's big unintended consequences », *Computer*, vol. 46, nº 6, p. 46-53, 2013.

[95] A. Andreevskaia et S. Bergler, « Semantic tag extraction from WordNet glosses », 2006.

[96] G. A. Miller, « WordNet: a lexical database for English », *Commun. ACM*, vol. 38, nº 11, p. 39-41, 1995.

[97] M. Zhao et H. Schütze, « A Multilingual BPE Embedding Space for Universal Sentiment Lexicon Induction », in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, p. 3506-3517.

[98] M. Lafourcade, N. Le Brun, et A. Joubert, « Construire un lexique de sentiments par crowdsourcing et propagation », 2016.

[99] A. Abdaoui, J. Azé, S. Bringay, et P. Poncelet, « Feel: a french expanded emotion lexicon », *Lang. Resour. Eval.*, vol. 51, nº 3, p. 833-855, 2017.

[100] S. M. Mohammad et P. D. Turney, « Crowdsourcing a word–emotion association lexicon », *Comput. Intell.*, vol. 29, n° 3, p. 436-465, 2013.

[101] D. Kandé, F. Camara, S. Ndiaye, et F. M. Guirassy, « FWLSA-score: French and Wolof Lexicon-based for Sentiment Analysis », in *2019 5th International Conference on Information Management (ICIM)*, 2019, p. 215-220.

[102] S. Li, L. Zhou, et Y. Li, « Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures », *Inf. Process. Manag.*, vol. 51, n° 1, p. 58-67, 2015.

[103] V. Hatzivassiloglou et K. R. McKeown, « Predicting the semantic orientation of adjectives », in *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, 1997, p. 174-181.

[104] K. Ravi et V. Ravi, « A survey on opinion mining and sentiment analysis: tasks, approaches and applications », *Knowl.-Based Syst.*, vol. 89, p. 14-46, 2015.

[105] J. A. Balazs et J. D. Velásquez, « Opinion mining and information fusion: a survey », *Inf. Fusion*, vol. 27, p. 95-110, 2016.

[106] S. Sun, C. Luo, et J. Chen, « A review of natural language processing techniques for opinion mining systems », *Inf. Fusion*, vol. 36, p. 10-25, 2017.

[107] G. Wang, D. Zheng, S. Yang, et J. Ma, « FCE-SVM: a new cluster based ensemble method for opinion mining from social media », *Inf. Syst. E-Bus. Manag.*, vol. 16, n° 4, p. 721-742, 2018.

[108] S. Riaz, M. Fatima, M. Kamran, et M. W. Nisar, « Opinion mining on large scale data using sentiment analysis and k-means clustering », *Clust. Comput.*, p. 1-16, 2017.

[109] V. Hatzivassiloglou et J. M. Wiebe, « Effects of adjective orientation and gradability on sentence subjectivity », in *Proceedings of the 18th conference on Computational linguistics-Volume 1*, 2000, p. 299-305.

[110] B. Pang et L. Lee, « Opinion mining and sentiment analysis », *Found. Trends® Inf. Retr.*, vol. 2, n° 1-2, p. 1-135, 2008.

[111] K. Sparck Jones, « A statistical interpretation of term specificity and its application in retrieval », *J. Doc.*, vol. 28, n° 1, p. 11-21, 1972.

[112] F. Husson, S. Lê, et J. Pagès, *Analyse de données avec R*. Presses universitaires de Rennes, 2016.

[113] P. Gonçalves, M. Araújo, F. Benevenuto, et M. Cha, « Comparing and combining sentiment analysis methods », in *Proceedings of the first ACM conference on Online social networks*, 2013, p. 27-38.

[114] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, et F. Benevenuto, « Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods », *EPJ Data Sci.*, vol. 5, n° 1, p. 1-29, 2016.

[115] S. L. Lo, E. Cambria, R. Chiong, et D. Cornforth, « Multilingual sentiment analysis: from formal to informal and scarce resource languages », *Artif. Intell. Rev.*, vol. 48, n° 4, p. 499-527, 2017.

[116] M. J. C. Samonte, J. M. R. Garcia, V. J. L. Lucero, et S. C. B. Santos, « Sentiment and opinion analysis on Twitter about local airlines », in *Proceedings of the 3rd International Conference on Communication and Information Processing*, 2017, p. 415-422.

[117] O. J. Gambino et H. Calvo, « Modeling distribution of emotional reactions in social media using a multi-target strategy », *J. Intell. Fuzzy Syst.*, vol. 34, n° 5, p. 2837-2847, 2018.

[118] C. Liao, C. Feng, S. Yang, et H.-Y. Huang, « A hybrid method of domain lexicon construction for opinion targets extraction using syntax and semantics », *J. Comput. Sci. Technol.*, vol. 31, n° 3, p. 595-603, 2016.

