UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES ET INFORMATIQUE
APPLIQUÉES

PAR
ALI HASNAOUI

IDENTIFICATION D'INDICATEURS STRATÉGIQUES DANS LES DOCUMENTS

FÉVRIER 2019

PRÉSENTATION DU JURY

Ce mémoire a été évalué par un jury composé de:

- Ismail Biskri directeur de recherche;

Professeur au département de Mathématiques et Informatique
Université du Québec à Trois-Rivières.

- Boucif Amar Bensaber, évaluateur;

Professeur au département de Mathématiques et Informatique
Université du Québec à Trois-Rivières.

- Mhamed Mesfioui, évaluateur;

Professeur au département de Mathématiques et informatique
Université du Québec à Trois-Rivières.

REMERCIEMENTS

En tout premier lieu, je remercie Dieu, le tout puissant et le miséricordieux, qui m'a donné la patience et la force pour accomplir ce travail de maîtrise.

Tout d'abord, je tiens à remercier mon directeur de recherche, Monsieur Ismail Biskri, pour ses précieux conseils qui ont fait progresser cette étude, son soutien précieux et ses encouragements. Je tiens également à lui exprimer ma reconnaissance pour sa grande disponibilité, sa rigueur scientifique et aussi surtout pour avoir cru en moi.

Mes plus vifs remerciements aux membres du jury, pour l'intérêt qu'ils ont eu vis à vis de mon travail et pour leurs conseils qui m'ont permis d'aller vers l'avant.

Je tiens également à remercier les professeurs du département d'informatique de l'Université du Québec à Trois rivières, pour la qualité de leurs enseignements lors de mes études.

Une grande reconnaissance à mes chers parents que Dieu les protège, sans lesquels je n'aurais jamais pu devenir ce que je suis pour tous les efforts et les sacrifices qu'ils ont fait pour moi et pour toujours.

Un grand merci à mes frères Azer, Maousser et spécialement à mon grand frère Amir pour son soutien, sa compréhension et ses encouragements.

Le mot merci ne suffit pas pour remercier mon cousin Hafedh et sa femme Mélissa qui m'ont apporté leurs soutiens moraux et intellectuels pendant mon parcours de maîtrise et depuis que je suis installé au Québec.

Je remercie également tous mes amis qui m'ont apporté leurs soutiens moraux et mes collègues au Laboratoire de Mathématiques et Informatiques Appliquées (LAMIA), pour tous ces moments passés ensemble à parler de nos travaux de recherche respectifs.

RÉSUMÉ

Au cours de la dernière décennie on assiste à une émergence accrue des technologies d'informations et de communications (TIC). Ces TIC ont donné naissance au Web 2.0 caractérisé et dominé par les médias sociaux.

Ces réseaux sociaux génèrent des volumes énormes de données caractérisées par leurs bruits (symboles, sms, smileys ...). Ces bruits rendent souvent l'analyse manuelle des données générées par des réseaux sociaux très complexe, ce qui entraîne l'utilisation d'un éventail de techniques pour extraire les connaissances utiles à partir de ces ensembles de données.
Par conséquent, les techniques d'exploration de données fournissent aux chercheurs les outils nécessaires pour analyser ces données. Ces techniques utilisent des processus de prétraitement, d'analyse et d'interprétation de données. Malheureusement, ces techniques possèdent aussi des limites telles que, l'aspect multilingue associé aux réseaux sociaux qui demeure une problématique éminemment complexe. De plus, les messages des réseaux sociaux sont caractérisés par les bruits, l'orthographe inhabituelle et la multiplication d'abréviations. Également, ces techniques ne représentent pas les résultats sous un format compréhensible facilement par des gestionnaires en quête d'informations fiables pour mieux décider.
Pour remédier à ces problématiques, nous proposons une approche méthodologique et globale qui permet d'identifier des indicateurs dans des documents multilingues. En effet, notre méthode est une chaîne de traitement qui se déroule en trois phases. La première phase est préalable à l'analyse, la deuxième phase est le processus habituel d'analyse des textes et la dernière phase, est celle de présentation des indicateurs dégagés dans un tableau de bord
L'application de notre approche sur un corpus de texte extrait de la plateforme Web de voyages Tripadvisor nous a permis de détecter les éléments caractérisant le tourisme en Tunisie après la révolution.

Mots clés : Analyse de données textuelles, text mining, classification, traduction automatique, mots hors-vocabulaire, réseaux sociaux, indicateur stratégique, tableau de bord, tropes, Deepl, avis en ligne.

ABSTRACT

Over the past decade, there has been an increased emergence of information and communication technologies (ICTs). These ICTs have given rise to Web 2.0 characterized and dominated by social media.

These social networks generate huge volumes of data characterized by their noise (symbols, sms, smileys...). These noises often make manual analysis of data generated by social networks very complex, resulting in the use of a range of techniques to extract useful knowledge from these data sets.

Therefore, data mining techniques provide researchers with the necessary tools to analyze these data. These techniques use pre-processing, analysis and data interpretation processes. Unfortunately, these techniques also have limitations such as the multilingual aspect associated with social networks, which remains an extremely complex issue. In addition, social network messages are characterized by noise such unusual spelling and multiple abbreviations. Also, these techniques do not represent results in a format that can be easily understood by managers seeking reliable information to make better decisions.

To address these issues, we propose a methodological and global approach that allows indicators to be identified in multilingual documents. Indeed, our method is a processing chain that is based on three phases. The first phase that has been added is named as a prerequisite for the analysis, the second phase is the usual text analysis process and the last phase, which has also been added, is the presentation phase of the indicators identified in a dashboard

The application of our approach on a corpus of text extracted from the Tripadvisor travel web platform allowed us to detect the elements characterizing tourism in Tunisia after the revolution.

Keywords: Text data analysis, text mining, classification, machine translation, off-vocabulary words, social networks, strategic indicator, dashboard, tropes, Deepl, online opinion.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTES DES FIGURES

LISTE DES ALGORITHMES

LISTE DES ABRÉVIATIONS

TIC : Technologies d'Informations et de Communications

TAL : Traitement Automatique de la Langue

TA : Traduction Automatique

NMT : Traduction Automatique Neuronale

TB : Tableaux de Bord

SMS : Short Message Service

HTML : L'HyperText Markup Language

TALN : Traitement Automatique de la Langue Naturelle

API : Interface de Programmation Applicative

BLEU : Bilingual Evaluation Understudy

IA : Intelligence Artificielle

FD : Fouille de Données

AA : Apprentissage Automatique

SVM : Support Vector Machine

K-NN ou KNN : k plus proches voisins

ACP : Analyse en Composantes Principales

CAH : Classification Ascendante Hiérarchique

Introduction

La Tunisie, terre de paix et d'hospitalité est devenue en janvier 2011, date de la révolution populaire, l'épicentre d'une vague de transitions politiques, sociales et économiques. Depuis, le pays vit une période de transformation profonde qui a créé de nouveaux défis et opportunités, en particulier pour l'économie du pays. Plusieurs secteurs, piliers de l'économie, ont subi des conséquences majeures à la suite de la dégradation de la situation sociale et économique causée par les multiples grèves, les revendications surréalistes, les contestations et la désobéissance. Certains secteurs, tels que le tourisme, l'industrie pétrolière et le secteur minier connaissent actuellement une situation critique, voir alarmante.

En effet, en termes de contribution à la croissance économique, la participation du tourisme ne dépasse pas le seuil de 0,1%. Il est clair qu'il s'agit d'un secteur en crise : « 3,3 millions de visiteurs sur les six premiers mois en 2012, une fréquentation en baisse de 18% par rapport à la même période en 2010. » [Camille Lafrance, 3 août 2012].
Les décideurs et les investisseurs tunisiens espèrent remédier à la situation vu l'importance de ce secteur. Pour ce faire, il est opportun de déterminer les caractéristiques et les éléments clés du tourisme tunisien sur lesquels il faut agir pour ramener ce secteur à son niveau avant la révolution. D'où une analyse approfondie des expériences des touristes s'avère utile pour déterminer certains indicateurs sur le tourisme tunisien pouvant contribuer à améliorer la situation.

Pour déterminer ces éléments et ces indicateurs une analyse exploratrice des forums de discussion est élaborée. En effet, l'utilisation de méthodes telles le « text mining » permettent d'analyser des textes et d'en extraire des connaissances favorisant une meilleure prise de décision. Toutefois, comme toute autre méthode informatique, le text mining présente quelques limites méthodologiques.

Les textes provenant des forums et des réseaux sociaux, même riches en opinions et en informations, ne sont pas structurés d'une façon appropriée pour appliquer un traitement

automatisé permettant d'avoir des connaissances concluantes. De plus, les textes des forums et des réseaux sociaux sont souvent multilingues, ce qui représente un défi tout en sachant que les méthodes de « text mining » sont majoritairement unilingues. De plus, la présentation des résultats constitue un grand défi pour les méthodes de « text mining » car les résultats ne sont pas formulés d'une manière à être facilement compréhensible par un décideur. En effet, les résultats sont présentés sous un angle informatique qu'un angle affaires.

Pour remédier à ces trois défis des méthodes de « text mining », nous avons parcouru la littérature et consulté des bibliothèques de logiciels afin de les combiner ensemble dans une démarche unifiée et globale d'analyse d'informations et de présentation de résultats, ce qui constitue l'objectif principal de ce mémoire.

Le présent mémoire est structuré en 8 chapitres comme suit :
Un premier chapitre qui présente l'exploration de la littérature sur le thème principal des méthodes d'analyse des données. Le chapitre 2 présente l'une de ces méthodes, soit le « text mining ». Le chapitre 3 présente les moyens préconisés pour remédier à la limite des textes multilingues, en présentant des moyens de traduction automatique et en comparant quelques outils de traduction parmi les plus utilisés de nos jours (Deepl, Google traducteur et Microsoft traducteur). Le chapitre 4 présente les solutions pour remédier à la problématique de structuration des textes des forums et des réseaux sociaux. Le concept de la classification des documents sera abordé dans ce chapitre. Le chapitre 5 est consacré à la limite de représentation des résultats dans un format compréhensible par un gestionnaire décideur. Ce chapitre présente le concept du tableau de bord qui sera alimenté par des indicateurs stratégiques déduits des résultats d'analyse des données.
Le chapitre 6 représente notre approche globale de traitement des données. Il représente notre démarche méthodologique depuis la sélection des textes des forums et des réseaux sociaux jusqu'à la présentation des indicateurs stratégiques dans un tableau de bord.

Le chapitre 7 représente une illustration pratique de notre démarche méthodologique en sélectionnant des textes dans un forum sur les expériences de voyages en Tunisie afin de déterminer les éléments clés du tourisme tunisien sur lesquels on peut investir pour redonner à ce secteur sa place dans l'économie tunisienne. Le dernier chapitre est réservé à la conclusion. On y présente les éléments synthèses du travail accompli ainsi que les perspectives futures.

Selon Statistiques Canada, 82% des canadiens vivent dans des milieux urbanisés qui offrent un accès facile au Web en 2018. Effectivement, 91% de la population canadienne utilisent internet, soit 33 millions de canadiens [Kabane.ca, Février 2018]. Conséquemment, on assiste de nos jours à un grand volume d'information sur le Web. Avec l'émergence des réseaux sociaux et des forums de discussion, qui font de plus en plus partie de nos vies quotidiennes, l'analyse des comportements humains suscite un intérêt grandissant.

Les réseaux sociaux sont devenus récemment le moyen de communication le plus utilisé sur le Web (Ku et al, 2013). Ils ont révolutionné les modes de communication où une personne se trouve connectée 24 h/24, 7j/7 avec son entourage pour partager des informations, des discussions, des photos et des vidéos.

Quant aux forums de discussion, ce sont des moyens de communication et d'interaction avec d'autres utilisateurs permettant d'échanger des points de vue, des opinions et des expériences de vie. Plus précisément « Les forums de discussion sont des espaces numériques de discussion qui permettent à des utilisateurs de gérer des activités intellectuelles collectives, que ce soient des simples discussions ou des processus complexes de résolution de problèmes ou d'aide à la décision. Les forums de discussion sur l'internet (de type usenet) offrent à celui qui les analyse la possibilité d'observer de nombreux phénomènes intéressants » [Michel Marcoccia, 2001].

Toutefois, malgré la richesse en information de ces deux moyens de communication et de connectivité, le langage utilisé sur ces plateformes est non structuré (des abréviations, textes en mode sms, discussion multilingue). Son analyse devient de plus en plus un fardeau pour les méthodes d'analyse classique de texte. En effet, les méthodes habituelles de

traitement automatique de la langue ont des difficultés d'analyse à cause du bruit et de l'orthographe inhabituelle [Atefeh Farzindar, Mathieu Roche, Mai 2015].

Dans une étude d'analyse des critères de choix des destinations touristiques, Jean et Younes (2016) ont sélectionné des corpus provenant d'un forum de voyages, soit www.tripadvisor.com. Selon Moscarola et Boughzala (2016), les corpus d'avis en ligne qui les intéressent présentent l'avantage d'être abondants, indépendants et spontanés mais, néanmoins non structurés et hétérogènes. La qualité des données peut rendre leur mobilisation très laborieuse moyennant des méthodes traditionnelles de l'analyse de contenu. Ils concluent que l'usage des approches qualitatives traditionnelles est pratiquement impossible. » [Jean Moscarola, Younès Boughzala, juin 2016]. De plus, leur choix de la plateforme internationale Tripadvisor, une des plateformes les plus réputées dans l'industrie touristique, les met face à un défi de langue (multilinguisme). Pour éviter ce problème, ils ont limité leur choix aux avis des touristes français pour avoir uniquement des avis en français. Ainsi, ils ont sélectionné 600 commentaires en ligne de touristes français évaluant des hébergements en Algérie, au Maroc et en Tunisie. Certes, ce choix leur facilite le processus d'analyse de texte, mais ils se privent d'un grand volume de données à cause de la langue, ce qui constitue un frein à la généralisation de leurs résultats et constitue un défi majeur pour leur validité. Ainsi, il serait opportun dans notre approche de diversifier les langues pour garantir des résultats plus généralisés et faciliter leur validité. Pour y arriver, le recours aux traducteurs offerts sur le marché serait un incontournable.

Dans le même sens, Farzindar et Roche (2015) ont confirmé aussi la difficulté de l'application du TAL habituel pour analyser des corpus avec du bruit. Ils ont déduit que « Les méthodes classiques de TAL utilisées dans le contexte des médias sociaux se butent à l'orthographe inhabituelle, au bruit, aux fautes et aux limites de ses fonctionnalités. Certaines techniques de TAL, dont la normalisation, l'expansion morphologique, la sélection améliorée de caractéristiques et la réduction du bruit ont été proposées pour améliorer les performances de classification automatique… » (Farzindar, Roche, 2015).
Pour résoudre le problème des erreurs typographiques, le langage propre au clavardage et la répétition de lettres qui sont devenus un phénomène prédominant dans les nouveaux

modes de communication et les modes d'écriture (forum, facebook, Twitter…), Farzindar et Roche (2015) ont suggéré une adaptation des outils traditionnels pour prendre en compte les nouvelles variations comme la répétition des lettres (par exemple, "merciiiiiiii ou beauuuuucoup, goooood, yessss"), dont la normalisation, l'expansion morphologique, la sélection améliorée de caractéristiques et la réduction du bruit.

Dans la même logique que précédemment, (Meishan et al., 2007) et (Freddy Chongtat Chua, 2013) préconisent de procéder au résumé automatique en extrayant des phrases représentatives à partir des commentaires d'internautes afin de minimiser les bruits et à améliorer les performances de classification. D'autre part, Fernández et al. (2014) se sont basés sur la normalisation de chaque commentaire afin de résoudre le problème des bruits des textes. Ils ont eu recours à la transformation de tous les caractères du texte en minuscules et à l'élimination des caractères répétés (s'il est plus de 3 fois les répétitions suivantes sont supprimées).

Tous ces cas soulignent l'importance d'ajouter une étape préalable à l'analyse qui permet de filtrer et de supprimer les commentaires et autres contenus non pertinents. À noter que cette étape est tout à fait différente de l'étape du prétraitement du processus habituel d'analyse des textes. Cette étape ajoutée permet de faciliter l'analyse et de gagner beaucoup de temps lors du prétraitement des données.

Tel que constaté précédemment, plusieurs auteurs se sont penchés à résoudre la problématique des textes non structurés, des bruits et du multilinguisme. Toutefois, la question de visualisation n'a pas profité de cet intérêt. En effet, plusieurs travaux proposent une visualisation basique, ce qui est clairement insuffisant dans le contexte du text mining. Dans ce sens, Mohamed Dermouche et al (2015) ont proposé « une approche pour visualiser les résultats d'analyse d'opinions, basée sur l'utilisation de termes clés par un "nuage de termes" construit à partir de l'ensemble des termes discriminants responsables de la classification. Chaque terme discriminant est ainsi représenté par une taille proportionnelle à sa fréquence dans le corpus des textes. Ils ont présente également une visualisation temporelle du nuage de termes en utilisant la technique de fisheye. ».

En réponse aux limites des techniques de classification traditionnelles des textes non structurés provenant des forums et des réseaux sociaux (fautes d'orthographes inhabituelles, bruit, textes sous forme des SMSs, multilinguisme), nous proposons une approche qui permet d'analyser des textes multilingues et avec des bruits permettant d'extraire des indicateurs stratégiques et les présenter dans un format lisible et compréhensible par un gestionnaire décideur.

CHAPITRE I : LE TEXT MINING

De nos jours l'analyse et le traitement d'informations textuelles sont devenus un enjeu majeur avec l'explosion du Web : environ 90% des données accessibles est sous forme textuelle (bibliothèques électroniques, pages HTML, forums de discussion, réponses ouvertes à des enquêtes, actualités, formulaires Web, etc). Cependant les tâches d'exploration et de récupération de l'information dans ces réservoirs de connaissances deviennent extrêmement complexes. Ce volume de données représente un défi pour de nombreuses organisations qui souhaitent trouver la méthode leur permettant de collecter, d'étudier et d'exploiter ces données. Face à ce problème, la fouille de texte, ou text mining, sert à faciliter l'extraction des connaissances cachées dans des grands volumes de données. Ce domaine de recherche essaie de mettre à profit la surabondance d'informations textuelles en utilisant des techniques d'informatique linguistique, de data mining, d'apprentissage automatique et de statistiques.

I. Le text mining

1) Définitions

Le Text Mining est un domaine de recherche considéré comme une des disciplines du traitement automatique du langage naturel (TALN). Il permet de traiter un volume important de données textuelles provenant d'internet (Ronen Feldman, 1995).

Plusieurs définitions dans la littérature décrivent le text mining sous différents angles. Fayyad et al. 1996 trouvent que le text mining est un processus non trivial d'extraction d'informations implicites, précédemment inconnues, et potentiellement utiles, à partir de données textuelles non structurées dans de grandes collections de textes. De Lassence (2006) définit le text mining comme un processus automatique d'extraction d'informations à partir de données textuelles permettant d'améliorer les décisions prises par des gestionnaires.

D'un autre côté, Monino et Sedkaoui (2016), précisent que le text mining est une technique permettant d'automatiser le traitement de grands volumes de données pour en extraire des connaissances. Quant à Gardarin (2009), il explique que le text mining est un procédé consistant à synthétiser (classer, structurer, résumer, …) des textes en analysant les relations, les patterns et les règles entre unités textuelles (mots, groupes, phrases, documents).

2) Catégorisation de textes

La catégorisation de texte permet de déterminer une liaison fonctionnelle entre un ensemble de textes et un ensemble de catégories (étiquettes, classes). Cette liaison fonctionnelle est connue sous le nom de modèle de prédiction. Elle est estimée par un apprentissage automatique.

Il est conseillé de posséder un ensemble de textes déjà étiquetés, à partir desquels nous déterminons les paramètres du modèle de prédiction le plus efficace, autrement dit le modèle qui fournit le moins d'erreurs en prédiction.

Certainement, la catégorisation de texte nécessite d'associer une valeur booléenne à chaque paire $(D_j, C_i) \in D \times C$, (avec D est l'ensemble des textes et C est l'ensemble des catégories).

- La valeur V (vrai) est alors associée au couple (D_j , C_i) si le texte dj appartient à la classe
- La valeur F (faux) lui sera associée dans le cas contraire.

L'objectif principal de la catégorisation de texte est de créer une procédure (modèle, classifieur) $\Phi : D \times C \rightarrow \{V, F\}$ qui associe une ou plusieurs étiquettes (catégories) à un document D_j telle que la décision donnée par cette procédure « coïncide le plus possible » avec la fonction $\Phi : D \times C \rightarrow \{V, F\}$, la vraie fonction qui retourne pour chaque vecteur D_j une valeur C_i.

3) Les types de catégorisation de textes

Trois types de catégorisation de textes sont prévus :

- Catégorisation binaire :

Ce type de catégorisation correspond au filtrage, il permet, par exemple, de répondre aux questions suivantes « le document est pertinent ou non ? », « le courriel est un spam ou non? ».

- Catégorisation multi-catégorie disjointes :

C'est le cas le plus général de la catégorisation à n classes. Le système doit affecter 0, 1 ou plusieurs catégories à un même document. Ce type de catégorisation répond par exemple au problème d'affectation automatique des codes CIM[3] aux comptes rendus médicaux.

- Catégorisation multi-catégories :

C'est une catégorisation à n classes mais le document doit être affecté à une seule catégorie. On trouve ce type de catégorisation par exemple dans le routage de courriels.

Le figure ci-dessous donne une vue globale des trois paradigmes de catégorisation de textes

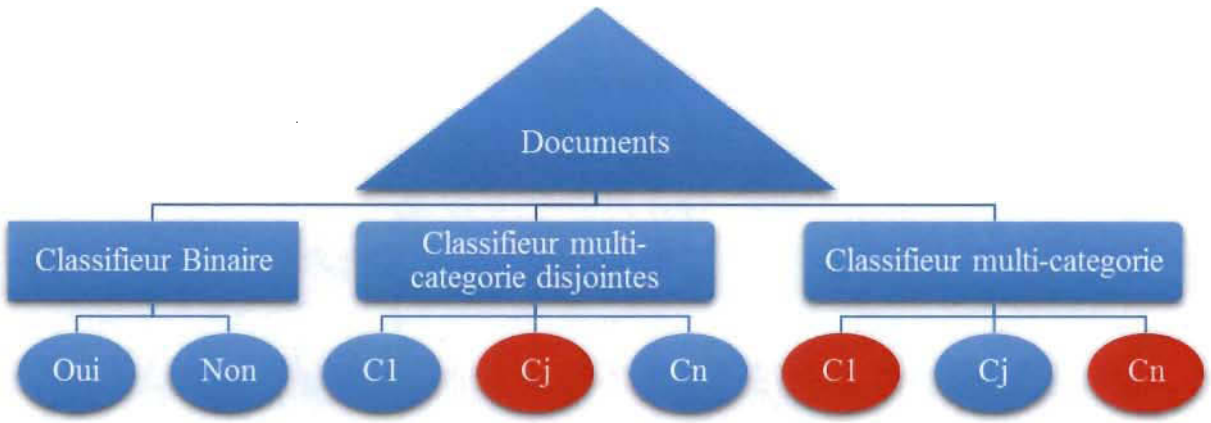

Figure 1: Les trois paradigmes de la catégorisation de textes

4) Le processus du Text Mining

Le processus du Text Mining se fait en cinq étapes :

1) La **sélection** ou la création d'un ensemble de données à étudier ;
2) Le **prétraitement** qui permet d'éliminer le bruit et traiter les données manquantes ;
3) La **transformation** ou la définition des structures optimales de représentation des données ;
4) La **fouille de données** et la détermination de la tâche (classification, recherche de modèles, etc.) en définissant les paramètres appropriés ;
5) L'**interprétation** et l'**évaluation** durant laquelle les éléments extraits sont analysés pour aboutir à des connaissances stockées dans une base de connaissances.

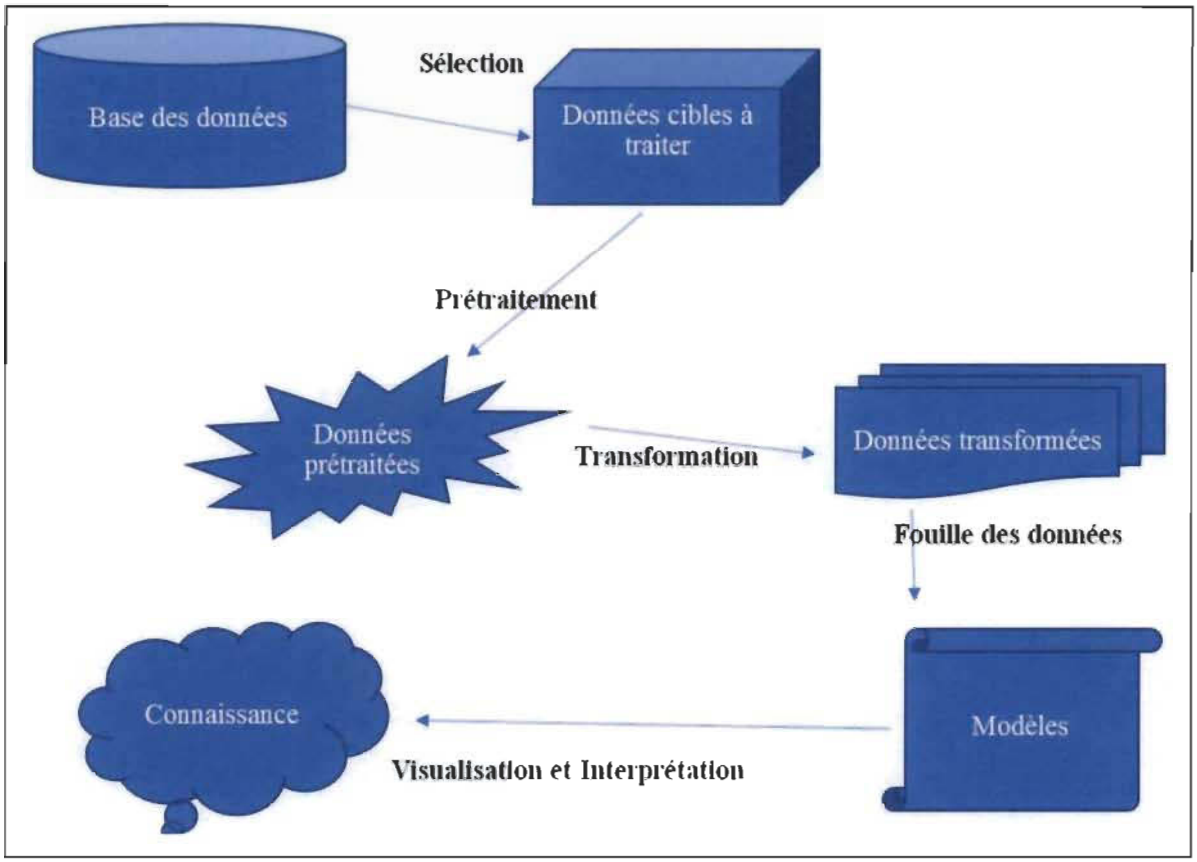

Figure 2: Le processus du Text Mining

5) Cooccurrence de mots

La notion de cooccurrence fait référence au phénomène général par lequel des mots sont susceptibles d'être utilisés dans un même contexte [Chris Manning, Hinrich Schütze, 1999]. Une cooccurrence est confirmée lorsque deux ou de plus de mots (ou autres unités linguistiques) sont présents simultanément dans le même énoncé (Texte, paragraphe, article). Ci-après un exemple très simple d'une cooccurrence : « bicyclette » et « cycliste », deux mots qui, selon toute vraisemblance, sont utilisés la plupart du temps dans un contexte commun, celui de cyclisme. Souvent, si on retrouve un de ces mots dans un texte, on peut prévoir que le deuxième va aussi être présent. Ces deux mots ne sont ni synonymes ni antonymes. Le sens de l'un n'est pas inclus dans l'autre (donc pas d'hyperonymie) et l'un n'est pas une partie de l'autre (donc pas de méronymie). Pourtant, il est évident que ces deux mots partagent quelque chose. On choisit donc de dire qu'ils sont cooccurrents.

D'autres exemples sont très faciles à trouver : « élève » et « maître », « Plage » et « soleil », « cuisinier » et « plat », etc.

6) La complexité des données textuelles

L'application des méthodes d'apprentissage automatique pour traiter les données textuelles est plus compliquée que celle des données numériques. Le langage naturel, contrairement au langage informatique, n'est pas univoque : « un langage univoque se dit d'une correspondance, d'une relation dans laquelle un terme entraine toujours le même corrélatif (aussi biunivoque). » [Lefèvre, 2000].

Le langage naturel est équivoque : il y a plusieurs façons d'exprimer la même idée (la redondance), ce qui est exprimé possède souvent plusieurs interprétations (l'ambiguïté) et tout n'est pas exprimé dans le discours (l'implicite). Ajoutant à ces particularités la grande dimensionnalité des descripteurs, et la subjectivité de la décision prise par les experts qui déterminent la catégorie dans laquelle il faut classer un document. [Bensbaa Abdelaziz, 2018].

a) Grandes dimensions

- Complexité de l'algorithme : Dans la catégorisation de textes, la majorité des algorithmes d'apprentissage sophistiqués sont sensibles au temps, tel que le k plus proches voisins et les arbres de décision. À noter que le temps est considéré comme un paramètre de complexité de l'algorithme, c'est pour cela qu'il est conseillé d'utiliser une méthode de réduction de dimension avant d'estimer les paramètres d'un classifieur.
- Sur-apprentissage : on peut l'appeler aussi sur-ajustement, c'est un phénomène qui apparaît lors de l'exécution d'un algorithme d'apprentissage et plus précisément lorsque celui-ci prend beaucoup de temps pour l'exécution à cause du ralentissement du réseau occasionné par le grand volume de données. Ainsi

b) Déséquilibrage

Dans la pratique, les effectifs des classes sont souvent déséquilibrés et pour certaines classes, le nombre d'exemples positifs est faible en le comparant à celui des exemples négatifs. Ceci crée une difficulté supplémentaire car les classes peu nombreuses sont mal représentées.

c) Ambiguïté

L'ambiguïté, c'est un même mot ou une suite de mots qui peuvent avoir plusieurs significations. Elle existe quatre formes :

- L'ambiguïté syntaxique: On peut le reconnaitre lorsque le syntagme ou bien la structure de la phrase pourrait avoir plusieurs et différentes significations.
- L'ambiguïté référentielle : On le constate lorsqu'il est difficile d'identifier le référent.
- L'ambiguïté lexicale : Il apparait lorsqu'un mot renvoie à plusieurs significations
- L'ambiguïté morphologique : on peut le savoir lorsqu'un mot renvoie à plusieurs catégories.

d) Synonymie

Selon Stephen Ullmann (1975) « On parle des synonymes quand deux ou plusieurs mots différents ont le même sens ». En effet, des mots dits synonymes sont des mots qui appartiennent à la même classe, qui peuvent être remplacés l'un par l'autre et qui ont sensiblement le même sens. Ils sont toujours de même nature. On peut citer comme exemple :

- Le petit garçon **aime** l'hiver.
- Le petit garçon **adore** l'hiver.
- Le petit garçon **apprécie** l'hiver.

II. Les types des données

On estime aujourd'hui que seulement 15 % des données des entreprises sont des données structurées, le reste sont des données non-structurées. Cette abondance des données non-structurées est due à l'époque actuelle où les e-mails, forums de discussion, textes bruts et les fichiers PDF constituent une partie essentielle de l'information collectée par les entreprises. Ces données sont très peu exploitées, elles méritent tout autant d'attention que les données structurées, car nous devons également être en mesure de profiter de ces informations. De façon générale, les données non structurées sont des données textuelles.

1) Les données et leur structure

a) Donnée structurée

Les données structurées sont des informations organisées et classées en vue de faciliter leur lecture et leur traitement. Elles sont organisées aussi de façon à être comprises par des machines. Dans la plupart des cas, ces données sont stockées dans des bases de données relationnelles et affichées en colonnes et lignes, ce qui y facilite l'accès et l'analyse via une recherche aux algorithmes et autres outils d'exploration des données.

b) Donnée semi-structurée

C'est un mélange de données structurées et non structurées. Les données sont structurées sans un modèle de données stricte telles que les données de journaux d'événement.

c) Données non structurées

Il s'agit de données complexes, difficilement exploitables avec les outils classiques, malgré le fait qu'elles renferment une véritable mine d'or d'informations. Ces données non structurées nécessitent des traitements particuliers. On peut les résumer dans le tableau suivant :

Produites par toutes sortes d'acteurs	Existent sous toutes sortes de forme	Correspondent à toutes sortes de support
➢ Entreprise ➢ Consommateur ➢ Internautes ➢ Prospects	➢ Phrases longues et complexes dans une langue soutenue ➢ Phrases courtes avec des fautes d'orthographe ou de grammaire ➢ Multilingue (arabe, français, anglais, chinois…)	➢ Pages de sites internet ou intranet ➢ Articles de blog ➢ Réponses sur les réseaux sociaux (Facebook, twitter...) ➢ Forum de la discussion ➢ Journaux ➢ Documents numérisés online ou offline ➢ E-mails

Tableau 1 : Les données non structurées

2) La différence entre les structures des textes

Le plus grand défi des données non structurées est qu'elles ne sont pas organisées de façon ou dans un format qui permet d'y accéder et de les traiter plus facilement. En effet, très peu de données sont complètement non structurées : même dans des cas où il y a des éléments

considérés comme non structurés, tels que des documents et images, ces derniers sont structurés dans une certaine mesure.

Contrairement aux données non structurées, les données structurées ont été reformatées et leurs éléments sont réorganisés par rapport à une structure permettant à chacun d'être traité, organisé et manipulé selon diverses combinaisons, afin de mieux exploiter les informations.

Les données semi-structurées, quant à elles, représentent une forme intermédiaire (entre les deux formats) : elles ne sont pas organisées d'une façon complexe rendant possible un accès et une analyse sophistiquée. En effet, on peut leur associer certaines informations, telles que des balises de métadonnées, qui permettent d'adresser des éléments qu'elles renferment.

Dans notre cas, nous nous intéressons aux données de type texte, qu'on appelle verbatims, qui sont des commentaires ou des avis laissés par des touristes pour exprimer leur opinion sur leur voyage et expliquer leur satisfaction ou à l'inverse leur insatisfaction. Ces données sont considérées comme étant non structurées et pouvant nuire à la prise de décision.

III. Les méthodes des analyses textuelles

L'analyse d'une langue naturelle subit des opérations communes aux linguistes et aux informaticiens. « Le texte exprime une gamme vaste et riche d'information, mais encode cette information dans une forme qui est difficile à déchiffrer automatiquement. » [Marti A. Hearst, 1999]. Parmi les méthodes d'analyse on peut citer :

1) Analyse lexicale

L'analyse lexicale est pareil à l'étude du vocabulaire d'un discours (richesse, redondances…). Elle sert à réunir les symboles en lexèmes (le lexème est une unité minimale de signification qui appartient au lexique de cette langue, racine du mot) et permet ainsi de mesurer le nombre de mots différents utilisés. L'analyse lexicale se focalise sur les mots plutôt que sur le texte.

D'un point de vue informatique, on peut aussi appeler l'analyse lexicale, tokenization ou segmentation. Il faut faire attention au terme tokenization qui est tout à fait différent du

terme utilisé dans la sécurité informatique. Elle permet de convertir une chaîne de caractères en une liste de symboles.

2) Analyse linguistique

L'apparition de l'approche d'analyse linguistique textuelle remonte aux années 1950. Il était apparu en même temps que l'analyse du discours. Cette discipline rejette tout rapport avec la grammaire du texte.

Les analyses linguistiques ont permis l'apparition de l'établissement de cinq domaines distincts d'étude. Ils sont considérés comme étant les domaines d'analyse traditionnels de la linguistique. Ces domaines sont :

- Sémantique : « raisonner sur le sens des mots et des phrases, et non
- Seulement sur le nombre d'apparition d'un même mot clé. » [Nina Khayat, 16 Janv 2015]
- Phonétique : « La phonétique est l'étude scientifique des sons du langage humain » [Greg Lessard, 24 décembre 2008]
- Phonologie : «la phonologie est d'établir quelles sont les classes de sons qui sont importantes dans la communication pour une langue donnée et d'expliquer la variation entourant ces classes » [Martin Beaudoin, Août 2002].
- Morphologie : « La morphologie est l'étude des formes et des mots. » [Jukka Havu, 2014].
- Syntaxe : « consiste à mettre en évidence la structure d'un texte, généralement une phrase écrite dans une langue naturelle… » [Ekué Wélédji Kpognon, 30 Octobre 2015]

3) Analyse thématique

L'analyse thématique est considérée comme une méthode qualitative de dépouillement utilisée lors d'une analyse de textes. Elle consiste à classer le discours par thème et aussi à calculer leurs fréquences et leurs interactions pour permettre de comprendre l'articulation de la pensée d'un être humain. En effet, « L'analyse thématique est un travail de réduction ou de synthèse sur un corpus dont la taille est toujours variable. La synthèse réduit la taille du corpus à des proportions gérables d'où l'on peut voir simplement exprimer l'essentiel de ce qui a été dit et son importance. Cette réduction est possible grâce à la dénomination [Paillé & Mucchielli, 2003] ou à la catégorisation [Mucchielli, 2010].

On peut définir aussi l'analyse thématique comme étant un processus systématique de repérage, de regroupement et d'examen des propos d'un corpus, son objectif est « la transposition d'un corpus donné en un certain nombre de thèmes représentatifs du contenu analysé, et ce, en rapport avec l'orientation de recherche (la problématique) » [Mucchielli, 2010, p. 124].

En effet Paillé et Mucchielli « font une distinction importante entre rubrique et thèmes, distinction sous laquelle, disent-ils, il existe de nombreuses confusions. La différence entre les deux se situe au niveau du degré de généralité et d'objectivité de la rubrique. Le nom que le thème appose sur des propos est toujours un nom qui renseigne sur l'orientation ou la teneur de ces propos sans les n'interpréter ni les théoriser. La dénomination pour une rubrique assimile cette dernière à un titre de journal. » [Jo Katambwe et al, Automne 2014].

IV. Domaines des applications du Text Mining

On peut appliquer le texte mining dans plusieurs domaines. Elle permet en particulier de :

- Répondre à des questionnements de décideurs sur des questions telles que la détection de fraude ou encore la recherche médicale.
- La compréhension accrue du phénomène détecté lors de l'extraction des connaissances.
- Analyser les performances des produits et services qu'offrent les entreprises à leurs consommateurs. Cela leur permet également de découvrir des informations sur leurs marchés et leurs concurrents en analysant les revues de presse ou d'autres sources pertinentes.
- Analyser les opinions publiées sur Internet (Réseaux sociaux, forum sites de e-commerce, avis en ligne etc.),
- Analyser les sentiments d'individus, soit positifs ou négatifs, exprimés sur différents sujets (satisfaction, insatisfaction…)
- Identifier des groupes de clients qui ont les mêmes caractéristiques et les mêmes intérêts, ou encore, de savoir les besoins d'achat des clients au fil du temps.
- Rediriger automatiquement les demandes précises vers le service approprié, ou de donner des réponses immédiates aux questions les plus souvent posées.

V. Conclusion

Le text mining est une discipline qui tendra à se développer dans l'avenir car l'usage des documents textuels électroniques est devenu populaire et vulgarisé et ce d'une façon croissante. Cette discipline repose sur une diversité de techniques et technologies (traduction automatique, intelligence artificielle, statistiques, théorie de l'information, bases de données, ...) qui requièrent des compétences variées et de haut niveau.

La traduction automatique de texte, qui sera traitée dans notre prochain chapitre, est considérée parmi les solutions innovantes proposant des outils de traitement du document textuel pour faciliter l'accès à l'information.

Chapitre III : TRADUCTION AUTOMATIQUE

Les débuts de la traduction automatique remontent aux années 50. Elle est l'un des outils les plus indispensables de notre époque à cause de la multitude des langues qui existent au monde. « Il y a aujourd'hui plus de 6000 différentes langues parlées dans le monde. Avec 200 pays et plus de 7 milliards de personnes sur notre globe, la traduction est un secteur en constante progression. Le marché de la traduction se situerait à 45 milliards de dollars en 2020 avec une croissance annuelle aux alentours de 6%. Ce marché représente plus de 640000 traducteurs ou interprètes et plus de 18 000 de sociétés. Ce marché en pleine expansion connait une véritable révolution technologique avec l'apparition d'Internet et l'usage de nouvelles technologies. » [Aurelien Deixonne, 6 Mars 2018].

I. Traduction automatique

1) Définition

La traduction automatique (TA) est une traduction qui se fait par ordinateur. Elle consiste à utiliser un logiciel informatique pour transcrire un texte d'une langue naturelle (par exemple l'anglais vers français).

Pour traduire un texte il faut comprendre son sens pour être restitué dans la langue cible. Ce processus est simple d'apparence mais en réalité, il est très compliqué. La traduction n'est pas seulement une simple substitution mot à mot. Le traducteur doit analyser et interpréter le texte et aussi comprendre les relations entre les mots qui peuvent en influencer le sens. Ceci nécessite une compétence ainsi qu'une connaissance de la grammaire, de la syntaxe (structure de la phrase) et de la sémantique (sens des mots), à la fois dans la langue source et aussi dans la langue cible. Dans la même veine de réflexion « La traduction n'est pas un travail sur la langue, sur les mots, c'est un travail sur le message, sur le sens. » [Florence Herbulot, 02 Juin 2004].

Pour avoir une bonne traduction, plusieurs révisions devraient être effectuées dans tous les niveaux. En effet, le défi de la traduction automatique est de produire des traductions comparables où pareilles à des traductions humaines. Dans ce sens, la traduction automatique se définit comme étant un mécanisme permettant de traduire entièrement un texte, à l'aide d'un ou plusieurs systèmes informatiques, sans qu'un traducteur humain n'ait à intervenir dans le processus.

2) L'architecture linguistique d'un système de traduction automatique

L'architecture linguistique d'un système de traduction automatique est presque pareille pour toutes les approches, même les plus récentes. Le « triangle de Vauquois » illustré dans la figure ci-dessous représente les différentes architectures linguistiques possibles d'un système de TA. Chaque chemin dans le triangle correspond à une architecture linguistique :

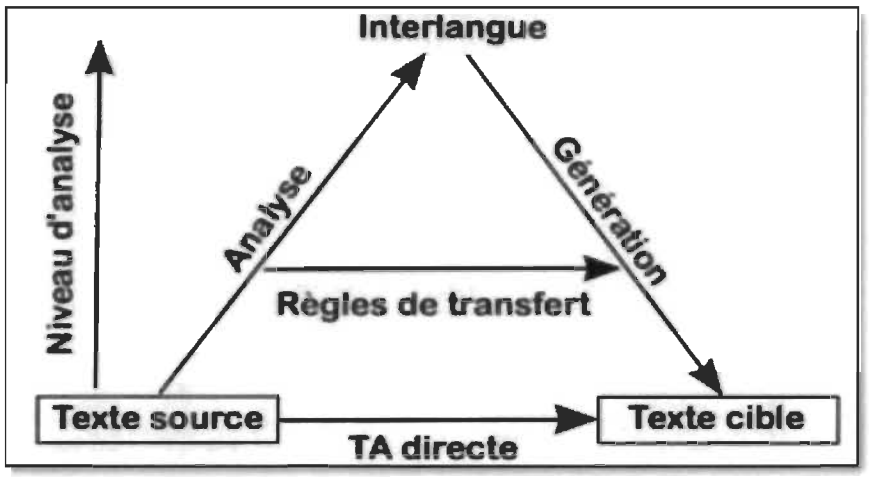

Figure 3:Triangle de Vauquois, représentation des différentes architectures linguistiques (Imane et Al, 2014)

3) Fonctionnement de la traduction automatique

Pour traduire un texte automatiquement, il existe plusieurs outils qui reposent sur des systèmes différents, à savoir :

a) Les systèmes qui reposent sur les règles

Ces systèmes reposent sur des règles associant des règles grammaticales, des règles linguistiques et des dictionnaires de mots courants. Ce système est « Développé en 1980, traduit d'une langue source vers une langue cible en utilisant des règles de transformation entre les deux grammaires : source et cible. Les règles de transformation sont définies manuellement par des linguistes qui sont experts en langue source et cible à la fois ». [Achraf Othman, Mohamed Jemni, 13 mars 2017].