[119] H. Keshavarz et M. S. Abadeh, « ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs », *Knowl.-Based Syst.*, vol. 122, p. 1-16, 2017.

[120] L. Faty, M. Ndiaye, I. Diop, et K. Drame, « The complexity of comments from Senegalese online presses face with opinion mining methods », in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, 2019, p. 1-6.

[121] C. Tremblay, « Qu'est-ce que le plurilinguisme? », *Bull. Eur. Sci. Soc.*, vol. 12, p. 39-57, 2015.

[122] M. Heller et Á. Rényi, « Discourse strategies in the new Hungarian public sphere: From the Populist-Urban controversy to the Hungarian–Jewish confrontation », *Öffentl. Konfliktdiskurse Um Restit. Von Gerechtigk. Polit. Verantwort. Natl. Ident. FrankfurtMain Peter Lang Verl.*, p. 373-392, 1996.

[123] M. Heller, D. Némedi, et Á. Rényi, « Structural changes in the Hungarian public sphere under state socialism », *Comp. Soc. Res.*, vol. 14, p. 157-171, 1994.

[124] P. Pupier, « Une première systématique des évaluatifs en français », *Rev. Québécoise Linguist.*, vol. 26, nº 1, p. 51-78, 1998.

[125] E. N. Sarr, O. Sall, A. Maiga, L. Faty, et R. M. Marone, « Automatic Segmentation and tagging of facts in French for automated fact-checking », in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, p. 5439-5441.

[126] B. Sagot, D. Nouvel, V. Mouilleron, et M. Baranes, « Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel », in *TALN-Traitement Automatique du Langage Naturel*, 2013, p. 407-420.

[127] E. Riloff et J. Wiebe, « Learning extraction patterns for subjective expressions », in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, p. 105-112.

[128] R. Navigli, « Word sense disambiguation: A survey », *ACM Comput. Surv. CSUR*, vol. 41, nº 2, p. 1-69, 2009.

[129] A.-J. Arnaud, « Jean-Louis Le Moigne, La modélisation des systèmes complexes, 1990 », *Droit Société*, vol. 19, nº 1, p. 424-424, 1991.

[130] L. Faty *et al.*, « News Comments Modeling for Opinion Mining: The Case of Senegalese Online Press », in *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2020, p. 1-5.

[131] A. Rochfeld et J. Morejon, *La méthode MERISE*. Les Éditions d'organisation, 1989.

[132] P. Roques et F. Vallée, « UML 2 en action », *L'analyse Besoins I Concept. J2EE*, p. 15, 2004.

[133] T. O'reilly, *Web 2.0: compact definition*. 2005.

[134] É. NEVEU, *Sociologie du journalisme*. La Découverte, 2019.

[135] A. FOUCRET, *NoSQL*, Http://www.smile.fr/Ressources/Livres-Blancs/Culture-Du-Web/Nosql. .

[136] A. Dey, A. Fekete, et U. Röhm, « Scalable transactions across heterogeneous NoSQL key-value data stores », *Proc. VLDB Endow.*, vol. 6, nº 12, p. 1434-1439, 2013.

[137] K. Chodorow, *MongoDB: the definitive guide*. O'Reilly Media, Inc., 2013.

[138] J. Han, E. Haihong, G. Le, et J. Du, « Survey on NoSQL database », in *Pervasive computing and applications (ICPCA), 2011 6th international conference on*, 2011, p. 363-366, Consulté le: août 04, 2017. [En ligne]. Disponible sur: http://ieeexplore.ieee.org/abstract/document/6106531/.

[139] « Classification des systèmes de stockage NoSQL », *Sogilis*, nov. 18, 2014. http://sogilis.com/blog/classification-systemes-stockage-nosql/ (consulté le mars 08, 2017).

[140] K. Morfonios, S. Konakas, Y. Ioannidis, et N. Kotsis, « ROLAP implementations of the data cube », *ACM Comput. Surv. CSUR*, vol. 39, n° 4, p. 12, 2007.

[141] P. Windrum, G. Fagiolo, et A. Moneta, « Empirical validation of agent-based models: Alternatives and prospects », *J. Artif. Soc. Soc. Simul.*, vol. 10, n° 2, p. 8, 2007.

[142] M.-C. Daniel-Vatonne et C. De la Higuera, « Les termes: un modèle algébrique de représentation et de structuration de données symboliques », *Mathématiques Sci. Hum.*, vol. 122, p. 41-63, 1993.

[143] F. E. Stiftung, *Barometre des medias Africains : Première analyse locale du paysage médiatique en Afrique*. Fesmedia Afrique Windhoek, 2013.