En effet, l'utilisation des dictionnaires spécialisés contenant la terminologie utilisée dans certains secteurs ou disciplines permet d'améliorer la qualité de la traduction de contenus afin d'être précise. Ainsi « la traduction à base de règles fonctionne grâce à l'application de règles à divers niveaux d'analyse linguistique (lexicale, syntaxique et grammaticale). Elle intègre la gestion d'un très grand nombre de cas particuliers et d'exceptions. Les textes

ainsi produits sont cohérents, mais finalement peu adaptés à des éléments trop spécialisés. » [Raphaël Dahl, 09 Juin 2017].

b) Les systèmes basés sur des statistiques

Contrairement aux systèmes précédents, les systèmes basés sur des statistiques n'appliquent aucune règle linguistique pour effectuer la traduction. « La traduction automatique statistique (Koehn,2009) a vu le jour à la fin des années 1980 quand une équipe d'IBM essaie d'appliquer à un problème de traduction des techniques issues de la reconnaissance de la parole ». [Thierry Poibeau, Mars 2016].

Ces systèmes serrent à stocker et à analyser des grandes quantités de données pour chaque paire de langues. Ils peuvent également intégrer des données spécifiques à un secteur ou un domaine précis pour que la traduction de documents spécialisés soit affinée. Dans la même veine de réflexion « La traduction statistique se base quant à elle sur une analyse statistique d'un grand volume d'exemples déjà traduits. Elle identifie les transformations de groupes de mots, d'une langue vers une autre, pour reproduire celles estimées les plus probables sur les nouvelles phrases à traduire. La traduction par ce modèle est adaptée à des contenus spécifiques, mais s'avère peu fluide. » [Raphaël Dahl, 09 Juin 2017].

c) Les systèmes basés sur des algorithmes neuronaux

La traduction automatique qui repose sur des algorithmes neuronaux (NMT), est une nouvelle approche. Elle est considérée comme la révolution dans le domaine de la traduction. Dans un temps réel, elle peut traduire des millions d'informations avec une précision et une fiabilité désormais proche de celle d'un être humain. L'idée principale de ce troisième système est de ne plus agir par mot ou expression pour traduire, elle considère chaque phrase comme un bloc à traduire. Ce n'est pas pareille pour les deux autres modes de traduction qu'on a vus précédemment (Les systèmes qui reposent sur des règles et les systèmes basés sur des statistiques).

Cette approche « s'est bien développée et semble vraiment surpasser tout ce qui a pu se faire jusque-là. Elle est rapidement en train de devenir une technologie de pointe ». [Matthew Carrozo, 26 Octobre 2017]. Elle est de plus en plus intéressante pour les chercheurs et les développeurs du secteur de la traduction automatique, car les systèmes de la traduction automatique neuronale commencent à fournir de meilleures performances en traduction (avec plusieurs paires de langues) que l'approche de traduction automatique basée sur les statistiques.

De plus « pour qu'un système de traduction neuronal fonctionne, il faut d'abord lui faire ingurgiter d'énormes volumes de textes traduits par l'homme. Le système analyse ensuite chaque mot dans son contexte et le classe dans un système abstrait de représentation numérique. Ensuite, chaque fois qu'il retrouve ce mot, il cherche la représentation qui correspond le mieux dans son classement, en fonction de ce qu'il a enregistré auparavant. La traduction neuronale traduit bien mieux les phrases longues que la traduction automatique statistique. » [Tradonline,2 Août 2018].

II. Des exemples des traducteurs automatiques

De nos jours, il existe de nombreuses solutions de traduction en ligne telles que google traducteur, Deepl et Microsoft traducteur. Ceux-ci sont considérés comme les meilleurs outils de traduction actuellement.

1) Les Traducteurs en ligne

a) Google traducteur

Google Translate est un service de traduction automatique multilingue gratuit lancé en 2005. Il est développé par Google, pour traduire des textes. En avril 2006, Google Translate a lancé le service de traduction automatique statistique. Il offre une interface de site Web, des applications mobiles pour Android et iOS, et une API qui aide les développeurs à créer des extensions de navigateur et des applications logicielles. Google Translate prend en charge plus de 100 langues à différents.

Plutôt que de traduire directement les langues, il traduit d'abord le texte en anglais, puis dans la langue cible. Au cours d'une traduction, il recherche des modèles dans des millions de documents pour aider à décider de la meilleure traduction.

En novembre 2016, Google a annoncé que Google Translate passerait à un moteur de traduction automatique neuronale - Google Neural Machine Translation (GNMT) qui traduit des phrases entières à la fois, plutôt que mot par mot. « En utilisant ses propres processeurs spécialement conçus pour l'intelligence artificielle et les réseaux neuronaux multicouches. Selon Google, la même phrase traitée par la méthode LSTM qui prenait 10 secondes à traduire ne nécessite aujourd'hui que 300 millisecondes. » [Marc Zaffagni,30 septembre 2016].

b) DeepL

DeepL Translator est un service de traduction automatique paru récemment en 29 août 2017 par DeepL GmbH. Il est capable de fournir des traductions tout en identifiant les nuances de langage les plus subtiles. Actuellement, le service prend en charge seulement sept grandes langues européennes (Anglais, Français, Allemand, Espagnol, Italien, Polonais et Néerlandais).

La société, allemande, existe en fait depuis 2009 sous le nom de Linguee, qui a été le premier moteur de recherche de traduction sur Internet. C'est un des outils les plus utiles et populaires du Web, « Utilisé par plus de 300 millions d'utilisateurs en 2016, Linguee est sans doute l'un des dictionnaires multilingues en ligne les plus complets. Sa force vient de son fonctionnement, qui consiste essentiellement à proposer des traductions contextualisées plutôt qu'un équivalent mot pour mot. Un service qui s'appuie sur plus d'un milliard de textes traduits par des humains, issus de nombreuses sources. » [Eri,23 octobre 2017]. Il est basé sur l'intelligence artificielle, le service utilise des réseaux neuronaux à convolution construits sur la base de données linguee.

c) Microsoft traducteur

La première version du système de traduction automatique a fait son apparition au début des années 2000 au sein de Microsoft Research. Microsoft traducteur offre également la traduction de texte et de la parole.

Le service prend en charge 60 systèmes linguistiques à partir d'août 2018. Il prend également en charge 10 systèmes de traduction vocale. Dernièrement et en mai 2018, les améliorations continuent, une mise à jour de l'API a été introduite. Cette nouvelle version offrait la traduction automatique neuronale comme méthode de traduction par défaut.

En plus de la traduction, la nouvelle version comprend la translittération et un dictionnaire bilingue pour rechercher des mots afin de trouver des traductions alternatives et de voir des exemples dans les phrases.

2) L'évaluation de la qualité des traductions

Une fois que la traduction est réalisée, il faut évaluer sa qualité. Il existe deux façons d'évaluer, soient :

- Une évaluation manuelle,
- Une évaluation automatique.

a) L'évaluation manuelle

Cette évaluation, nommée aussi « subjective », est réalisée par l'être humain. Elle exige l'intervention des experts bilingues, qui doivent évaluer une très grande quantité de traductions selon des critères de qualité bien précis, tels que la fluidité, la fidélité au sens du texte et les corrections grammaticales.

Ce type d'évaluation est dispendieux puisqu'elle exige un travail manuel fastidieux. Chaque expert doit évaluer une grande quantité de traductions et chaque traduction doit être évaluée par plusieurs experts afin de s'assurer de la fiabilité des résultats, ce qui explique le degré de complexité de l'évaluation manuelle.

b) L'évaluation automatique

L'évaluation automatique est réalisée à l'aide de métriques automatiques. Contrairement à l'évaluation manuelle, elle permet de réaliser des résultats instantanés à faible coût. Les métriques comparent la traduction générée automatiquement avec une traduction de référence, c'est à-dire une traduction réalisée par l'être humain.

Il existe plusieurs métriques mais la métrique la plus utilisée, c'est le BLEU (Bilingual Evaluation Understudy) qui a été proposée par Papineni et al. au début des années 2000. Plus précisément, « BLEU » est une mesure de précision, dont le principe est de calculer le degré de similitude entre une traduction automatique et une ou plusieurs références en se basant sur la précision n-gramme: si une traduction automatique est identique à une des références, alors le score BLEU est égal à 100. Par contre, si aucun des n-grammes de la traduction n'est présent dans aucune référence, alors le score BLEU est égal à 0. » [Souhir Gahbiche-Braham, Octobre 2013] .

3) Comparaison des traducteurs automatiques

Afin de comparer les traducteurs, un texte en anglais a été sélectionné pour être soumis aux services de traduction automatique de Deepl, Microsoft traducteur et Google traduction afin d'avoir une traduction au français

Ainsi, le texte source est un extrait d'un article de Karolak Magdalena, soit :

« The aim of this paper is to analyse the use of social media in the stages of uprising, democratic transition and democratic consolidation using the case study of Tunisia. While the impact of social media in uprisings has been widely documented in past research about the MENA region, Tunisia provides new evidence to the use of Internet in the processes of democratization. » [Karolak Magdalena, 01 Juin 2018].

- **Traduire manuellement**

« L'objectif de cet article est d'analyser l'utilisation des médias sociaux dans les phases de soulèvement, de transition démocratique et de consolidation démocratique en utilisant l'étude de cas de la Tunisie.

Alors que l'impact des media sociaux dans les soulèvements a été largement documenté dans les recherches précédentes sur la région MENA, la Tunisie apporte de nouvelles preuves de l'utilisation d'internet dans les processus de démocratisation. »

- **Traduire avec Deepl**

Figure 4 : Capture d'écran avec Deepl

- **Traduire avec Microsoft traducteur**

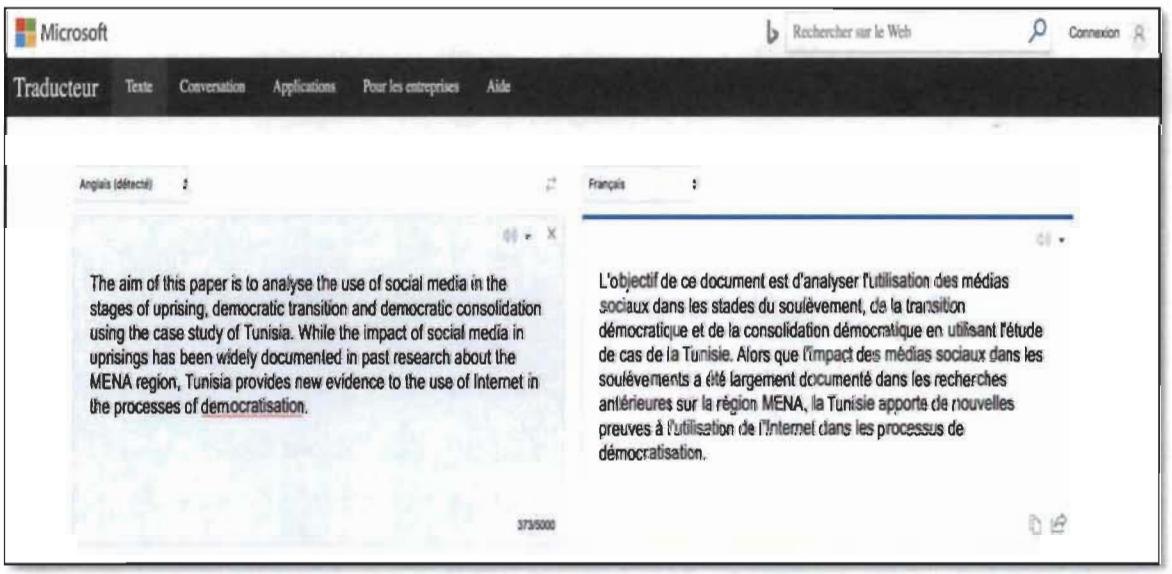

Figure 5 : Capture d'écran avec Microsoft traducteur

➢ **Traduire avec Google traducteur**

Figure 6 : Capture d'écran avec Google

Comparés côte à côte, certaines phrases sont quasiment identiques, mais là où les phrases sont interprétées différemment, celles de DeepL sont souvent plus justes.

- Cet article : le mot « paper » est utilisé dans le contexte d'un article. Le mot document est plus général et vague.
- En utilisant l'étude : le verbe « analyser » dans la phrase fait que le mot « en utilisant » est plus pratique que le mot « à l'aide ».
- Alors que : la traduction de « while » dans cette phrase ne peut qu'être « alors que »
- Les recherches passées: la traduction du mot « past » est strictement le mot « passé », le mot « antérieures » est, à mon avis, utilisé quand deux actions sont comparées dans l'ordre chronologique.
- Apporte : est plus exacte pour garder le sens de la phrase. « Fournir » est lié à une demande (le mot le plus proche en anglais serait « supply ») or que « apporter » est sans conditions.

Déjà cette hypothèse a été confirmée par Michel Courcelles, « Les journalistes du journal le Monde ont réalisé quelques tests de performance non exhaustifs pour montrer les différences entre cinq services de traduction (Deepl, les services de traduction de Google,

Bing, Yandex et Baidu). Deepl performe relativement mieux que ses concurrents. La société propriétaire de Deepl a communiqué qu'elle estime que Deepl est trois fois plus performant que Google translate. » [Michel Courcelles, 31 Août 2017].

De plus, cette hypothèse a été confirmée aussi en utilisant le test BLEU (Bilingual Evaluation Understudy) pour évaluer la qualité des traductions. (Voir figure publiée par les sites www.rtl.fr, www.deepl.com, www.letelegramme.fr ...)

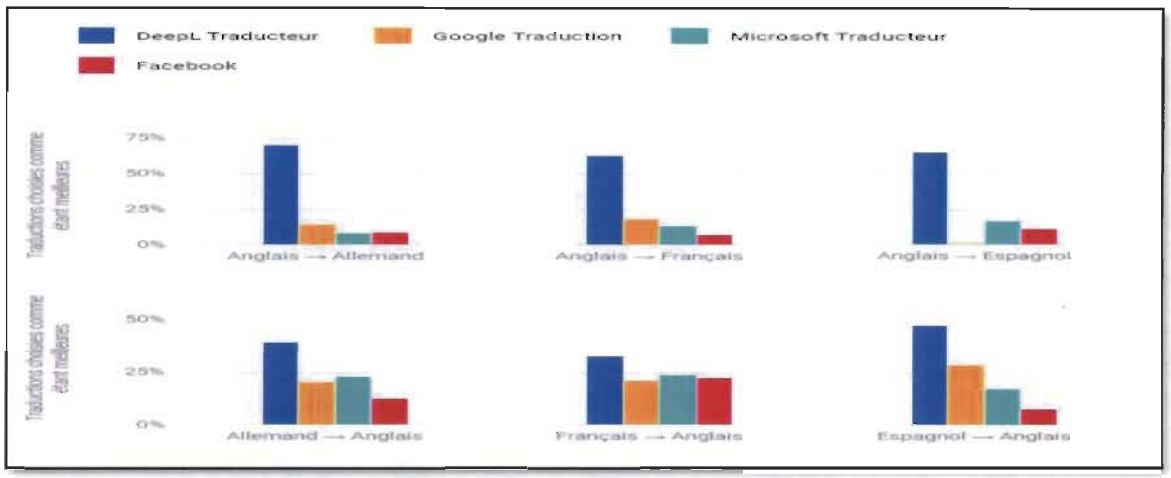

Figure 7 : Test BLEU 100 traductions ont été évaluées par des traducteurs professionnels

Conséquemment, il est évident de conclure que DeepL est extrêmement prometteur et il sera le traducteur utilisé pour la suite de ce travail.

III. Conclusion

La traduction consiste à porter un texte écrit dans une langue naturelle, la langue source, vers une autre langue. Il est intéressant de voir que la TA pourrait permettre aux chercheurs de rehausser leur productivité et la qualité de l'analyse des données. Selon les fonctionnalités qu'offraient les différents outils analysés, nous avons choisi Deelp comme traducteur pour notre travail puisque celui-ci répondait à tous les critères voulus. Il traduit nos propos grâce à des algorithmes d'Intelligence Artificielle et une volumineuse banque de données de texte déjà traduits. Cependant, il est important de rappeler que l'Intelligence Artificielle évolue car elle apprend à chaque nouvelle analyse. Dans notre prochain chapitre, nous présenterons l'apprentissage automatique et la classification des données

CHAPITRE IV: L'APPRENTISSAGE ET LA CLASSIFICATION

L'apprentissage artificiel (machine learning en anglais), est un ensemble de méthodes permettant d'établir, à partir de données, des modèles de prise de décision, de prédiction ou de classification.

Tel qu'illustré dans la figure ci-dessous, la construction de modèles se basent sur plusieurs techniques issues de l'intelligence artificielle (IA) et la fouille de données (FD).

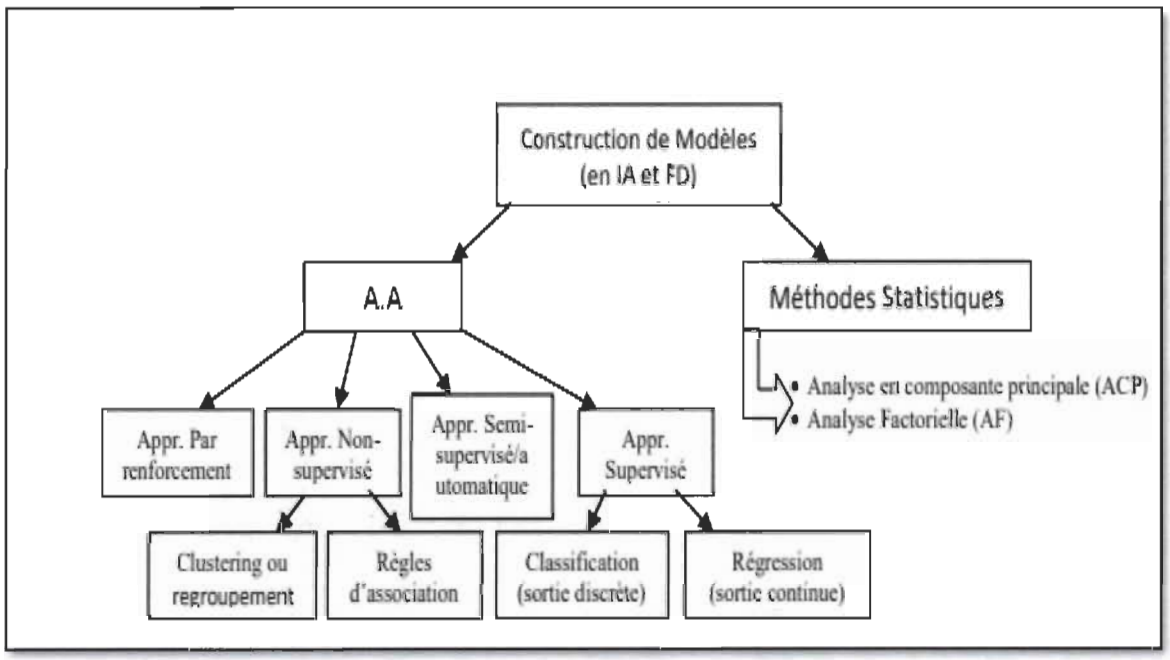

Figure 8 : Les différentes techniques issues de l'IA et FD pour la construction de modèles de données (Mokhtar Taffar, 2013)

Ces techniques sont des techniques d'apprentissage automatique (AA) ou de méthodes statistiques. Dans notre travail, nous nous intéressons spécifiquement aux techniques du AA soient l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage semi-supervise, l'apprentissage par transfert, l'apprentissage par renforcement et la classification.

I. La classification des données

1) Définition de la classification des données

La classification (clustering) est une méthode mathématique d'analyse de données qui consiste à attribuer une classe ou catégorie à chaque objet (ou individu) à classer, en se basant sur des données statistiques. Plus précisément « La classification des données est un processus de catégorisation cohérente des données sur la base de critères spécifiques et

prédéfinis, afin qu'elles puissent être utilisées et protégées plus efficacement. » [Vincent Dely, juin 2018].

Le champ d'application de la classification est très vaste. Elle traite une multitude de problème dans plusieurs domaines, entre autres :

- Analyse d'ADN dans le domaine Bio-Informatique
- Analyses spatiales des données
- La recherche documentaire (text mining)

2) Les étapes de la classification des données

La classification des données s'effectue sur trois étapes :

- Le choix des données :

  Le choix des données varie d'un sujet à autre et il dépend de leur disponibilité.

- Le choix d'un algorithme de classification et l'exécution :

  Le choix des méthodes tient compte de la pertinence et la convenance de l'algorithme. L'algorithme devrait être correct, mais aussi efficace, c'est-à-dire : rapide (en termes de temps d'exécution) et économe en ressources. Également, d'autres critères sont pris en compte, soit la représentation des résultats. En effet, ils doivent être sous une forme permettant une meilleure compréhension.

- L'interprétation des résultats obtenus :

  C'est l'étape de l'évaluation de la qualité de classification ainsi la description des classes obtenues.

Le résultat d'une classification prend traditionnellement soit la forme d'un graphique de points ou d'un dendogramme. La figue ci-dessous présente un aperçu général sur les résultats d'une classification :

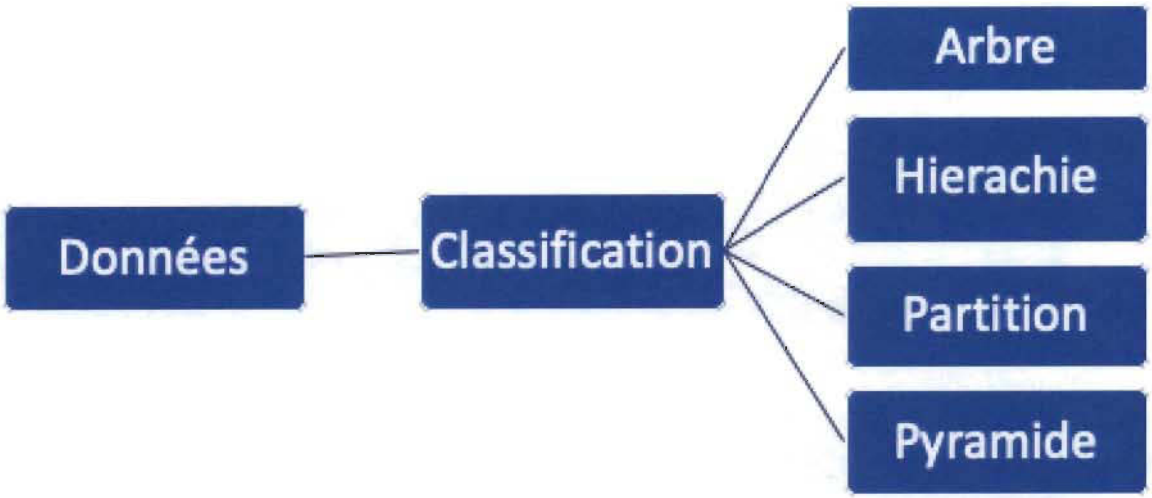

Figure 9 : Un aperçu des résultats d'une classification

II. L'apprentissage automatique

1) Définition

L'apprentissage automatique est le fait d'apprendre un ensemble de relations entre les critères caractérisant l'élément à classer et sa classe cible. Dans ce sens il permet de comprendre la structure des données et de les intégrer dans des modèles qui peuvent être compris et employés par tout le monde.

En 1959 Arthur Samuel a défini l'apprentissage automatique comme étant « ...la discipline donnant aux ordinateurs la capacité d'apprendre, sans qu'ils soient explicitement programmés » [Frédéric Camps, juin 2018]. Selon [Jean-Francis Roy, 2018] : « L'apprentissage automatique est une science qui consiste à développer des algorithmes d'apprentissage, qui apprennent à résoudre une tâche. ». Plus précisément, [L'UC Berkeley] a défini l'apprentissage automatique comme étant « ... la branche de l'intelligence artificielle (IA) qui explore les moyens d'amener les ordinateurs à améliorer leurs performances en fonction de leur expérience. »

2) Les domaines de l'application

Vue l'importance de l'apprentissage automatique dans la vie de l'être humain, il a été utilisé dans plusieurs secteur tel que :

- La technologie de reconnaissance faciale qui permet aux plateformes de médias sociaux d'aider les utilisateurs à marquer et partager des photos d'amis, ainsi qu'aux moteurs de recherche de créer des solutions agréablement personnalisées pour les consommateurs en suggérant des films ou des émissions de télévision selon les préférences de l'utilisateur.
- Dans les grandes surfaces en disséquant les données collectées en magasin sur le comportement des consommateurs. Ainsi, ils peuvent réorganiser leurs rayons pour booster leurs ventes.
- Les voitures autonomes qui utilisent l'apprentissage automatique permet d'identifier en temps réel, les éventuels obstacles qui se présentent sur une route.

III. Les méthodes d'apprentissage automatique

Il existe différents types d'apprentissage mais les deux modes les plus utilisées sont l'apprentissage supervisé et l'apprentissage non supervisé. L'apprentissage supervisé se base sur des algorithmes alimentés par des données d'entrée et de sortie étiquetées par l'homme. L'apprentissage non supervisé ne fournit pas à l'algorithme des données étiquetées pour lui permettre de trouver une structure et de découvrir une logique dans données entrées.

1) L'apprentissage supervisé

Le but principal de la classification supervisée est de définir des règles qui permettent de classer des objets dans des classes à partir de variables qualitatives ou quantitatives qui les caractérisent. Il existe de nombreux algorithmes et techniques utilisés pour la classification supervisée, tels que :

a) La méthode de Boosting

- Définition

La méthode de Boosting permet de créer un ensemble de classifieurs et de fusionner leurs décisions pour réaliser la classification. Les classifieurs sont traités par une méthode pareille d'apprentissage de façon séquentielle de manière que les classifieurs précédents sont considérés comme exemple pour accroitre la performance des classifieurs suivants. Les algorithmes Boosting comme AdaBoost ont donné des réponses très efficaces pour la tâche de classification de textes.

➢ Avantages

- Très bonne performance en pratique,
- Un seul paramètre à régler (le nombre T d'itérations),
- Simple et aussi facile à programmer.

➢ Limites

- Difficile d'incorporer des connaissances à priori,
- Difficile de savoir comment régulariser,
- Les frontières de décision en utilisant des méthodes parallèles aux axes sont souvent très irrégulières (non interprétables).

b) Machine à vecteurs de support

➢ Définition

Les machines à vecteurs de support, « support vector machine (SVM) » en anglais, sont destinées à résoudre des problèmes de classification. Les SVM visent une séparation linéaire entre les groupes dans un espace étendu par rapport à l'espace où les descripteurs sont définis.

L'objectif général des machines à vecteurs de support est la construction d'une fonction f qu'a un vecteur d'entrée x fait correspondre une valeur y.

$$f(x) = y$$

Avec x = l'individu à classer

Et y = la classe à laquelle correspond l'individu en entrée.

Comme tout autre méthode de classification, elle fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle.

➢ Avantages

- Très bonne performance en pratique,
- Traitement des données à grandes dimensions,
- Il y a peu de paramètres à régler.

➢ Limites

- Demande des données négatives et positives en même temps,
- Problèmes de stabilité des calculs dans la résolution de certains programmes quadratiques à contraintes.

c) Réseau de neurones

➢ Définition :

Les réseaux de neurones sont particulièrement bien adaptés pour résoudre des tâches d'apprentissage complexe telles que la reconnaissance de formes afin d'identifier et de classer des objets ou des signaux dans des systèmes de parole, de vision et de contrôle. Ils peuvent également être utilisés pour prévoir des événements futurs et des modélisations de séries chronologiques.

Plus précisément « Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau. » [Claude Touzet, Juin 2016].

➢ Avantages

- Ils possèdent une grande capacité et efficacité de classification,
- Ils peuvent travailler sur des données incomplètes ou bruitées,
- La possibilité de représenter n'importe quelle fonction, linéaire ou pas, simple ou complexe
- Bonne performance en pratique.

➢ Limites

- Toutes les valeurs des variables doivent être encodées d'une façon standardisée en prenant des valeurs entre 0 et 1, et cela pour les variables catégorielles,
- Incompréhensibilité du modèle (le réseau de neurones fonctionne comme une boite noire).

d) Méthode des k plus proches voisins

➢ Définition

La classification K-plus proche voisin (kNN) est l'une des méthodes de classification les plus utilisées grâce à sa simplicité. Elle a été développée pour répondre au besoin d'effectuer une analyse discriminante lorsque des estimations paramétriques fiables des densités de probabilité sont inconnues ou difficiles à déterminer. Elle est basée sur la distance, dans lesquels l'ensemble d'apprentissage est mémorisé, de façon qu'une classification pour un nouvel enregistrement non classé puisse être trouvé simplement en le comparant aux enregistrements les plus similaires de l'ensemble d'apprentissage.

➢ Avantages

- Robuste par rapport au bruit,
- Ne requiert aucune phase d'entrainement (tout le travail se fait au moment de classification),
- Conception très simple et facile à implémenter.

➢ Limites

- Coût de classification élevée
- Les paramètres de de l'algorithme doivent être choisis avec discernement (nbre de voisins K et la taille de voisinage)

e) Arbre de décision

➢ Définition

Une méthode attractive de classification induit la construction d'un arabe de décision. Son objectif est de créer un modèle qui prédit la valeur d'une variable cible en fonction de plusieurs variables d'entrée. Elle est constituée d'un ensemble de nœuds de décision connectés par des branches, s'étendant vers le bas, du nœud racine jusqu'à des nœuds feuilles. Les variables sont testées dans les nœuds de décision, et chaque résultat est représenté en nœud feuille.

➢ Avantages

- Compréhensibilité du modèle,
- La préparation des données est facile (pas de normalisation, de valeurs vides à supprimer, ou de variable muette),
- Elle est très économique en termes de ressources de calcul,
- Un arbre de décision peut être lu et interprété facilement et aussi directement.

➢ Limites

- Instabilité de l'algorithme, en effet à la suite à une petite perturbation des données, l'arbre produit peut-être très différent.
- Une détection difficile des interactions entre les variables.

f) Classification naïve bayésienne

➢ Définition :

Les classificateurs Naïve Bayes sont très évolutifs et nécessitent un certain nombre de paramètres linéaires dans le nombre de variables (caractéristiques / prédicteurs) d'un

problème d'apprentissage. Il s'agit d'une méthode de classification statistique. Ils sont particulièrement utiles pour classifier un ensemble d'observations selon des règles définies par l'algorithme lui-même.

Ils se basent sur le théorème de Bayes fonder sur les probabilités conditionnelles telles que la probabilité qu'un événement se produit sachant qu'un autre événement s'est déjà produit.

➢ Avantages

- Conception très simple,
- Les calculs de probabilité ne sont pas très coûteux,
- La possibilité de la classification même avec un petit jeu de données,

➢ Limites

- L'algorithme Naive Bayes Classifier suppose l'indépendance des variables. C'est une hypothèse forte et qu'est violée dans la majorité des cas réels.

2) L'apprentissage non-supervisé

L'objectif principal de l'apprentissage non-supervisé est de détecter les similarités à partir de jeux de données composés de données d'entrée non libellées. Il existe plusieurs algorithmes et techniques utilisés pour la classification non supervisée, on peut nommer par exemple :

a) Analyse en composantes principales

➢ Définition

L'analyse en composantes principales (ACP) est une procédure statistique qui utilise une transformation orthogonale pour convertir un ensemble d'observations de variables éventuellement corrélées (des entités prenant chacune différentes valeurs numériques) en un ensemble de valeurs de variables linéairement non corrélées, appelées composantes principales. Elle se base sur le calcul des moyennes, variances et coefficients de corrélation. « L'analyse en composantes principales (ACP) est un outil extrêmement puissant de synthèse de l'information, très utile lorsque l'on est en présence d'une somme importante de données quantitatives à traiter et interpréter. L'apparition au cours des dernières années de logiciels chaque fois plus performants et faciles à utiliser rend aujourd'hui accessible ce type d'analyses des données » [Marc Guerrien, 2003].

➢ Avantages

- La Simplicité de ces résultats grâce aux graphiques qu'elle fournit. En effet, l'utilisateur peut appréhender une grande partie des résultats d'un simple coup d'œil,
- Une souplesse d'utilisation qui s'explique essentiellement par la diversité des applications de l'ACP,
- Elle est très puissante puisqu'elle fournit en quelques opérations, un aperçu et une vue complète des relations existantes entre les variables quantitatives.

➢ Limites

- Tant que L'ACP est une méthode de projection, la perte d'information résultant par la projection peut entraîner des fausses interprétations.

b) Carte auto-adaptative

➢ Définition

Les cartes auto adaptatives ou encore cartes de Kohonen, le statisticien ayant introduit le concept en 1984, forment une classe de réseau de neurones artificiels. La capacité à s'auto-organiser offre de nouvelles possibilités d'adaptation aux données d'entrée

« Un réseau de Kohonen est constitué d'une couche d'entrée de N neurones connectés aux M neurones d'une couche de sortie elle-même interconnectée. Soit WjT = {wi,j} le vecteur des poids des N connexions reliant la couche d'entrée au neurone j de la couche de sortie. Soit X un vecteur d'entrées de N composantes. » [Michel Bret, juillet 2018]

➢ Avantages

- Une visualisation graphique pour les résultats obtenus,
- On peut présenter les données dans plusieurs dimensions.
- L'algorithme de Kohonen exploite des relations de voisinage dans la grille pour réaliser une discrétisation dans un temps très court,

➢ Limites

- Un énorme temps pour la convergence.

c) Des k-moyennes

➢ Définition

Le clustering K-means est un simple algorithme d'apprentissage non supervisé utilisé pour résoudre les problèmes de clustering. La procédure suit un moyen simple et facile à classer un ensemble de données en un certain nombre de grappes, définies par la lettre "k", qui est

fixée au préalable. Les groupes sont ensuite positionnés en tant que points et toutes les observations où tous les points de données sont associés au groupe le plus proche, calculés, ajustés, puis le processus recommence en utilisant les nouveaux ajustements jusqu'à ce qu'on atteigne un résultat souhaité.

Plus précisément « La méthode des k-moyennes est une méthode de classification qui permet de mettre au jour une éventuelle structure de groupes dans un ensemble de données. Cette méthode n'est pas récente (les premiers articles traitant d'aspects théoriques remontent aux années cinquante, avec notamment les travaux de Cox, 1957, et de Fisher, 1958) » [Christel Ruwet,2012].

➢ Avantages

- C'est une méthode simple, robuste et facile à comprendre,
- Elle permet d'avoir rapidement un premier résultat,
- On peut l'appliquer à des données de grandes tailles, et aussi à plusieurs types de données.

➢ Limites

- Les clusters dépendent de l'initialisation et de la distance choisie,
- Le nombre de classe doit être fixé au départ,
- Elle ne peut pas détecter les données bruitées.

d) Regroupement hiérarchique

➢ Définition

La Classification Ascendante Hiérarchique (CAH) est une méthode qui consiste à mettre en évidence un regroupement naturel d'un ensemble d'individus décrits par des caractéristiques (les variables). Elle propose une série de partitions emboîtées représentées sous forme d'arbres appelés dendrogrammes.

Les stratégies de regroupement hiérarchique se divisent généralement en deux types:

➢ L'approche ascendante : chaque observation commence dans son propre groupe et des paires de groupes sont fusionnées au fur et à mesure que l'on monte dans la hiérarchie.

➢ L'approche descendante : toutes les observations commencent dans une grappe et les séparations sont effectuées de manière récursive au fur et à mesure que l'on descend dans la hiérarchie.

➢ Avantages

- Une représentation sous forme d'arbre qui met en évidence une information supplémentaire,
- On peut déterminer le nombre optimal de classes,
- Le principe est facilement compréhensible sans être forcément statisticien.

➢ Limites

- Complexité algorithmique (non linéaire),
- A chaque étape, le critère de partitionnement n'est pas global mais dépend des classes déjà obtenues,
- Coûteux en temps de calcul.

3) L'apprentissage semi-supervisé

L'apprentissage semi-supervisé se situe entre l'apprentissage non supervisé (sans données de formation étiquetées) et l'apprentissage supervisé (avec des données de formation entièrement étiquetées). De nombreux chercheurs en apprentissage automatique ont constaté que la combinaison de ces deux méthodes sert à améliorer significativement la qualité de l'apprentissage car les données non étiquetées, lorsqu'elles sont utilisées avec une petite quantité de données étiquetées, peuvent améliorer considérablement la précision de l'apprentissage.

Une autre utilité provient du fait que l'étiquetage de données exige l'intervention de l'être humain et lorsque les jeux de données deviennent très énormes, cette opération peut s'avérer ennuyante. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident.

Il existe deux types d'apprentissage semi-supervisés le Transductive et l'inductif :

- ➢ Transductif : Il offre le label seulement pour les données disponibles non-labellisées.
- ➢ Inductif : Il n'offre pas uniquement des labels pour données non labellisées, il produit aussi un classifieur.

4) L'apprentissage par transfert

Lorien Pratt (1993) a formulé l'algorithme de transfert basé sur la discrimination (DBT).

« L'apprentissage par transfert, consiste à utiliser un jeu de tâches pour influencer l'apprentissage et améliorer les performances sur une autre tâche. Cependant,

l'apprentissage par transfert peut en réalité gêner les performances si les tâches sont trop dissemblables. Un défi pour l'apprentissage par transfert est donc de développer des approches qui détectent et évitent le transfert négatif des connaissances. » [Ievgen Redko, Younès Bennani, février 2018].

5) L'apprentissage par renforcement

Le but principal de l'apprentissage par renforcement, c'est d'entraîner un agent à se comporter de façon intelligente dans un environnement donné. Un agent interagit avec l'environnement en choisissant, à chaque temps donné, d'exécuter une action parmi un ensemble d'actions permises. Le comportement intelligent que doit apprendre cet agent est donné implicitement via un signal de renforcement qui, après chaque décision de l'agent, indique s'il a bien ou mal agi et l'agent a comme entrée un ensemble d'indicateur ou de caractéristique décrivant l'environnement.

Par conséquent « La méthode d'apprentissage par renforcement est considérée comme la plus faible des techniques d'apprentissage automatique, puisqu'elle est très lente et demande de faire tourner les options des millions de fois. » [In Principio, 2017].

IV. Comparaison des techniques d'apprentissage

L'apprentissage supervisé et l'apprentissage non supervisé sont deux approches différentes. Ces deux approches serrent à améliorer l'automatisation et l'intelligence artificielle.

La différence entre ces deux apprentissages consiste au fait que l'apprentissage non-supervisé cherche à trouver des partitions de modèles par lui-même. C'est la machine qui sert à apporter de meilleures performances sans l'intervention de l'être humain. De plus, l'extraction des données est descriptive.

Contrairement à l'apprentissage supervisé, l'apprentissage non-supervisé demande une intervention humaine pour une meilleure automatisation. Il est utilisé quand l'utilisateur sait labelliser les informations. L'extraction des données est prédictive.

« … Les exemples d'apprentissage sont alors constitués de données et du résultat attendu. Par exemple, pour créer un système d'identification vocale, les données sont des échantillons de voix, et le résultat attendu sont leurs propriétaires (…) dans le cadre d'une approche non supervisée, les exemples sont constitués uniquement de données, sans

résultat attendu, et l'apprentissage se fait donc par similarité entre elles. » [Hubert Wassner, 2017].

La différence entre les différents types d'apprentissage se résume comme suit :

Apprentissage Supervisé	Apprentissage non- supervisé
Le nombre de classes est déjà connues	Le nombre de classe n'est pas connues
L'extraction des données est descriptive	L'extraction des données est prédictive
Utilisé pour classifier des données futures	Utilisé pour mieux comprendre et explorer les données
L'être humain qui réagit	La machine qui réagit

Tableau 2 : La différence entre les deux approches

V. Choix d'une technique d'apprentissage

La résolution des problèmes de classification est très délicate et décisive dans de nombreuses applications comme la classification de texte.

C'est pour cela « pour chaque contexte il existe un classifieur optimal selon le critère du taux d'erreur, mais aucun n'est optimal dans tous les cas. Comme il existe de nombreux classifieurs, l'utilisateur préférera souvent choisir un classifieur généraliste, dont l'ajustement et l'exploitation sont à sa portée, en espérant que celui-ci fait presque aussi bien que l'optimal. » [Gilles R. Ducharme, 21 Jun 2018]

Parmi les critères les plus importants sur lesquels nous nous basons dans le choix du classifieur, c'est la rapidité et la simplicité.

Souvent, le choix du classifieur se fait en fonction des résultats qu'on souhaite obtenir. Par exemple, si notre but est de donner une explication ou une justification qui sera ensuite présentée à un décideur ou un expert, alors on préférera les méthodes qui produisent des modèles compréhensibles tels que les arbres de décision ou les classifieurs à base de règles.

VI. Conclusion

Ce chapitre, a permis de faire le tour de la classification des données et des différents algorithmes d'apprentissage automatique.

Nous avons présenté des cas d'utilisation de l'apprentissage automatique, des méthodes courantes et des approches populaires utilisées sur le terrain, des langages de programmation appropriés pour l'apprentissage automatique.

Nous avons pu voir comment fonctionne l'apprentissage automatique ainsi que ses qualités et ses limites. De plus, nous avons présenté la différence entre les différentes méthodes d'apprentissage.

Nous avons remarqué qu'il n'y a pas une méthode d'apprentissage mieux que les autres. L'efficacité de l'apprentissage ou encore, de sa méthode d'application dépend essentiellement de son utilisation et du type de traitement que l'on veut gérer.

Toutefois, la représentation des résultats demeure une limite car le format des résultats demeure difficile et méconnaissable pour un gestionnaire décideur.

En effet, notre prochain chapitre présente le concept du tableau de bord qui sera alimenté par des indicateurs stratégiques déduits des résultats d'analyse des données pour résoudre ce problème et bien présenter les résultats de l'apprentissage d'une manière s'adressant à des responsables, des décideurs et des cadres d'affaires.

CHAPUTRE V : INDICATEURS STRATEGIQUES ET TABLEAU DE BORD

Aujourd'hui, on est confronté à une croissance exponentielle des données. L'analyse de ces données génèrent des résultats qui sont difficiles à comprendre parfois. Pour faciliter leur compréhension, ces résultats sont traduits en indicateurs. Le choix et la présentation de ces indicateurs représente un défi pour les gestionnaires. Ce défi est résolu grâce au recours à un outil de gestion qui s'avère simple mais qui offre une représentation et une visibilité accrue des résultats, c'est le tableau de bord.

Nous présentons dans ce qui suit deux sections présentant les indicateurs stratégiques et le tableau de bord.

I. Indicateurs stratégiques

1) Définition

L'indicateur stratégique est une information ou une mesure, il permet d'expliquer une situation évolutive, une action ou les conséquences d'une action. Il est construit à partir des données disponibles dans les bases de données. Il existe un grand nombre de définitions, à savoir :

- « Un indicateur est un élément ou un ensemble d'éléments d'information représentative par rapport à une préoccupation ou un objectif, résultant de la mesure tangible ou de l'observation d'un état, de la manifestation d'un problème, d'une réalisation » (Pierre Voyer, 1999).
- Un indicateur est « toute mesure significative, relative ou non, utilisée pour apprécier les résultats obtenus, l'utilisation des ressources, l'état d'avancement des travaux ou le contexte externe» (Conseil du trésor du Québec, 2002).
- Pour [Salma Bougar et al, Mars 2018] « Un indicateur stratégique, par définition, identifie la nature de l'information nécessaire pour contrôler la réalisation des objectifs stratégiques afin d'évaluer le degré de la performance recherchée. L'indicateur est un indice par rapport auquel il est vérifié le degré d'atteinte d'un objectif ».
- Pour que [Aurélien Boutaud, 30 novembre 2015] affirme qu'« il existe de nombreuses définitions de la notion d'indicateur. Toutes convergent plus ou moins autour de l'idée qu'un indicateur est la traduction d'un concept ou d'un phénomène sous la forme d'un signal (par exemple un code couleur) ou plus souvent encore d'un chiffre. Cette « traduction » a la plupart du temps pour but :

- De simplifier une information (parfois complexe) pour la rendre compréhensible et utilisable par un public cible (gestionnaires, décideurs, grand public...) ;
- De décrire une situation à un moment et un endroit donné puis, par réplication, de permettre des comparaisons dans le temps et/ou dans l'espace. »

2) Rôle des indicateurs stratégiques

Les indicateurs sont également des outils de communication qui servent à simplifier l'information souvent sous une forme quantifiée pour la rendre plus lisible et signifiante auprès du phénomène cible. Ils ont pour objectif de :

- Suivre un phénomène ou une action,
- Faciliter la communication par un langage et référentiel commun,
- Évaluer un programme,
- Aider à la décision,
- Décrire un élément d'une situation ou une évolution d'un point de vue quantitatif,
- Permettre de synthétiser une grande quantité d'informations et la réduire à quelques éléments clés stratégiques tout en conservant les informations.

3) Les types des indicateurs stratégiques

Il existe deux types d'indicateurs stratégiques, soit qualitatif ou quantitatif. Ils sont mesurés par :

- Quantité (nombre, pourcentage, volume, taux…)
- Qualité (valeur, niveau, cote, degré…)
- Montant (coûts, frais, montants…)
- Temps (délai moyen, nombre de jours…)

a) Indicateur stratégique qualitatif

Un indicateur stratégique qualitatif est « un constat, une indication, une appréciation d'une situation, d'un phénomène afin de l'expliquer, de le comprendre » [Samuel Legault-Mercier Michèle St-Pierre,2011].

b) Indicateur stratégique quantitatif

Un indicateur stratégique quantitatif : « mesure de la qualité ou appréciation chiffrée d'un phénomène pour offrir une certaine objectivation de la réalité étudiée. » [Samuel Legault-Mercier Michèle St-Pierre,2011].