[144] E. N. Sarr, O. Sall, et A. Diagne, « SenFact Algorithm: Fact-checking by the confrontation of opinions », in *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2017, p. 2235-2241.

[145] E. N. Sarr, O. Sall, et A. Diallo, « SnVera: A New Algorithm for Automation of Fact-Checking in Web Journalism Context », in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018, p. 342-348.

[146] L. Faty, M. Ndiaye, K. Dramé, I. Diop, A. Diédhiou, et O. Sall, « An Architecture for Opinion Mining on Journalistic Comments: Case of the Senegalese Online Press », in *World Conference on Information Systems and Technologies*, 2020, p. 395-403.

[147] L. Faty, M. Ndiaye, I. Diop, et K. Drame, « The complexity of comments from Senegalese online presses face with opinion mining methods », 2019, p. 1-6.

[148] L. Faty *et al.*, « Opinion Scraper: A News Comments Extraction Tool for Opinion Mining », in *2020 3rd International Conference on Big Data and Computational Intelligence (ICBDCI)*, 2020, p. 1-9.

[149] H. Wickham et G. Grolemund, *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc., 2016.

[150] F. M. Hasan, N. UzZaman, et M. Khan, « Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla », in *Advances and innovations in systems, computing sciences and software engineering*, Springer, 2007, p. 121-126.

[151] A. Radziszewski, « A tiered CRF tagger for Polish », in *Intelligent tools for building a scientific information platform*, Springer, 2013, p. 215-230.

[152] L. Faty *et al.*, « SenOpinion: A New Lexicon for Opinion Tagging in Senegalese News Comments », in *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 2020, p. 1-6.

[153] M. Khoule et M. N. Thiam, « Towards the establishment of a LMF-based Wolof language lexicon (Vers la mise en place d'un lexique basé sur LMF pour la langue wolof)[in French] », in *TALN-RECITAL 2014 Workshop TALAf 2014: Traitement Automatique des Langues Africaines (TALAf 2014: African Language Processing)*, 2014, p. 172-177.

[154] M. K. El Hadji Mamadou Nguer, M. N. Thiam, M. B. Thiam, O. Thiare, et M.-T. Cisse, « Dictionnaires wolof en ligne: Etat de l'art et perspectives », 2015.

[155] J. L. Diouf, *Dictionnaire Wolof: wolof-français, français-wolof*. Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA …, 2001.

[156] S.-M. Kim et E. Hovy, « Automatic detection of opinion bearing words and sentences », 2005.

[157] A. Esuli et F. Sebastiani, « Sentiwordnet: A publicly available lexical resource for opinion mining. », in *LREC*, 2006, vol. 6, p. 417-422.

[158] A. Adreevskaia et S. Bergler, « Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses », 2006.

[159] P. D. Turney et M. L. Littman, « Unsupervised learning of semantic orientation from a hundred-billion-word corpus », *ArXiv Prepr. Cs0212012*, 2002.

[160] P. D. Turney et M. L. Littman, « Measuring praise and criticism: Inference of semantic orientation from association », *ACM Trans. Inf. Syst. TOIS*, vol. 21, n° 4, p. 315-346, 2003.

[161] C. Brun, « Detecting opinions using deep syntactic analysis », in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 2011, p. 392-398.

[162] M. Vernier, L. Monceaux, B. Daille, et E. Dubreil, « Catégorisation sémantico-discursive des évaluations exprimées dans la blogosphère », 2009.

[163] L. Faty *et al.*, « SenOpinion: A New Lexicon for Opinion Tagging in Senegalese News Comments », présenté à 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020.

[164] S. Anthony, « De l'étiquetage morpho-syntaxique au super-chunking: Levée d'ambiguïtés à l'aide de méthodes hybrides et de ressources lexicales riches ».

[165] L. Faty *et al.*, « News Comments Modeling for Opinion Mining: The Case of Senegalese Online Press », 2020, p. 1-9.