4) Caractéristiques des bons indicateurs

Le choix d'indicateurs est une étape très importante, lors de la conception d'un tableau de bord. Selon, M. Hammer (2002), un indicateur stratégique « doit être précis et décrire réellement la situation à laquelle il s'applique. Il doit être objectif, ne pas prêter le flanc à la discussion ; compréhensible, facile à communiquer ; peu coûteux et simple à calculer ; disponible en temps utile. »

En effet, pour qu'on obtient un bon indicateur stratégique, il faut que cet indicateur dépende de certains critères :

- **La pertinence :** L'indicateur doit correspondre à une préoccupation, à un objectif ou à une attente. Il doit avoir une signification dans le contexte d'étude ou de gestion.
- **La qualité :** Pour la précision de sa mesure, un indicateur doit être formulé, précisément défini, ses paramètres bien établis et le tout est bien documenté.
- **La faisabilité** : Elle représente la possibilité de mesurer ou la disponibilité des données. Il faut s'assurer qu'il y a une personne qui va assumer la responsabilité d'alimenter, de produire et de fournir les indicateurs.
- **La convivialité** : Elle représente la possibilité opérationnelle, visuelle et cognitive d'utiliser correctement et confortablement l'indicateur.

En tant qu'outils d'appréciation et d'aide à la décision, les indicateurs souvent regroupés dans un tableau de bord considéré comme un outil de pilotage permettant à un ou plusieurs responsables d'être informés d'un coup d'œil, d'une situation donnée pour interpréter un phénomène.

II. Le tableau de bord

1) Définition d'un tableau de bord

Plusieurs spécialistes ont proposé différentes définitions des tableaux de bord ; on peut citer parmi eux : (Pierre Voyer, M. Gervais, H. Bouquin, Claude Alazard et Sabine Sépari)

- [Pierre Voyer, 1999] affirme que « Le tableau de bord mise principalement sur la qualité de l'information et non sur la quantité. Il met en évidence les résultats significatifs, les exceptions, les écarts et les tendances ; il fournit à son utilisateur un modèle cohérent en regroupant les indicateurs de façon à frapper son imagination. Ce schéma intégré permet d'enrichir d'autant l'analyse et l'interprétation de l'information ; il représente les indicateurs sous une forme compréhensible, évocatrice et attrayante, pour en faciliter la visualisation. »
- [M. Gervais,2000] estime que « Le tableau de bord confirme de façon structurée les impressions du responsable et lui indique la nécessité d'entreprendre une action ou une analyse plus approfondie. En cernant la zone à problème, il oriente des corrections à mener ou les pistes à explorer avant d'agir. »
- [H. Bouquin] mentionne que « [Le tableau de bord est] un ensemble d'indicateurs peu nombreux (5 à 10) conçus pour permettre aux gestionnaires de prendre connaissance de l'état de l'évolution des systèmes qu'ils pilotent et d'identifier les tendances qui les influenceront sur un horizon cohérent avec la nature de leurs fonctions. »
- [Claude Alazard et Sabine Sépari, 2010] « Un tableau de bord est un ensemble d'indicateurs organisés en système suivis par la même équipe ou le même responsable pour aider à décider, à coordonner, à contrôler les actions d'un service. Le tableau de bord est un instrument de communication et de décision qui permet au contrôleur de gestion d'attirer l'attention du responsable sur les points clés de sa gestion afin de l'améliorer ».
- [Gabriel Dabi-Schwebel, 22 juillet 2015] trouve que le « Tableau de bord (ou dashboard en anglais) est un outil informatique permettant de centraliser en un seul point un ensemble de données permettant de piloter une activité. Dans le domaine du web, on parle de tableau de bord web analytics ou digital analytics. »

2) Le rôle du tableau de bord

Pareil à un panneau de contrôle, le tableau de bord est un outil d'aide à la décision très important, il sert à ordonner et condenser l'information sous forme d'indicateurs. Il remplit notamment les rôles suivants :

- Fournir rapidement l'information essentielle, bien organisée et illustrée,
- Avertir l'analyseur à la manière d'un système d'alarme, de tout résultat ou écart indésirable,
- Faciliter aux décideurs la prise de connaissance de l'état et de l'évolution des systèmes qu'ils dirigent ainsi que pour suivre la mise en œuvre des objectifs fixés,
- Donner des sens à des données, et trouver des solutions ou des explications aux résultats obtenus,
- Mettre davantage l'accent sur les bons résultats en éliminant les données non nécessaires,
- Aider les dirigeants à prioriser les efforts sur les bons points (produits, clients, territoires...),
- Permettre de prendre des décisions éclairées basées sur des informations factuelles,
- Permettre d'anticiper sur les prospectives, pour que l'analyseur puisse voir plus loin.

3) Les modèles de tableau de bord

Assurément, on peut trouver plusieurs types de tableaux de bord ayant pour dénomination « tableau de bord de ... », et qu'on peut ajuster selon les phénomènes à traiter, parmi ces tableaux on cite :

- Le tableau de bord de tourisme tunisien,
- Le tableau de bord de gestion,
- Le tableau de bord des étudiants en informatique,
- Le tableau de bord de l'entreprise,
- Le tableau de bord de contrôle,
- Le tableau de bord financier,
- Le tableau de bord budgétaire, etc…

4) Caractéristiques de tableau de bord

Selon les chercheurs (Pierres Voyer, Ludovic Aubut-lussier, Alain Fernandez) pour qu'un tableau de bord soit bon et efficace, il faut qu'il obéît à la règle des « **3U** » :

- **UTILE**, Avoir un message clair (Savoir ce que l'on veut mesurer, ce que l'on veut suivre, assurer que l'indicateur est en lien avec ce qui nous intéresse) ;
- **UTILISABLE**, Avoir une représentation significative (Assurer la représentation de l'indicateur montre bien le message à communiquer, assurer que la représentation (diagramme ou autre) est comprise par les utilisateurs) ;
- **UTILISE :** Tester et raffiner le modèle (Le tableau de bord n'est pas statique, il faut accepter le fait que le tableau de bord est un outil qui évolue).

5) Processus d'élaboration d'un tableau du bord

La mise en place d'un tableau de bord suit un processus bien précis (voir figure suivante) :

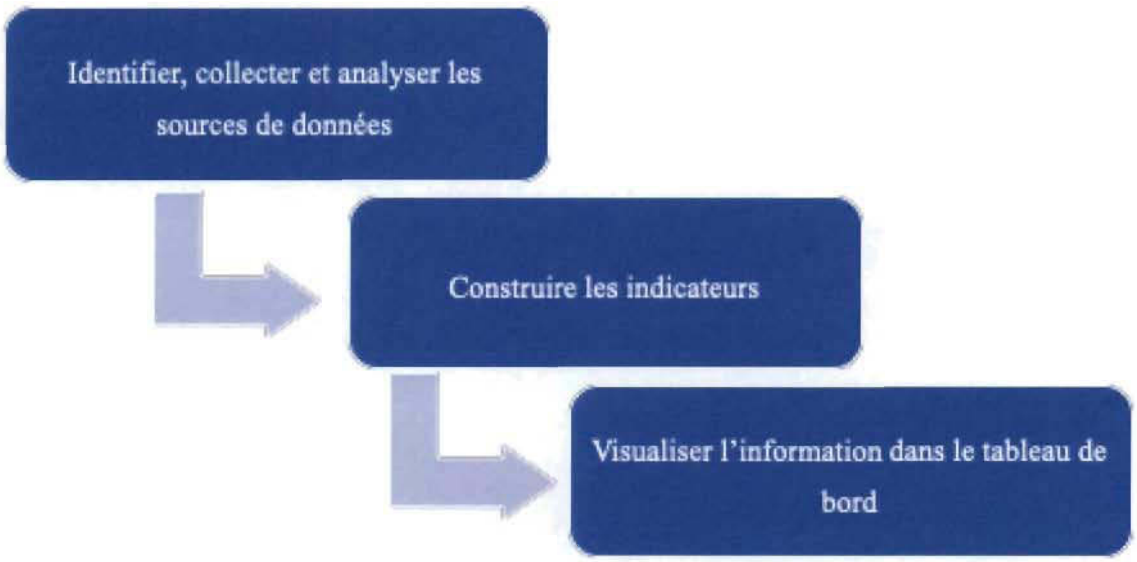

Figure 10 : Processus d'élaboration d'un tableau de bord

Identifie, collecte et analyse les sources de données :

La première étape consiste tout d'abord à identifier le problème à résoudre. Ensuite, on procède à la collection des données à traiter à partir d'une source fiable. Une fois les données sont collectées, un simple exercice de choix de logiciels informatiques est nécessaire pour bien sélectionner la meilleure solution d'analyse de données.

➢ Construire les indicateurs :

La construction des indicateurs varie selon le phénomène à étudier. Pour construire les bons indicateurs, il est important de répondre aux questions suivantes :

- Quelle est le type de mon indicateur ? (Quantitatif ou qualitatif)
- Comment se calcule-t-il? ainsi sa fréquence
- L'objectif auquel l'indicateur est lié
- Selon quels critères mon indicateur doit-il être étudié ?
- L'indicateur concerne-t-il toutes les instances d'un processus ou seulement une partie d'entre elles ?
- Que fait avec l'information fournie par l'indicateur ?

Ces indicateurs permettent d'anticiper, décider et contrôler les politiques et pratiques. L'efficacité globale du pilotage repose sur la fiabilité des informations, la pertinence des indicateurs et leur adaptation aux besoins spécifiques des différents décideurs et acteurs.

➢ Visualise l'information dans le tableau de bord :

La dernière étape, c'est la mise en place et l'organisation des indicateurs dans un tableau de bord. Un tableau de bord qui s'appuie sur un ensemble d'indicateurs issus d'informations disponibles de l'étape précédente est consultable en un seul coup d'œil. Il permet d'interpréter une opinion précise sur la situation de problème à résoudre sans réfléchir trop longuement.

Les représentations graphiques devraient être choisies avec soin en tenant compte de la nature de l'information et du message porté. Plusieurs options sont possibles pour présenter les informations :

- **Par groupe d'indicateurs**

Assemble les indicateurs par objectif en créant des sections.

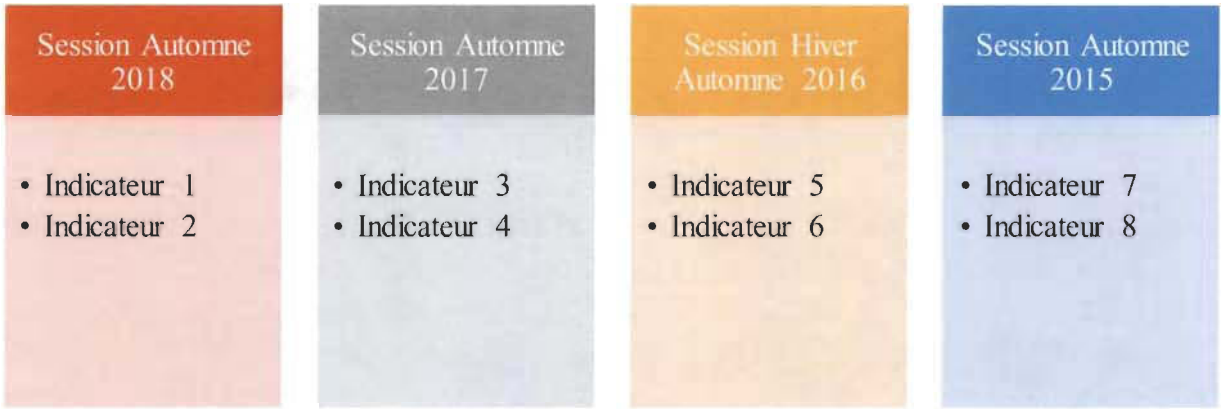

Figure 11 : Affichage des indicateurs par groupe

- **Par niveau de détail**

Représente une hiérarchie entre les indicateurs allant du plus synthétique au plus détaillé.

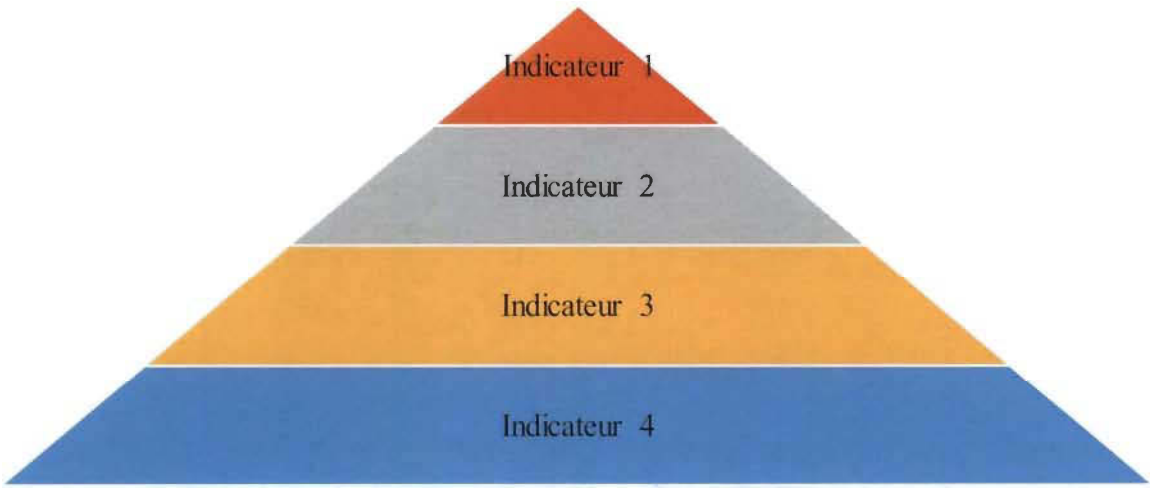

Figure 12 : Afficher les indicateurs par niveau de détail

- **Par lien de causalité**

Relie l'ensemble des indicateurs qui ont un effet l'un sur l'autre.

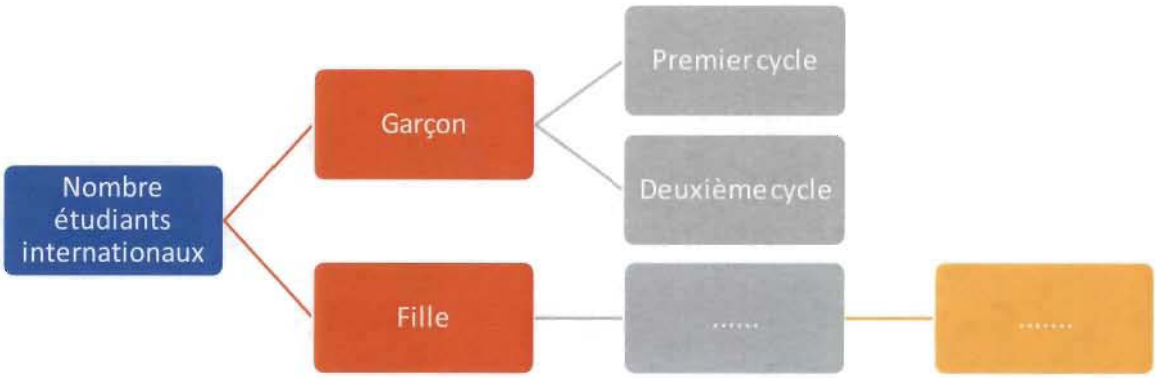

Figure 13 : Afficher les indicateurs par lien de causalité

« Un graphique permet d'interpréter une mesure d'un seul coup d'œil. Il complète avantageusement les informations clés affichés. Mais il faut faire attention aussi au choix de couleur car elles sont là pour aider et non brouiller le message. Il convient donc d'éviter les couleurs criardes [...] pour distinguer les chiffres positifs (des hausses ou des objectifs dépassés) des négatifs (des pertes, des objectifs non atteints), le vert et le rouge sont toujours de rigueur. » [L'équipe de Manager GO, 17 juillet 2018]. Toutefois, le choix d'un graphique significatif garantissant une visualisation claire et simple demeure un défi. Ci-après quelques recommandations avec les types de graphique de base :

- **Représente une proportion**

Pour représenter un pourcentage, le diagramme circulaire (camembert) est bien approprié. Toutefois, il faut limiter la représentation à un maximum de cinq valeurs pour ne pas rendre le graphique trop chargé et donc illisible. L'illustration ci-dessous est un graphique de type diagramme circulaire :

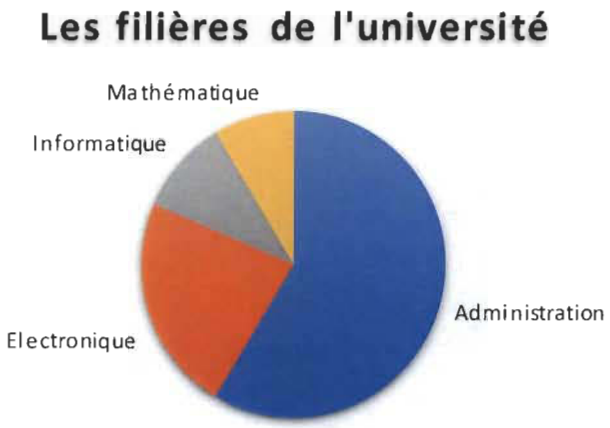

Figure 14 : Diagramme circulaire

- **Représente un ratio**

La jauge est plus pertinente que le diagramme circulaire, car elle affiche le positionnement entre 0% et 100%. Elle sert à indiquer la mesure d'un système surveillé à l'aide d'une aiguille ou d'un pointeur en se déplaçant le long d'une échelle étalonnée. L'illustration ci-dessous est un graphique de type jauge :

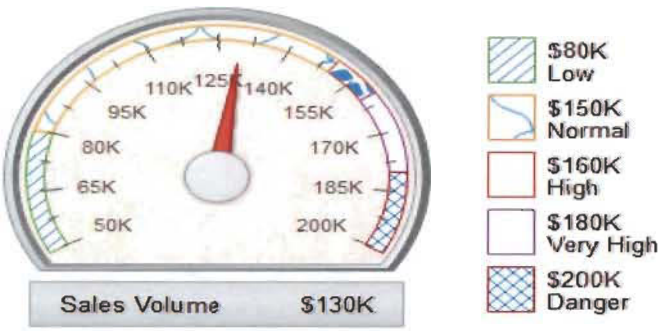

Figure 15 : Une représentation de jauge

- **Représente une progression**

Les graphiques de type ligne aident à visualiser rapidement les tendances, les progressions sur un espace-temps donné. On peut mettre en valeur des progressions positives et négatives à l'aide de couleurs pour distinguer entre les deux. Il affiche des informations sous la forme d'une série de points de données appelés « marqueurs » reliés par des segments de droite. C'est un type de base de graphique commun dans de nombreux

domaines. Il est similaire à un nuage de points, sauf que les points de mesure sont ordonnés (généralement par leur valeur sur l'axe des x) et joints à des segments de droite. L'illustration ci-dessous est un graphique de type ligne :

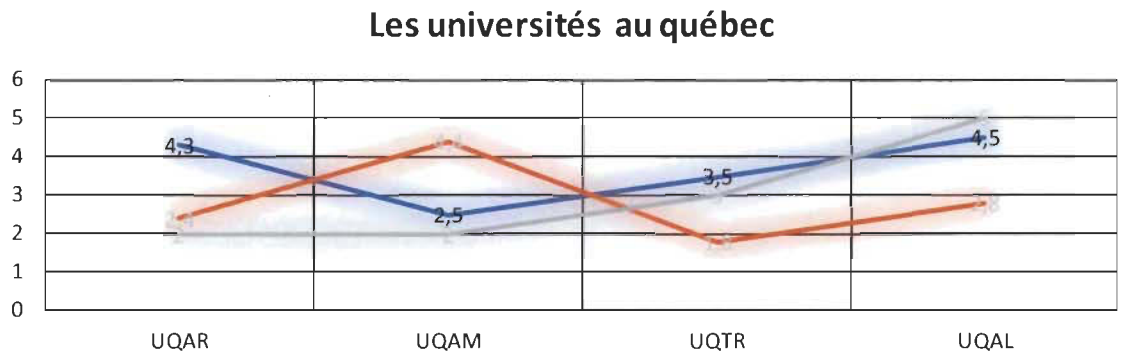

Figure 16 : Un graphique de type ligne

- **Compare des catégories, des niveaux**

L'histogramme (diagramme en barres), est une représentation graphique précise de la distribution des données numériques. Il est tout indiqué pour mettre côté à côté des données afin de faciliter la comparaison de valeurs. Il permet de représenter la répartition d'une variable continue en la représentant avec des colonnes verticales. L'illustration ci-dessous est un histogramme :

Figure 17 : Une représentation d'un histogramme

III. Conclusion

A la fin de cette partie, on peut conclure qu'une représentation graphique représente un moyen de condenser et de simplifier la communication des données. Elle permet de présenter certains résultats de manière claire et concise en transformant de façon efficace les données en information pertinente. Une visualisation de données réussie permet de donner une valeur importante du résultat obtenu par un analyste. Elle rapporte une histoire et accompagne la transmission de l'information grâce à des éléments de couleur et de contexte pour valoriser les informations importantes, et permette de les rendre actionnables et opérationnelles.

On a montré qu'il existe plusieurs formes possibles de présentation visuelle des résultats tels que: tableaux, graphiques, histogrammes, diagrammes à barres horizontales et verticales, diagrammes circulaires, cartes, pictogrammes et bandes dessinées. Le choix dépend de la nature de l'information à présenter et de l'auditoire visé.

CHAPITRE VI : APPROCHE PROPOSEE

I. L'approche proposée

1) Description de l'approche

Compte tenu des limites de processus d'analyse des textes non structurées et multilingues, nous proposons une approche se composant de trois grandes phases. La première phase est une étape préalable à l'analyse qui consiste à détecter et à supprimer les textes mal structurés qui n'ont aucune signification. De plus, elle consiste à structurer le texte en traduisant certains mots ou tout le texte pour unifier la langue. Ces actions permettent de réduire le temps de prétraitement de la prochaine phase ainsi d'obtenir un résultat d'analyse plus précis. La deuxième phase consiste à ppliquer le processus habituel d'analyse de donnée. La troisième phase consiste à transformer les résultats qui se dégagent de la phase 2 en indicateurs qui seront présentés dans un tableau de bord permettant une meilleure visualisation des résultats de manière qu'ils soient intelligibles et exploitables facilement. Au bout de ces trois phases, l'approche préconisée permettra d'alimenter un outil de gestion, soit le tableau de bord, avec des indicateurs permettant à un gestionnaire de mieux décider.

La figure ci-dessous illustre en détaillant la démarche à suivre pour l'application de notre approche. Toutes ces étapes seront expliquées par la suite :

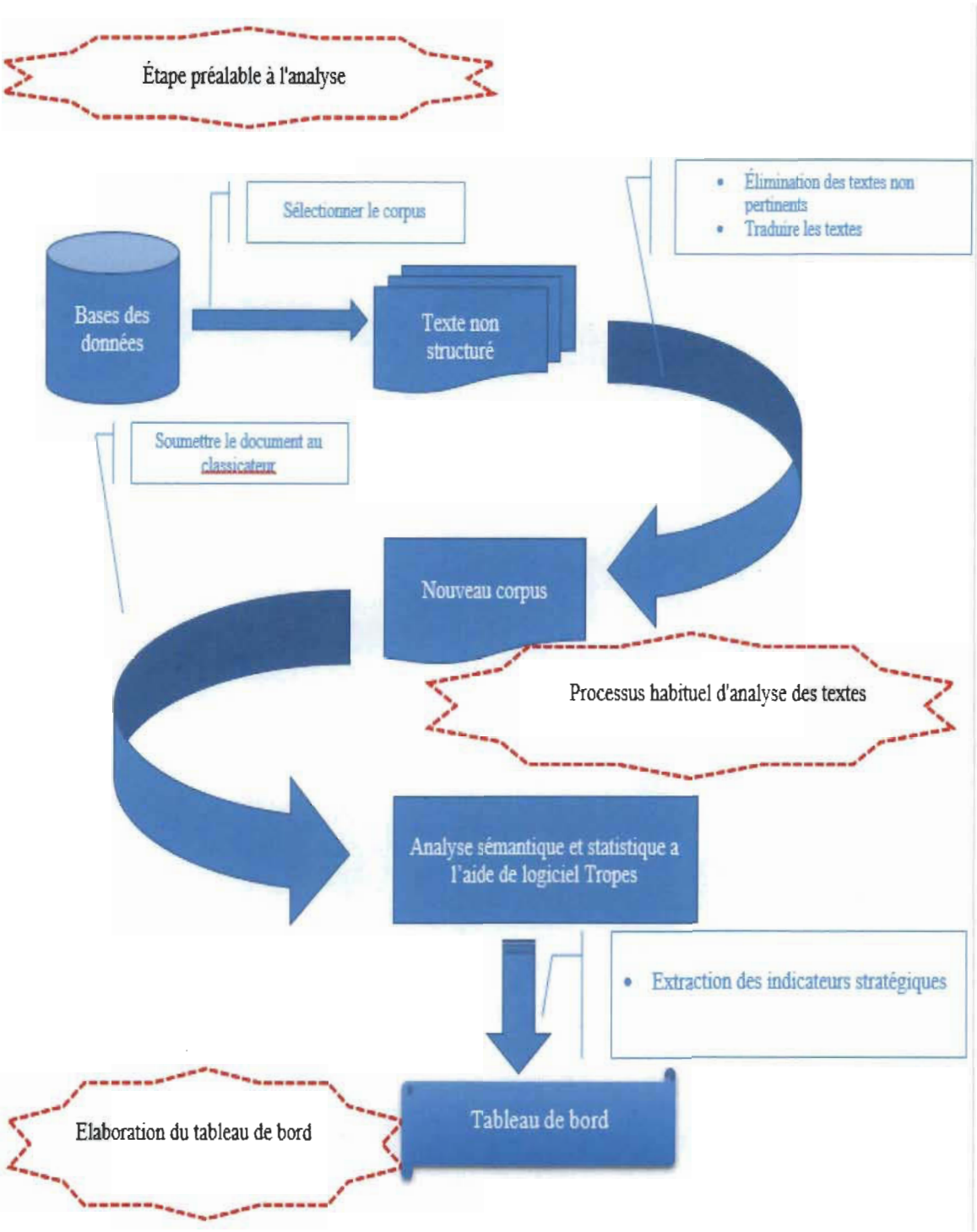

Figure 18 : La démarche de notre approche

2) L'algorithme de l'approche proposée

L'algorithme suivant explique toutes les démarches de notre approche proposée :

Algorithme : L'approche proposée

Entrée : Des données textuelles
Sortie : Tableau de bord

1) **Sélection de corpus**
2) **Détection des textes mal structurés et incompréhensibles**
    a) Si le texte est bon, alors on le retient
    b) Sinon on le supprime
3) **Détection de la langue**
    a) Si le commentaire est écrit en français, on le retient
    b) Sinon on le traduit
4) **Traduire le texte**
    a) Si le texte est multilingue alors la traduction sera faite par un expert de la langue
    b) Sinon la traduction sera faite par le traducteur Deepl
5) **Sélection de nouveau corpus**

a) On sélectionne seulement les textes qui sont retenus et qui sont écrits en français

6) **Processus habituel d'analyse des textes à l'aide de logiciel d'analyse tropes**
    a) Exécuter le processus de prétraitement
        i. La tokenisation
        ii. Mots vides (stop-word)
        iii. Normalisation lexicale
        iv. La racinisation
    b) Exécuter le processus de la classification
        i. La pondération
        ii. La représentation vectorielle
        iii. Exécution de tropes
    c) L'analyse des résultats
7) **Identifier les indicateurs stratégiques**
8) **Construire un tableau de bord**

Algorithme 1 : L'approche proposée

3) Les différentes phases de l'approche

a) La première phase :

Cette phase consiste à sélectionner notre corpus en tenant compte de la chronologie des commentaires et aussi à éliminer les textes mal structurés. Ces textes seront traduits à l'aide de l'outil DeepL, si le texte est écrit avec une seule langue autre que le français. Sinon on le traduit manuellement, si notre le texte est multilingue. Cela nous permet de minimiser le

bruit dans les textes multilingues et par conséquent de réduire le temps de prétraitement de la phase suivante.

- **Détection et élimination des textes mal structurés :**

Les n-grammes sont une méthode permettant de détecter des mots hors vocabulaire (OOV). Ils sont des chaînes de caractères qui peuvent être composées d'un seul caractère (1-grammes), deux caractères (2-grammes) et jusqu'à n caractères (n-grammes). La longueur la plus fréquemment utilisée est celle de 3 caractères.

Le tableau ci-dessous donne un exemple pour le traitement de mots « Québec »:

1-gram	q	u	é	b	e	c
2-gram	qu	ué	éb	be	ec	
3-gram	que	uéb	ébe	bec		
4-gram	québ	uébe	ébec			

Tableau 3 : Les traits (n-gram)

De nombreux travaux ont montré l'efficacité des n-grammes comme méthode de représentation des textes pour leur classification [Adeline et al, 2018 ; Nicolas Despres et al, 2016 ; Yves Bestgen, 2014]. Toutefois, malgré son efficacité, cette méthode peut échouer lorsque le commentaire comprend des mots qui n'étaient pas présents dans le dictionnaire ou la base de données de notre système.

Il existe une autre solution pour éviter l'échec, elle consiste à introduire une étape supplémentaire qui exploite des ressources web pour créer des dictionnaires dynamiques, mais elle prend assez de temps, c'est pour cela, une intervention humaine est nécessaire pour éviter cet échec.

L'algorithme proposé pour supprimer les textes mal structurés.

Algorithme : Suppression des textes mal structurés

Entrée : Commentaire structuré ou non structuré
Sortie : Commentaire structuré

1) **Détection des phrases structurées**

a) Pour chaque classe $\boldsymbol{j}$, recherche tous les n-grammes dans tous les textes de l'ensemble d'apprentissage.

b) Constituer le tableau croisé $(\boldsymbol{N_{ij}})$ des occurrences des n-grammes $\boldsymbol{i}$ dans la classe $\boldsymbol{j}$,

c) Calculer les fréquences $\boldsymbol{f_{ij}}$ correspondantes : $\boldsymbol{f_{ij}} = \frac{N_{ij}}{N}$

d) Calculer les contributions de $(\boldsymbol{ij})$ à la statistique du $\boldsymbol{\chi^2}$:

a. $\boldsymbol{\chi^2_{ij}} = \frac{\left(N_{ij} - \frac{N_i \times N_j}{N}\right)^2}{\frac{N_i \times N_j}{N}} = \boldsymbol{N} \times \frac{(f_{ij} - f_i \times f_j)^2}{f_i \times f_j}$

e) Calculer le $\boldsymbol{\chi^2_{ij}} \times \boldsymbol{signe}(\boldsymbol{f_{ij}} - \boldsymbol{f_i} \times \boldsymbol{f_j})$

f) Trier le tableau de $\boldsymbol{\chi^2_{ij}}$ dans l'ordre croissant

g) Pour chaque classe $\boldsymbol{j}$ faire

i. Déterminer la liste $\{\boldsymbol{gram_{ij}}\}$ des $\boldsymbol{k}$ premiers n-grammes de la classe

ii. Pour chaque $\boldsymbol{gram_{ij}}$ faire

- Chercher tous les mots $(\boldsymbol{mot_{jk}})$ tels que $\boldsymbol{gram_{ij}} \subseteq \boldsymbol{mot_{jk}}$
- Calculer le nombre $\boldsymbol{nb_{mots_{jk}}}$ des répétitions de $\boldsymbol{mot_{jk}}$ dans la classe

i. Si le nombre $\boldsymbol{nb_{mots_{jk}}} \neq \boldsymbol{0}$, (on retient le commentaire)

2) **Suppression les phrases mal structurées**

ii. Si le nombre $\boldsymbol{nb_{mots_{jk}}} = \boldsymbol{0}$, (supprimer le commentaire)

Algorithme 2: La Suppression des textes mal structurés

L'idée générale de l'algorithme a était proposé par Radwan Jalam et Jean-Hugues Chauchat, pour répondre à la question de l'efficacité de n-grammes comme un outil de classement des textes. Nous avons modifié l'algorithme pour qu'il sert à détecter et à compter les mots incompréhensibles. Dans le cas où on trouve que les nombre de n-grammes est égale à zéro, ça veut dire que le texte ne contient que des mots incompressibles et hors vocabulaire (OV). Dans ce cas, on supprime le commentaire et si les nombres de n-gramme diffèrent de zéro on retient le commentaire.

- **Traduire les commentaires :**

Détecter la langue :

La détection automatique et avec précision de la langue utilisée dans le commentaire représente une étape importante et cruciale dans notre démarche d'analyse car une erreur à ce niveau voue à l'échec des étapes suivantes. Plusieurs approches ont été identifiées dans la littérature pour la détection de langue :

- Des approches manuelles telles que les guides destinées aux bibliothécaires,
- Des approches semi-automatiques basées sur l'apprentissage supervisé telles que les réseaux de neurones, les chaînes de Markov et les approches probabilistes.
- Des approches automatiques en utilisant les bases de connaissances,

Les approches utilisant des bases de connaissances sont plus précises et ils sont déployées dans les cas où les approches semi-automatiques ne parviennent pas à détecter la langue des textes très courts dont la longueur maximale est celle d'une phrase. Mais dans notre cas on a choisi l'approche manuelle car si un document est rédigé en plusieurs langues, une tentative de détection de la langue dominante du document est effectuée en ignorant les autres mots de multilingue et pour cela les résultats de l'analyse ne sont pas toujours satisfaisants.

Ainsi, dans le cas où les commentaires sont courts et multilingues, il sera mieux d'utiliser l'approche manuelle pour la détection de la langue. Elle sera faite par l'intervention d'un expert linguistique. La figure ci-dessous présente le processus de détection manuelle de la langue d'un commentaire :

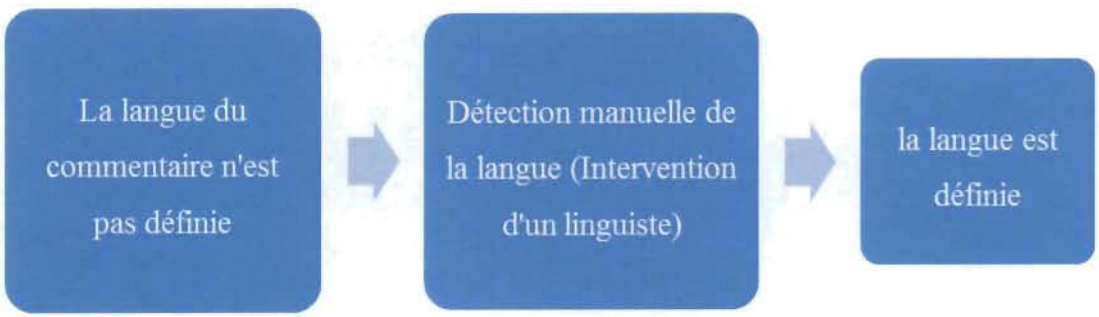

Figure 19 : Processus de détection manuelle de la langue d'un commentaire

Traduire les textes

Après l'étape de la détection de la langue en utilisant l'approche manuelle, on a trois possibilités pour traduire nos textes :

- Soit d'une façon manuelle,
- Soit d'une façon semi-automatique,
- Soit d'une façon automatique.

Dans le cas où le commentaire est écrit avec une seule langue autre que le français, on utilise Deepl pour le traduire en français (la façon automatique). La figure ci-dessous montre le Processus de traduction d'un commentaire à l'aide de l'outil Deepl :

Figure 20 : Processus de traduction d'un commentaire à l'aide de l'outil Deepl

Dans le deuxième cas où le commentaire est multilingue, l'internaute a écrit en français mais aussi il a utilisé une autre langue dans son commentaire. On traduit les mots manuellement vers le français à l'aide d'un expert linguistique (la façon manuelle) où on peut utiliser l'outil Deepl pour traduire ces mots lors de l'absence de l'expert (la façon semi-automatique).

La figure ci-dessous illustre le Processus manuel de la traduction d'un commentaire :

Figure 21 : Processus manuel de la traduction d'un commentaire

A part la détection manuelle, on a proposé une technique qui permet de détecter la langue. Cette technique se base sur l'analyse de fréquence de l'ensemble des lettres de l'alphabet. Elle a été utilisée par les linguistes comme technique rudimentaire d'identification de la langue. Elle est particulièrement efficace pour indiquer si un système d'écriture inconnu est alphabétique, syllabique ou idéographique. Elle est constituée de trois étapes :

- Éliminer de la ponctuation et les espaces dans un texte;
- Transformer le texte en lettres minuscules;
- Analyser la fréquence de l'ensemble des lettres de l'alphabet.

L'algorithme proposé de la première étape, consiste à supprimer la ponctuation ainsi que les espaces dans nos commentaires :

Algorithme : Suppression de la ponctuation et les espaces dans un commentaire

Entrée : Un commentaire
Sortie : Commentaire transformé
Début de l'opération
nbrcar prend la valeur de longueur(commentaire) ;
Alphabétique = "abcdefghijklmnopqrstuvwxyzàœçèéêëîïñßù" ;
Commentaireformate prend la valeur 0 ;
Pour k **allant de** 1 **à nbcar faire**
j prend la valeur 0 ;
Trouve prend la valeur 0 ;
Tant que j < longueur(alphabétique) et trouve = 0 faire
Si commentaire[k] = alphabétique[j] **alors**
Trouve = 1 ;
fin
Si trouve = 0 **alors**
j prend la valeur de j + 1 ;
fin
fin
Si trouve = 1 **alors**
Commentaireformate = commentaireformate + commentaire[k] ;
fin
fin
Fin

Algorithme 3: La suppression de la ponctuation et les espaces dans un commentaire

L'algorithme suivant présente la deuxième étape, il sert à transformer un commentaire qu'on a supprimé de tout signe de ponctuation en lettres minuscules :

Algorithme : Transformer le commentaire en lettres minuscules

Entrée : Un commentaire
Sortie : Commentaire formaté
Début de l'opération
nbrcar prend la valeur de longueur(commentaire) ;
Commentairetransformer prend la valeur 0 ;
Pour k **allant** de 1 à nbrcar **faire**
Si ord(commentaire[k]) <97 **alors**
Commentairetransformer = commentairetransformer+chr(ord(texte[k]) +32) ;
Sinon
commentairetransforme prend la valeur de commentairetransforme + texte[k] ;
Fin
Fin
Fin

Algorithme 4: La Transformation des commentaires en lettres minuscules

Après l'exécution de deux algorithme précédents, l'algorithme suivant permet d'afficher la fréquence de chacune des lettres de l'alphabet dans nos commentaires :

Algorithme : La fréquence des lettres de l'alphabet dans un commentaire
Entrée : Un commentaire
Sortie : La fréquence des lettres

Début de l'opération
nbrcar prend la valeur de longueur(Commentaire) ;
Alphabétique prend la valeur "abcdefghijklmnopqrstuvwxyzàœçèéêëîïñßù"
Pour j **allant de** 1 **à** 38 **faire**
Compteur prend la valeur 0 ;
Pour k **allant de** 1 **à** nbrcar **faire**
Si texte[k] = alphabétique[j] **alors**
Compteur prend la valeur de compteur+1 ;
Fin
Fréquence prend la valeur compteur nbrcar $= \frac{Compteur}{\text{nbrcar}}$
Afficher fréquence
Fin
Fin
Fin

Algorithme 5: La Transformation des commentaires en lettres minuscules

Le tableau ci-dessous illustre les distributions de fréquence des 38 caractères les plus courantes dans les langues (Français, allemand, espagnol, portugais, italien, anglais). Toutes ces langues utilisent un alphabet similaire de 26 caractères.

Lettre	Français	Allemand	Espagnol	Portugais	Italien	Anglais
a	7.636%	6.51%	12.53%	14.63%	11.74%	8.167%
b	0.901%	1.89%	1.42%	1.04%	0.92%	1.492%
c	3.260%	3.06%	4.68%	3.88%	4.5%	2.782%
d	3.669%	5.08%	5.86%	4.99%	3.73%	4.253%
e	14.715%	17.40%	13.68%	12.57%	11.79%	12.702%
f	1.066%	1.66%	0.69%	1.02%	0.95%	2.222%
g	0.866%	3.01%	1.01%	1.30%	1.64%	2.015%
h	0.737%	4.76%	0.70%	1.28%	1.54%	6.094%
i	7.529%	7.55%	6.25%	6.18%	11.28%	6.966%
j	0.545%	0.27%	0.44%	0.40%	0.00%	0.153%
k	0.049%	1.21%	0.01%	0.02%	0.00%	0.772%
l	5.456%	3.44%	4.97%	2.78%	6.51%	4.025%
m	2.968%	2.53%	3.15%	4.74%	2.51%	2.406%
n	7.095%	9.78%	6.71%	5.05%	6.88%	6.749%
o	5.378%	2.51%	8.68%	10.73%	9.83%	7.507%
p	3.021%	0.79%	2.51%	2.52%	3.05%	1.929%
q	1.362%	0.02%	0.88%	1.20%	0.51%	0.095%
r	6.553%	7.00%	6.87%	6.53%	6.37%	5.987%
s	7.948%	7.27%	7.98%	7.81%	4.98%	6.327%
t	7.244%	6.15%	4.63%	4.74%	5.62%	9.056%
u	6.311%	4.35%	3.93%	4.63%	3.01%	2.758%
v	1.628%	0.67%	0.90%	1.67%	2.10%	0.978%
w	0.114%	1.89%	0.02%	0.01%	0.00%	2.360%
x	0.387%	0.03%	0.22%	0.21%	0.00%	0.150%
y	0.308%	0.04%	0.90%	0.01%	0.00%	1.974%
z	0.136%	1.13%	0.52%	0.47%	0.49%	0.074%
à	0.486%	0	0	0	0	0
œ	0.018%	0	0	0	0	0
ç	0.085%	0	0	0.53%	0	0
è	0.271%	0	0	0	0.263%	0
é	1.904%	0	0	0.337%	0	0
ê	0.225%	0	0	0	~ 0%	0
ë	0.001%	0	0	0	0	0
î	0.045%	0	0	0	0	0
ï	0.005%	0	0	0	0	0
ñ	0	0	0.31%	0	0	0
ß	0	0.31%	0	0	0	0
ù	0.058%	0	0	0	0	0

Tableau 4 : Fréquences relatives des lettres (source : Wikipédia)

Le tableau ci-dessous, c'est la liste des cinq premières lettres en fréquence et les plus utilisées. Si le message est écrit en majuscule, l'ordre change car on va réunir par exemple les fréquences du a et du à. C'est pour cela on a ajouté l'algorithme qui transforme toutes les lettres en minuscule.

Français	e	s	a	i	t
Anglais	e	t	a	o	i
Espagnol	e	a	o	s	r
Allemand	e	n	i	s	r
Italien	e	a	i	o	n
Portugais	a	e	i	s	r

Tableau 5: Liste des 5 premières lettres les plus fréquents

Par conséquent, grâce à cette technique, on peut définir le pourcentage d'apparition de chaque lettre dans un commentaire. Ainsi on peut déduire la langue utilisée.

b) La deuxième phase :

La deuxième phase consiste à faire une analyse sémantique et une autre statistique des avis en ligne. Le fait de travailler avec un logiciel d'analyses textuelles constitue un premier choix méthodologique fort. C'est dans ce contexte, nous avons choisi Tropes parmi une offre diversifiée de logiciels d'analyses.

L'analyse automatique du logiciel Tropes permettra d'identifier les termes clés du tourisme tunisien. Ce logiciel permet de déterminer, au sein d'un ou plusieurs textes, qu'elles sont les variables du tourisme tunisien et quelles sont les relations qui les lient. Il permettra également de repérer et de dénombrer les occurrences et les co-occurrences des éléments.