PUBLICATIONS

PUBLICATIONS TIREES DE LA THESE

- **L. Faty**, M. Ndiaye, E.N. Sarr, et O. Sall « OpinionScraper: A News Comments Extraction Tool for Opinion Mining », *in 4th International Workshop on Sentiment Analysis and Mining of Social Networks (SAMSN)* 2020, (accepted).
- **L. Faty**, M. Ndiaye, E.N. Sarr, et al. « SenOpinion: A New Lexicon for Opinion Tagging in Senegalese News Comments », in 2020 *15th Iberian Conference on Information Systems and Technologies (CISTI)*, 2020, p. 1-6.
    - ***IEEE :*** https://ieeexplore.ieee.org/xpl/conhome/9137058/proceeding
- **L. Faty**, M. Ndiaye, E.N. Sarr, et al. « News Comments Modeling for Opinion Mining: The Case of Senegalese Online Press », in 2020 *International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2020, p. 1–5.
    - IEEE : https://ieeexplore.ieee.org/abstract/document/9155069
- **L. Faty**, M. Ndiaye, K. Dramé, I. Diop, A. Diédhiou, et O. Sall, « An Architecture for Opinion Mining on Journalistic Comments: Case of the Senegalese Online Press », in *World Conference on Information Systems and Technologies (WorldCIST)*, 2020, p. 395–403.
    - Springer : https://link.springer.com/chapter/10.1007/978-3-030-45688-7_41
- **L. Faty**, M. Ndiaye, I. Diop, et K. Drame, « The complexity of comments from Senegalese online presses face with opinion mining methods », in 2019 *14th Iberian Conference on Information Systems and Technologies (CISTI)*, 2019, p. 1–6.
    - IEEE: https://ieeexplore.ieee.org/xpl/conhome/8755443/proceeding

L. Faty et M. Ndiaye**,** « Extraction d'itemsets fréquents à partir des données complexes »**,** in Actes de la *8ieme édition de la ConfereNce sur la Recherche en Informatique et ses applications (CNRIA)* 2018, Article jeune chercheur.

AUTRES PUBLICATIONS

- E. N. Sarr, O. Sall et **L. Faty** « Part-Of-Speech Tagging in French: State-of-the-Art and Obstacles », in 2020 *4th International Workshop on Advances in Natural Language Processing (ANLP)*, (accepted)
- E. N. Sarr, S.N. Mbaye, **L. Faty** et al., « News Articles and Facts Modeling: Linked Data or Linked Fact », in 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2020, p. 1–6.

    - IEEE : https://ieeexplore.ieee.org/abstract/document/9154931
- K. Drame, G. Sambe, I. Diop, et **L. Faty**, « Approche supervisée de calcul de similarité sémantique entre paires de phrases », in Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes, 2020, p. 49–54.
- K. Dramé, I. Diop, **L. Faty**, et B. Ndoye, « Indexation et appariement de documents cliniques avec le modèle vectoriel », DEFT, p. 91, 2019.
    - irit.fr : https://www.irit.fr/pfia2019/taln-recital/
- E. N. Sarr, O. Sall, A. Maiga, **L. Faty**, et R. M. Marone, « Automatic Segmentation and tagging of facts in French for automated fact-checking », in 2018 IEEE International Conference on Big Data (Big Data), 2018, p. 5439–5441.
    - IEEE: https://ieeexplore.ieee.org/document/8622168

ACTIVITÉS SCIENTIFIQUES

- **Du 24 au 27 juin 2020** : Participation et présentation au ***15th Iberian Conference on Information Systems and Technologies (CISTI)*** en ligne.
- **Du 22 au 24 juin 2020** : Participation et présentation au ***Sixth International Conference on Advances in Computing & Communication Engineering (ICCACE)*** en ligne.
- **Du 29 au 30 janvier 2020** : Participation et présentation à la **première édition des journées scientifiques du Laboratoire d'Informatique et d'Ingénierie pour l'Innovation (LI3)** à l'Université Assane Seck de Ziguinchor.
- **Du 01 au 05 Juillet 2019** : Participation à l'école d'été ***Statistique & Sciences des Données pour les jeunes chercheurs de l'Afrique Francophone*** à AIMS (African Institute for Mathematical Sciences) Sénégal Mbour
- **Du 19 au 22 Juin 2019 :** Participation et présentation au ***14th Iberian Conference on Information Systems and Technologies (CISTI)*** à Coimbra/Portugal
- **Du 18 au 21 Avril 2018 :** Participation et présentation à la 8ieme édition de la **ConfereNce sur la Recherche en Informatique et ses applications (CNRIA)** à Ziguinchor/Sénégal
- **Du 22 au 24 mars 2018 :** Participation et présentation aux doctoriales de l'**Ecole Doctorale Sciences, Technologies et Ingénierie (ED-STI)** de l'Université Assane Seck-Ziguinchor
- **Du 30 juillet au 01 Août 2018** : Participation à la formation sur « l'Entreprenariat » à l'Université Assane Seck-Ziguinchor
- **Du 26 au 28 juillet 2018** : Participation à la formation sur « la Recherche documentaire et outils de gestion bibliographique à l'Université Assane Seck-Ziguinchor
- **Du 06 au 08 novembre 2012** : Participation à la formation sur « Google Apps EDU » à l'Université Assane Seck-Ziguinchor