Nous nous sommes basés sur l'hypothèse que les adjectifs et les verbes étaient les deux traits grammaticaux les plus utilisés pour exprimer des opinions et jugement. Les étapes de processus d'analyse des textes sont illustrées par le schéma suivant :

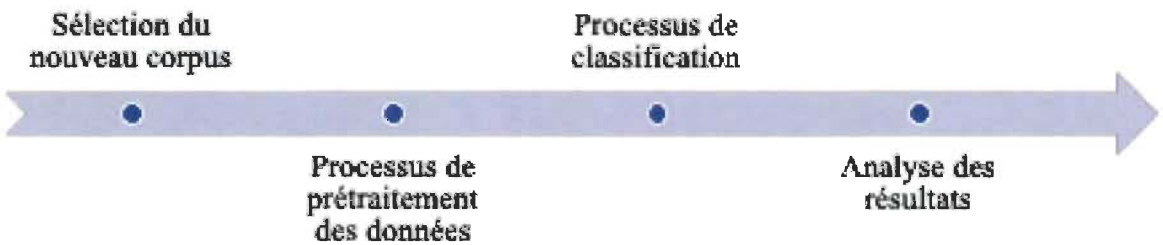

Figure 22 : Processus d'analyses des textes

Étape 1 : Sélection du nouveau corpus :

Après l'élimination des textes mal structurés, incompréhensibles, et aussi après l'unification de la langue dans la première étape de notre approche proposée (étape préalable à l'analyse). On a maintenant un nouveau corpus qu'est tout à fait différent à notre premier corpus sélectionné.

Étape 2 : Processus de prétraitement des données

La deuxième étape d'un processus d'analyses des textes est l'étape de prétraitement des données. De nombreuses recherches dénoncent la complexité et le côté fastidieux du prétraitement des données. Sur ce constat, les internautes et surtout celles qui utilisent les forums et les réseaux sociaux en général utilisent un style d'écriture incompréhensible par la machine et commettent souvent des fautes d'orthographe et de grammaire. En effet, ils ont estimé que le prétraitement occupe 60% du temps total du processus d'analyse de textes. Plus précisément « les prétraitements sont aujourd'hui des tâches qui en plus de demander des compétences techniques complexes, sont très gourmandes en temps et peuvent introduire des aléas dans les analyses qui suivront. » [Daras et al, 2017]

Autrement dit,l'étape de prétraitement, c'est une étape dont l'objectif est de préparer les textes bruts pour les étapes suivantes, ainsi il s'agit d'effectuer des transformations des données pour les rendre plus utilisables par les algorithmes d'apprentissage. Selon [Mathieu Feuilloy, mars 2010] « C'est pourquoi, avant de chercher à obtenir de bonnes performances de classification, les données doivent subir un prétraitement afin d'éliminer toute incertitude sur leur légitimité à apparaître dans la base de données ».

Toutes ces caractéristiques linguistiques et orthographiques des avis en lignes des internautes nous obligent à procéder à une étape de prétraitement. Il sert à faire :

- La vérification de la qualité,
- La tokenisation
- La cohérence des données,
- La correction des erreurs et des données erronées,
- La suppression des valeurs manquantes et les mots vides,
- La racinisation
- Normalisation lexicale

L'étape de prétraitement peut aussi contenir les opérations d'échantillonnage et de transformation des données, comme par exemple l'ajout de données extraites du Web ou d'une autre source de données, le calcul d'écart, moyennes, somme etc…

La figure ci-dessous montre le premier aperçu de nuage de mots-clés qu'est une sorte de condensé sémantique d'un document dans lequel qui permet de présenter l'information par rapport à l'occurrence des mots-clés contenus dans notre corpus.

Le Processus de prétraitement des données qu'on a proposé se résume comme suit :

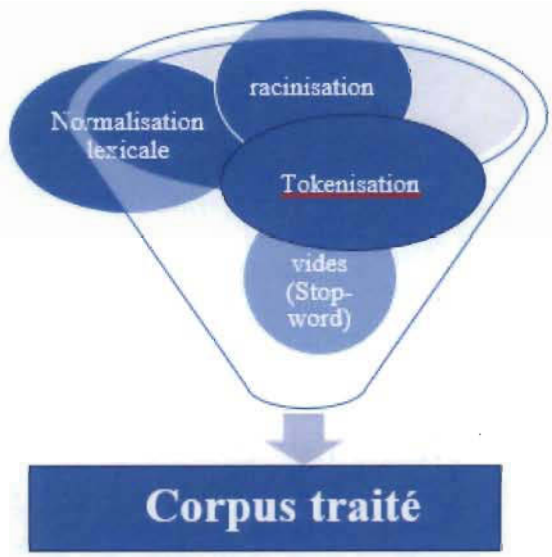

Figure 23 : Processus de prétraitement proposé

- Tokenisation

La tokenisation est un processus qui permet de découper un texte ou un commentaire en mots, phrases et d'autres éléments significatifs appelé token. Un token est une unité définie comme une séquence de caractères comprise entre deux séparateurs (espace, signe de ponctuation, guillemets, parenthèses…). Il existe plusieurs méthodes de tokenisation dont la séparation en phrase ou en mots, qui est la plus utilisée.

Le tableau ci-dessous représente un exemple de la tokenisation d'une phrase en mots :

Phrase simple	Identification d'indicateurs stratégiques dans les documents
Tokenistation d'une phrase	« Identification », « de », « indicateurs », « stratégiques », « dans », « les », « documents »

Tableau 6 : Tokenistation d'une phrase

L'algorithme proposé de la Tokenistation des commentaires :

L'algorithme de la Tokenistation d'un commentaire

Entrée : Un commentaire

Sortie : Commentaire tokenisé

1. Les mots des commentaires sont stockés dans un tableau.
2. Remplacer les caractères suivants par des espaces **: ;,**
3. Supprimer les caractères suivants s'ils sont suivis d'un espace ou s'ils sont procédés par espace **: ;,**
4. Remplacer les paires de crochets **( )** et **[ ]** par des espaces
5. Supprimer les paires de crochets **( )** et **[ ]** si le crochet ouvert est précédé d'un espace où le crochet de fermeture est suivi d'un espace
6. Remplacer les guillemets simples **' où " " où ' ' où ‹ › où " "** par un espace
7. Supprimer les guillemets simples **' où " " où ' ' où ‹ › où " "**dans le cas où ils sont précédés d'un espace ou s'ils sont suivis d'un espace
8. Remplacer la barre oblique / par un espace
9. Supprimer la barre oblique / dans le cas où elle est suivie d'un espace ou elle est précédée d'espace
10. Remplacer les caractères suivants par un espace: **! " # %$ % & * < = > ? @ \ | ~ ' '**
11. Supprimer les caractères suivants par des espaces **! " # %$ % & * < = > ? @ \ | ~ ' '** s'ils sont précédés d'un espace où dans le cas ils sont suivis d'un espace

Algorithme 6 : La Tokenistation des commentaires

- Mots vides (Stop-word)

Les mots vides sont les mots les plus courants dans une langue et non significatif figurant dans un texte. Ces mots représentent 1/3 du contenu d'un texte. On peut constater les mots vides facilement suite à la tokenisation d'un commentaire. Ces mots vides existent dans tous les langues :

- ✓ En anglais, certains de ces mots sont « the », « all », « and », etc.
- ✓ En français, certains de ces mots sont « le », « la », « ce », etc.

La présence de ces mots n'apporte absolument aucune différence ni sur le plan lexical, ni sur le plan sémantique. Ainsi, leur utilisation pour la classification s'avère inutile. En d'autres termes, leur suppression réduit la dimension de notre document vecteur, cela nous aide à améliorer le temps de traitement et aussi réduire considérablement le temps d'apprentissage.

Dans le tableau suivant, on applique un exemple pour la suppression des mots vides :

Tokenistation d'une phrase	« Identification », « de », « indicateurs », « stratégiques », « dans », « les », « documents »
Suppression des mots vides	« Identification », « indicateurs », « stratégiques », « documents »

Tableau 7: Suppression des mots vides dans une phrase

Pour calculer le pourcentage de la présence des mots vide dans un document et pour savoir l'importance de cette étape, on applique cette formule :

$$\% \text{ Mv} = = \frac{100 * \text{NMs}}{\text{NMt}}$$

$$\% \text{ Mv} = = \frac{100 * 3}{7} = 42{,}85\%$$

Avec Mv : Mots vides

NMs : Nombre des mots vides

NMt : Nombre des mots totaux

Mais dans certains cas et selon le sujet traité, il faut faire attention en supprimant les mots vides puisqu'ils peuvent influencer notre analyse et affecter la précision du résultat.

On peut prendre un exemple simple et significatif, lors de la suppression de mot « Est » dans ces deux cas, on constate que le sens de la phrase a complétement changé :

- ➢ Le tourisme **est** beau en Tunisie.
- ➢ Le tourisme en **Est** de la Tunisie **est** beau.

Pour la première phrase indique le tourisme en général en Tunisie contrairement à la deuxième phrase qui précis un lieu spécifique où se trouve le tourisme.

Pour l'algorithme proposé de la suppression des mots vides ci-dessous, la liste des mots vides est modifiable. Comme on a mentionné pour le mot « Est » et pour qu'il n'y aura aucun conflit avec le verbe être et celle de direction. On a ignoré ce terme de notre liste.

L'algorithme de la suppression des mots vides

Entrée : Un commentaire

Sortie : Commentaire sans mot vide

Mot_vide : (alors, au, aucuns, aussi, autre, avant, avec, avoir, bon, car, ce, cela, ces, ceux, chaque, ci, comme, comment, dans, des, du, dedans, dehors, depuis, devrait, doit, donc, dos, début, elle, elles, en, encore, essai, et, eu, fait, faites, fois, font, hors, ici, il, ils, je, juste, la, le, les, leur, là, ma, maintenant, mais, mes, mine, moins, mon, mot, même, ni, nommés, notre, nous, ou, où, par, parce, pas, peut, peu, plupart, pour, pourquoi, quand, que, quel, quelle, quelles, quels, qui, sa, sans, ses, seulement, si, sien, son, sont, sous, soyez, sujet, sur, ta, tandis, tellement, tels, tes, ton, tous, tout, trop, très, tu, voient, vont, votre, vous, vu, ça, étaient, état, étions, été, être)

1. Les mots de commentaires sont stockés dans un tableau.
2. Un seul mot vide sera sélectionné à partir de la liste de mots vides.
3. Le mot vide est comparé au commentaire choisi sous forme de tableau en utilisant la technique de recherche séquentielle.
4. Si cela correspond, le mot du tableau sera supprimé et le mot vide continue sa comparaison jusqu'à la fin du tableau.
5. On sélectionne un autre mot vide à partir de la liste et on recommence l'algorithme depuis l'étape 2.
6. L'algorithme fonctionne jusqu'à ce que tous les mots vides soient comparés.

Algorithme 7: La suppression des mots vides

- Normalisation lexicale

La tâche de la normalisation consiste à réduire les mots à leurs racines et permet de réécrire les commentaires de forum à leurs formes canoniques. Notre objectif n'est pas de faire la correction orthographique et syntaxique, mais de réécrire le texte en se basant sur les

erreurs lexicales fréquentes surtout dans les forums et aussi dans les réseaux sociaux (facebook, Twitter, Instagram).

Le tableau suivant présente quelques exemples de normalisation lexicale :

Numéro de tâche	Tâche	exemple	Correction
1	Élimination des caractères en doublons	➢ Tunisieeeee ➢ Sooooooleil ➢ Merciiiiiiii	➢ Tunisie ➢ Soleil ➢ Merci
2	La correction orthographique pour les erreurs fréquentes dans les forums	➢ Sa va	➢ ça va
3	La correction des mots écrits sous forme de SMS	➢ 2rl ➢ Mr6 ➢ bjr	➢ De rien ➢ Merci ➢ Bonjour

Tableau 8 : La normalisation lexicale

- La racinisation (Stemming)

La racinisation est un procédé qui sert à regrouper les mots ayant la même racine, c'est-à-dire des mots ayant une sémantique proche notamment pour les verbes à l'infinitif et les verbes conjugués. Ces mots ont le même sens mais chacun est considéré comme un descripteur à part. Dans ces cas on choisit uniquement la racine de ces mots plutôt que les mots en entier sans se soucier de l'analyse grammaticale.

Quelques exemples sont cités dans le tableau suivant :

Exemple des racines	Exemple des mots utilisés
Demander	Demande, demandé, demandant, demandent
Satisfaire	Satisfait, satisfaction, satisfis

Tableau 9 : La racinisation

L'algorithme proposé de la racinisation des commentaires :

Algorithme de racinisation

Entrée : Un commentaire

Sortie : Commentaire raciner

1. Transformer les commentaires en minuscule,
2. Séparer des caractères et des traits d'union qui existe dans le commentaire,
3. Supprimer les signes diacritiques,
4. Supprimer les doublements des lettres,
5. Convertir les mots qui sont écrits au pluriel en singulier,
6. Convertir les verbes du temps passé en temps présent,
7. Un processus vérification qui utilise un dictionnaire.

Algorithme 8: La racinisation

Étape 3 : Processus de classification

- La méthode de pondération

La première opération consiste à représenter les commentaires par une méthode algébrique de façon qu'ils peuvent être traités automatiquement par les classifieurs. Cette représentation est utilisée pour plusieurs objectifs, tels que la classification ou filtrage des informations, la recherche d'informations, l'indexation et les classements de pertinence

- La représentation vectorielle

Le modèle vectoriel (sémantique vectorielle ou Vector space model en anglais) est un modèle algébrique permettant de représenter des documents texte. Il consiste à représenter chaque document de la collection comme un point de l'espace. Les coordonnées correspondent en fait aux descripteurs composant le document. Dans la figure, trois documents sont symbolisés dans un espace à trois dimensions (chaque dimension correspondant à un terme).

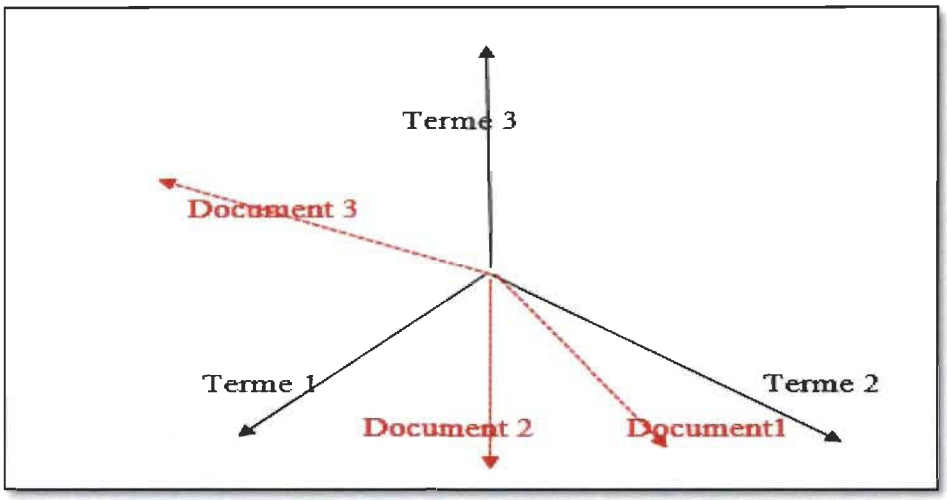

Figure 24 : Une représentation dans l'espace vectorielle avec trois termes

- Analyses des résultats à l'aide de logiciel tropes

Le logiciel Tropes est caractérise par ses analyses sémantiques des textes basées sur des techniques de fouille des données. Pierre Molette, est un co-auteur de logiciel Tropes indique en 1994 que Tropes est tout à fait différent des logiciels de lexicométrie, puisqu'il fait un appel à deux processus d'analyse (morphosyntaxique et sémantique) avant de faire des analyses statistiques. Il permet de reconstruire un réseau de liaisons sémantiques existant entre les différentes notions évoquées par l'internaute.

Ainsi le logiciel s'appuie sur différents indicateurs langagiers (verbes, adverbes, adjectifs, etc.) pour déterminer, d'une part, le style du discours, et d'autre part, les principaux univers sémantiques évoqués par les locuteurs. De plus, il sert à garantir la découverte de structures significatives d'un corpus de manière objective et à effectuer une analyse fine des énoncés discursifs.

Conséquemment, le logiciel Tropes serait idéal, pour extraire l'ossature signifiante d'une masse importante de données textuelles. Il a servi comme outil de décomposition du corpus textuel en mots-clés.

Parmi les principaux avantages de cet outil, les résultats sont présentés sous la forme de rapports ou de représentations graphiques permettant une appropriation aisée.

La figure ci-dessous représente une vue globale du processus d'analyse du logiciel Tropes:

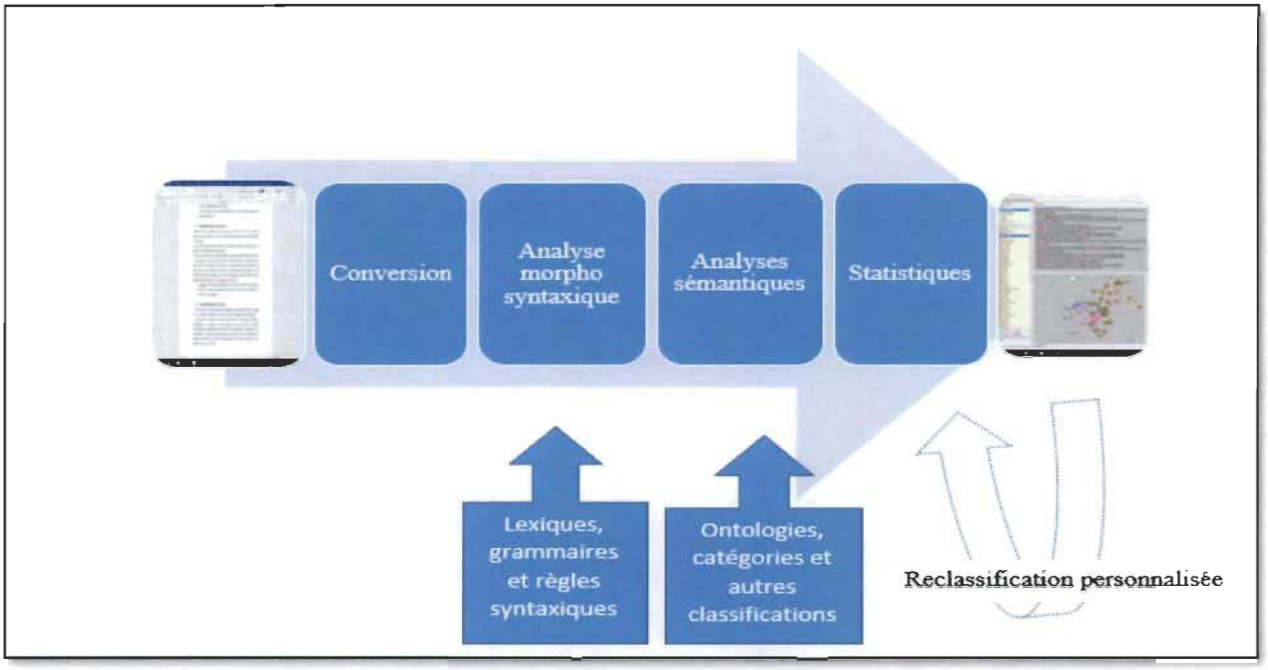

Figure 25 : Vue globale du processus d'analyse du logiciel Tropes

c) Troisième phase :

La Troisième phase consiste à transformer les résultats de l'analyse en indicateurs qui seront présentés dans un tableau de bord.

La plupart des approches proposées de visualisation des résultats ne sont pas très efficaces pour extraire des informations pertinentes qui peut servir à prendre des décisions instantanées. C'est pour cela dans notre approche on a proposé, l'élaboration d'un tableau de bord qui permet d'afficher les résultats de notre analyse. Cette technique est utilisée à la l'informatique décisionnelle (business intelligence) qui sert à une prise de décision stratégiques, tactiques et opérationnelles plus efficaces.

En d'autres termes, les tableaux de bord sont des technologies de visualisation des données ; c'est-à-dire, « des techniques de représentation graphique et d'exploration visuelle de données quantitatives permettant de traduire un ensemble de données brutes en information [...] pour améliorer la prise de décision » [Jean-Sébastien Vayre, janvier 2016].

Plusieurs formes existent pour concevoir un tableau de bord, soient des écarts, des ratios, des graphiques et autres. « Ces derniers sont utilisés dans le but d'attirer l'attention du responsable sur les informations clés pour faciliter l'analyse et le processus de décision. »

[Michel Leroy,2001]. Ainsi Selon la définition de [Stephen Few, 2006], « Un tableau de bord est une présentation visuelle des informations les plus importants nécessaires à l'atteinte d'un ou plusieurs objectifs. Ces informations sont consolidées et organisées sur un seul écran afin de pouvoir les contrôler rapidement », mais chaque tableau de bord est enrichi avec des indicateurs.

Il existe deux types de tableau de bord:

- Tableau de bord qui permet aux décideurs d'analyser et d'interpréter différemment les résultats. Dans ce cas, la connaissance est requise. Le résultat d'analyse n'est pas prédéfini dans le tableau de bord.
- Tableau de bord qui donne une information à admettre telle quelle par les décideurs. Le résultat d'analyse est prédéfini tableau de bord.

À titre d'exemple, on peut distinguer entre les deux modèles du tableau de bord dans le cas de la détection de vitesse avec l'application Waze, qui est une application mobile de navigation GPS et avec le tableau de bord d'une voiture. Ces deux figures ci-dessous illustrent la différence entre les deux :

Figure 26 : Capture d'écran tableau de bord d'une voiture

Le symbole « Check engine » dans tableau de bord de voiture est allumé. Ce témoin indique qu'il existe quelque part un problème mécanique lié au moteur, sans donner aucune autre précision. Chaque conducteur peut interpréter ce signal selon son expérience de conduite et sa connaissance de l'état mécanique de sa voiture. Ainsi, l'estimation ça diffère d'un conducteur à un autre, nous sommes donc dans le cas du premier type de tableau de bord.

Le résultat d'analyse n'est pas prédéfini. L'information n'est pas à admettre mais elle est à discuter selon l'analyste.

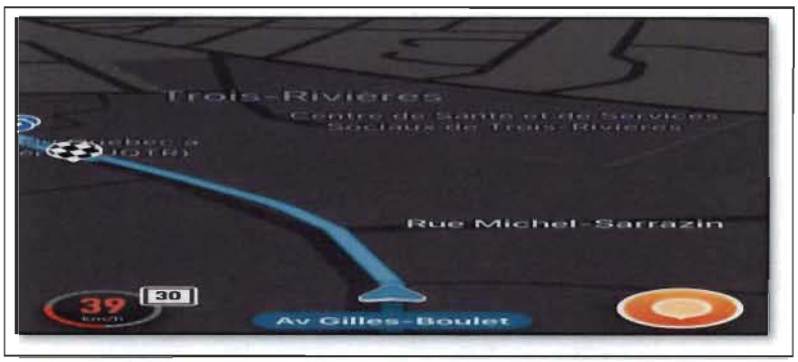

Figure 27 : Capture d'écran de l'application Waze

Waze affiche la limite de vitesse sur la portion de la route que vous empruntez. Il la met en relation avec votre vitesse actuelle. Une alerte en rouge prévient le conducteur qu'il a dépassé la limite de vitesse. Il s'agit du deuxième type du tableau de bord où le résultat d'analyse est prédéfini. L'information est à admettre.

Dans certains cas et dans le même tableau de bord, on peut trouver les deux cas :

- L'information est à admettre
- L'information n'est pas à admettre

La figure ci-dessous montre l'existence de ces deux informations. Celle de l'obligation de l'utilisation de ceinture de sécurité est en rouge et l'autre est le symbole « Check engine » qui est en jaune, ça veut dire il y a un problème mais chaque conducteur à son jugement.

Figure 28 : Capture d'écran tableau de bord d'une voiture

II. Validation de l'approche proposée

Avant d'appliquer notre approche sur le corpus des expériences de voyage en Tunisie, nous avons opter de la valider en l'appliquant manuellement sur un mini corpus des expériences des utilisateurs de le plateforme Apple de téléchargement d'applications.

1) Corpus

Le corpus est extrait de l'App Store, qui est un magasin d'applications distribué par Apple pour les appareils mobiles Iphone et Ipad . Chaque application a une page de présentation qui comporte :

- Un bouton pour l'acheter ou la télécharger
- Une interface permettant de visionner les captures d'écran du logiciel
- Les avis des utilisateurs (il est possible de poster le sien à condition d'avoir acheté l'application en question)
- Un bouton pour signaler un problème au développeur
- Un bouton pour recommander l'application par E-Mail
- Un bouton pour offrir l'application à un ami.

Après le choix du corpus, nous avons sélectionné quelques avis sur la dernière mise à jour de l'application de Facebook pour les utilisateurs d'IPhone (les avis des utilisateurs sont sélectionnés pendant la période du mois de juillet à septembre 2018). Les avis sont multilingues (espagnol, anglais et français)

Figure 29 : Capture d'écran pour un commentaire 1

Figure 30 : Capture d'écran pour un commentaire 2

Figure 31 : Capture d'écran pour un commentaire 3

Figure 32 : Capture d'écran pour un commentaire 4

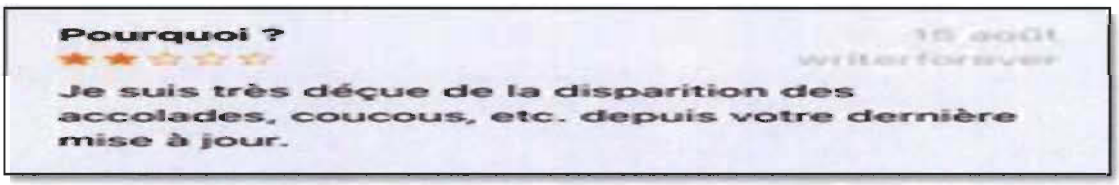

Figure 33 : Capture d'écran pour un commentaire 5

Figure 34 : Capture d'écran pour un commentaire 6

2) Élimination des textes mal écrits

Les commentaires sont souvent « bruités » en raison de l'écriture web 2.0 utilisée par les internautes (fautes d'orthographe, langage sms, etc.). En effet, les textes des médias sociaux ont des particularités linguistiques qui peuvent affecter la performance des classifieurs » [Mikeet al, Juin 2017].

Le principal but d'élimination des textes mal écrits est d'améliorer les résultats d'analyse et de réduire le temps d'analyse. Ci-après les commentaires supprimés car ils n'ont aucune signification ou sens.

Figure 35 : Capture d'écran d'un commentaire qu'a été supprimé

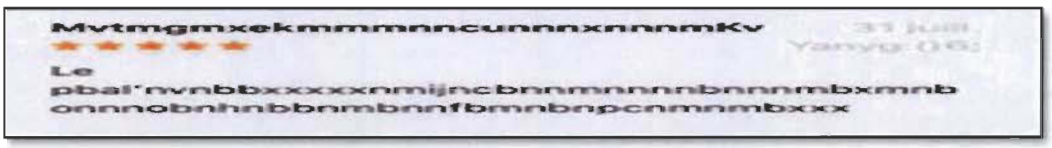

Figure 36 : Capture d'écran d'un commentaire qu'a été supprimé

3) Traduire les textes en français

« L'analyse sémantique des conversations portant sur un événement qui se déroule la même journée ou de la même semaine sur les réseaux sociaux, ou d'un ensemble de discussions sur un sujet apparenté se bute aux difficultés soulevées par les aspects multilingues du TAL. Cette spécificité est accrue avec les réseaux sociaux qui peuvent mêler différentes langues dans un même contenu. » [Atefeh Farzindar, Mathieu Roche, Mai 2015]

Pour résoudre le problème de multilinguisme (sur un même flux, dans un seul commentaire), on a ajouté l'étape de traduction des textes. Cette étape est primordiale pour l'analyse des textes et l'extraction d'information.

- **Détection de la langue :**

Ce qui est remarquable dans le corpus, il n'existe pas de multilinguisme dans les commentaires. Les langues détectées sont :

- L'anglais : Deux commentaires
- Le français : Un commentaire
- L'espagnol : Un commentaire

- **Traduction manuelle :**

L'anglais vers le français :

- Quand j'édite un effet photo, l'application se ferme rapidement.
- Cela prend une éternité pour charger les photos et les commentaires. Cette application est de pire en pire

L'espagnol vers le français :

- Quand je veux commenter quelque chose, il me dit que je ne peux pas.

4) Analyse sémantique et statistique des textes

L'analyse sémantique des médias sociaux a ouvert la voie à l'analyse de données volumineuses, discipline émergente inspirée de l'apprentissage automatique, de l'exploration de données, de la recherche documentaire, de la traduction automatique, du résumé automatique et du TAL plus globalement. » [Atefeh Farzindar, Mathieu Roche, Mai 2015].

Du point de vue marketing, il est important pour les responsables de la clientèle d'Apple de savoir les opinions de leurs utilisateurs pour améliorer leur expérience et garantir leur satisfaction. Du point de vue informatique, il est important pour les développeurs d'Apple de connaître les anomalies ou les points fort d'une application pour cibler les besoins des utilisateurs dans les développements futurs, pour corriger les erreurs et pour améliorer l'application. Positives, négatives ou neutres, les réactions des utilisateurs représentent une source importante de données qui devraient être analysées pour mieux comprendre l'expérience utilisateur et ajuster les applications selon les besoins et selon les tendances sur le marché.

- **L'étape de prétraitement**

Commentaire	Quand je veux commenter quelque chose, il me dit que je ne peux pas
La tokenisation	« Quand», « je», « veux», « commenter», « quelque», « chose», « il», « me», « dit», « que», « je», « ne», « peux», « pas».
Mots vides (stop-word)	« veux», « commenter», « chose», « dit», « peux».
Normalisation lexicale	« veux», « commenter», « chose», « dit», « peux»
La racinisation	« venir», « commenter», « chose», « dire», « pouvoir»

Tableau 10 : Prétraitement pour le commentaire 1

$$\% \text{ Mv} = = \frac{100 * \text{NMs}}{\text{NMt}} = \frac{100 * 9}{14} = 60{,}28\%$$

Avec **Mv** : Mots vides, **NMs** : Nombre des mots vides, **NMt** : Nombre des mots totaux

Commentaire	Quand j'édite un effet photo, l'application se ferme rapidement
La tokenisation	« Quand», « je», « édite », « un», « effet», « photo », « la », « application », « se », « ferme », « rapidement ».
Mots vides (stop-word)	« édite », « effet», « photo », « application », « ferme », « rapidement ».
Normalisation lexicale	« édite », « effet», « photo », « application », « ferme », « rapidement »
La racinisation	« éditer », « effet», « photo », « application », « fermer », « rapide ».

Tableau 11 : Prétraitement pour le commentaire 2

$$\% \mathbf{Mv} = = \frac{100 * \mathbf{NMs}}{\mathbf{NMt}} = \frac{100 * 5}{11} = 45{,}45\%$$

Avec **Mv** : Mots vides, **NMs** : Nombre des mots vides, **NMt** : Nombre des mots totaux

Commentaire	Cela prend une éternité pour charger les photos et les commentaires. Cette application est de pire en pire
La tokenisation	« Cela », « prend», « une», « éternité», « pour», « charger», « les», « photos», « et», « les», « commentaires», « cette», « application», « est», « de», « pire», « en», « pire».
Mots vides (stop-word)	« prend», « éternité», « charger», « photos», « commentaires», « application», « pire», « pire».
Normalisation lexicale	« prend», « éternité», « charger», « photos», « commentaires», « application», « pire», « pire»
La racinisation	« prendre», « éterniser», « charger», « photos», « commenter», « application», « pire», « pire»

Tableau 12 : Prétraitement pour le commentaire 3

$$\% \mathrm{Mv} = = \frac{100 * \mathrm{NMs}}{\mathrm{NMt}} = \frac{100 * 10}{18} = 55{,}55\%$$

Avec **Mv** : Mots vides, **NMs** : Nombre des mots vides, **NMt** : Nombre des mots totaux

Commentaire	Je suis très déçue de la disparition des accolades, coucous, etc. depuis votre dernière mise à jour
La tokenisation	« je», «suis», «très», «déçue», «de», «la», «disparition», «des», «accolades», «coucous», «etc», «depuis», «votre», «dernière», «mise», «à», «jour»
Mots vides (stop-word)	«déçue», «disparition», «accolades», «coucous», «dernière», «mise», «jour»
Normalisation lexicale	«déçue», «disparition», «accolades», **«coucou»,** «dernière», «mise», «jour»
La racinisation	«décuver», «disparaitre», «accolades», **«coucou»,** «dernière», «mettre», «jour»

Tableau 13 : Prétraitement pour le commentaire 4

$$\% \, \mathbf{Mv} = = \frac{100 * \mathbf{NMs}}{\mathbf{NMt}} = \frac{100 * 10}{17} = 58{,}82\%$$

Avec **Mv** : Mots vides, **NMs** : Nombre des mots vides, **NMt** : Nombre des mots totaux

Après l'étape de prétraitement, on obtient ce corpus :

venir commenter chose dire pouvoir
éditer effet photo application fermer rapide .
prendre éterniser charger photos commenter application pire pire
décuver disparaitre accolade mettre jour

5) Mise en place d'un tableau de bord

L'analyse du corpus révèle que les utilisateurs ne sont pas satisfaits à cause de la dernière mise à jour de Facebook. Ce désagrément est dû à des problèmes qui surviennent lors de la modification ou de la publication des nouvelles photos et par conséquent l'application reste figée (bug). De plus, un autre problème a paru et les utilisateurs ne peuvent même pas commenter des photos nouvelles. Nos deux indicateurs sont « commentaire » et « Photo ».

Le tableau de bord proposé pour notre cas est celui-ci :

Commentaire	
Photo	

Figure 37: Le tableau de bord des avis d'utilisateur concernant la dernière mise à jour de Facebook

 Mécontent.

 Satisfait

 Neutre

III. Conclusion

Notre approche a été présentée de manière détaillée. Elle se base sur le développement et la réutilisation des technologies existantes et consiste à les rendre complémentaires les unes des autres permettant ainsi l'extraction de l'information à partir de données non structurées. Pour cela, nous avons proposé des algorithmes que nos premières évaluations indiquent qu'ils sont efficaces.

Le chapitre suivant contient les tests et les expérimentations de validation de notre approche.

CHAPITRE VIII : EXPERIMENTATION ET ANALYSE DES RESULTATS

I. Expérimentation et analyse des résultats

1) Description du corpus

Dans le but d'analyser le tourisme tunisien et déterminer certains éléments clés qui le caractérisent, nous avons opté pour l'analyse de commentaires de certains touristes décrivant leur expérience ou posant des questionnements pour préparer leur voyage en Tunisie. Notre choix de la plateforme s'est porté à la plateforme Web TripAdvisor qui représente une des plus importantes tribunes des opinions de voyageurs au monde. En effet, cette plateforme Web a été créée au début des années 2000. Le site TripAdvisor.com présente des avis et des conseils touristiques émanant de consommateurs sur des hôtels, des restaurants et des villes. Outre les commentaires des voyageurs, il offre également des liens vers des informations correspondant à des articles de journaux, de magazines ou de guides de voyages ainsi qu'un forum accessible à ses membres. En effet, le site TripAdvisor.com « accueille plus de 315 millions de visiteurs uniques chaque mois et recueille plus de 500 millions d'avis et d'opinions » [Ryan Saad, Juin 2017].

Effectivement, nous avons sélectionné les avis des touristes de nationalités différentes ayant visité ou se préparaient pour visiter la Tunisie pendant la période entre 2007 et 2017 en tenant compte de la chronologie des commentaires.

Le tableau ci-après décrit les détails du corpus qui sera utilisé pour nos tests:

Nom	Taille
Pages	132
Mots	27880
Caractères (espace non compris)	130839
Caractères (espace compris)	157273
paragraphes	2805
ligne	4190
fichier	1200 Ko

Tableau 14 : Taille du corpus d'entrainement

2) L'analyse du corpus et discussion des résultats

a) Les outils utilisés

Les outils proposés ont été choisis en fonction des éléments suivants :

- Convivialité des interfaces et facilité d'interaction et d'intégration;
- Bonne côte dans la littérature [Stéphane Goria, 2018; Chloé et al, 2018; Marli et al, 2018]

➢ Tropes V8.4.2c

Le logiciel Tropes a été créé en 1994 par Pierre Molette et Agnès Landré. C'est un logiciel d'analyse sémantique et également pour faire du text mining. Il se base sur les travaux de Rodolphe Ghiglione. Ce logiciel permet de déterminer, au sein d'un ou plusieurs textes, qui sont les acteurs principaux, quelles sont les relations qui les lient. Il permet également de ressortir le sens global du texte. Parmi les fonctions proposées par Tropes, on peut citer :

- La classification arborescente de la référence,
- L'analyse chronologique du récit,
- Le diagnostic du style du texte,
- La catégorisation des mots-outils,
- L'extraction terminologique,
- L'analyse des acteurs,
- La constitution de résumés

➢ IRaMuTeQ V0 .7 alpha 2

Iramuteq a été créé en 2015. C'est un logiciel de traitement de données pour des corpus texte. Il est puissant pour travailler avec des données désordonnées. Il dispose d'un outil bien pratique permettant de nettoyer les textes. Il fonctionne en interface avec le langage R, son fonctionnement consiste à :

- Préparer les données;
- Écrire des scripts qui sont ensuite analysés dans le logiciel statistique R.

b) L'étape préalable

- Suppression des publicités et des images :

Après sélection du corpus, nous procédons à l'élimination des images et textes publicitaires pour réduire considérablement notre corpus.

- Suppression des commentaires non pertinents

Une première analyse des commentaires, nous a permis d'éliminer les commentaires jugés non pertinents et qui ne rajoutent pas d'informations à notre analyse. Effectivement, parfois les modérateurs du site Web éliminent un commentaire pour non-respect des règles. Les trois figures ci-dessous représentent quelques exemples des commentaires ignorés :

2. **Re : pas le moment d'y aller**
31 oct. 2011 14:51

-:- Message de l'équipe TripAdvisor -:-

Les modérateurs ont supprimé ce message parce qu'il ne respectait pas les règles d'utilisation des forums de TripAdvisor qui interdisent la publicité ou les sollicitations.

Nous demandons à tous nos utilisateurs de ne pas faire de publicité pour leurs produits/services, ni de sollicitation d'aucune sorte dans leurs messages sur nos forums. Les personnes affiliées à une entreprise touristique ne sont pas autorisées à mettre un lien vers un site Internet ni à donner leurs coordonnées dans leurs messages.

Vous pouvez consulter nos règles d'utilisation des forums en cliquant sur ce lien : http://www.fr.tripadvisor.ca/pages/forums_posting_guidelines.html

Nous supprimons les messages qui ne respectent pas les règles d'utilisation et nous nous réservons le droit de supprimer tout message, quelle qu'en soit la raison.

Supprimé le : 01 novembre 2011, 06:02

Figure 38 : Commentaire non sélectionné

30. **Re : Le Maroc mieux que la Tunisie ?**
28 juin 2012 18:13

-:- Message de l'équipe TripAdvisor -:-

L'équipe TripAdvisor a supprimé ce message à la demande de son auteur ou parce qu'il ne respectait pas les règles du forum de TripAdvisor.

Vous pouvez consulter nos règles d'utilisation des forums en cliquant sur ce lien : http://www.fr.tripadvisor.ca/pages/forums_posting_guidelines.html

Nous supprimons les messages qui ne respectent pas les règles d'utilisation et nous nous réservons le droit de supprimer tout message, quelle qu'en soit la raison.

Supprimé le : 29 juin 2012, 04:13

Figure 39 : Commentaire non sélectionné

Figure 40 : Commentaires non sélectionnés

À la fin de cette étape, notre corpus est réduit considérablement tel qu'illustré dans le tableau suivant. Il est à préciser que notre corpus a passé de 132 pages à 66 pages, ce qui montre l'importance de l'étape préalable que nous avons rajouté à notre démarche. Effectivement, cette étape aide à réduire le temps de l'étape suivante de prétraitement du processus habituel d'analyse qui représente un grand défi pour les analystes.

Nom	Taille
Pages	66
Mots	245387
Caractères (espace non compris)	111350
Caractères (espace compris)	134489
paragraphes	1351
ligne	2501
fichier	158 Ko

Tableau 15 : La taille du nouveau corpus d 'entrainement

- La traduction des commentaires

La deuxième tâche de l'étape préalable consiste à unifier la langue en traduisant toutes les autres langues en une seule langue pour permettre d'effectuer une analyse textuelle monolingue. Dans notre cas, nous traduisons les commentaires qui sont écrits avec une langue autre que le français et les commentaires multilingues au français pour obtenir les meilleurs résultats possibles.

Selon [Philipp Koehn, 2010], « la traduction automatique est dans ce cas un premier pas vers l'extraction d'information et l'analyse de grands volumes de textes étrangers.»
Certains commentaires en français comportent des mots écrits dans une langue autre que le français. Nous avons procédé à la traduction de ces mots conformément au tableau suivant :

Mot détecté	Mots traduit
Beach	Plage
Swimming pool	Piscine
Stay	Séjour
Safe	Sécurité
Incredible	Incroyable
Seaside	Balnéaire
Good	Bien
Fear	Peur
South	Sud
Uncertainty	Incertitude
Happy	Content
Sad	Triste
Holiday	Vacance
City	Ville
Unhappy	Mécontent

Tableau 16 : Exemple des mots qui sont traduits dans notre corpus

Certains commentaires sont écrits complètement dans une langue autre que le français. Nous avons fait recours au logiciel DeepL pour les traduire (voir figures ci-après)

- Un commentaire rédigé en anglais

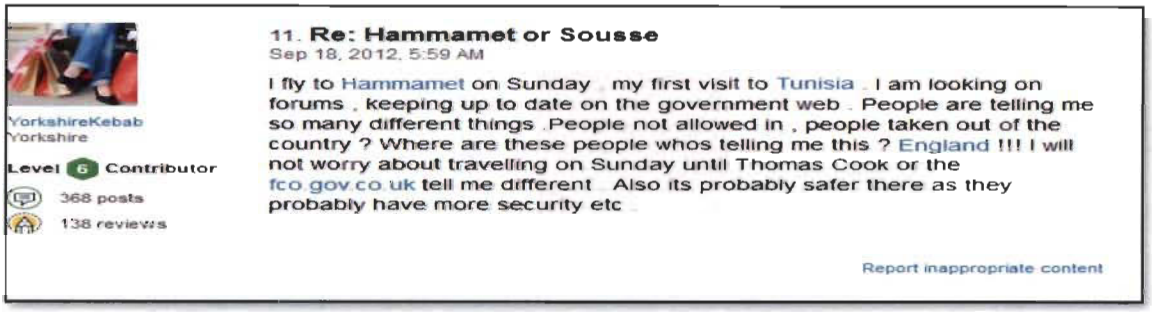

Figure 41 : Capture d'écran d'un commentaire en anglais

- La traduction avec Deepl

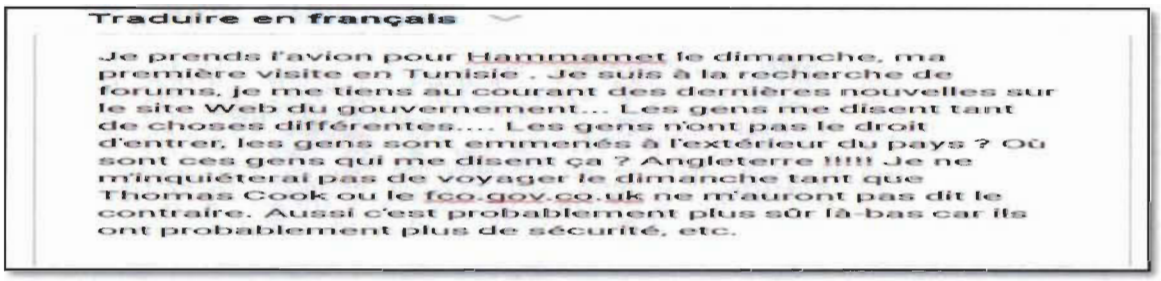

Figure 42: Capture d'écran d'un commentaire traduit à l'aide de Deepl

- Un commentaire rédigé en anglais

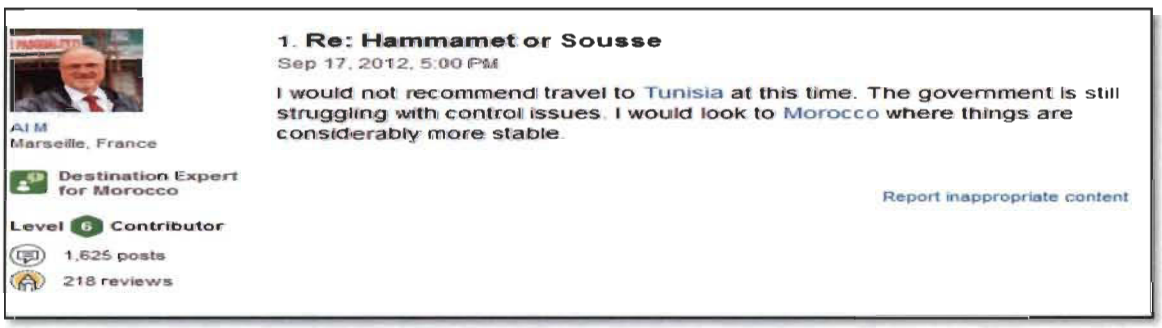

Figure 43 : Capture d'écran d'un commentaire en anglais

- La traduction avec Deepl

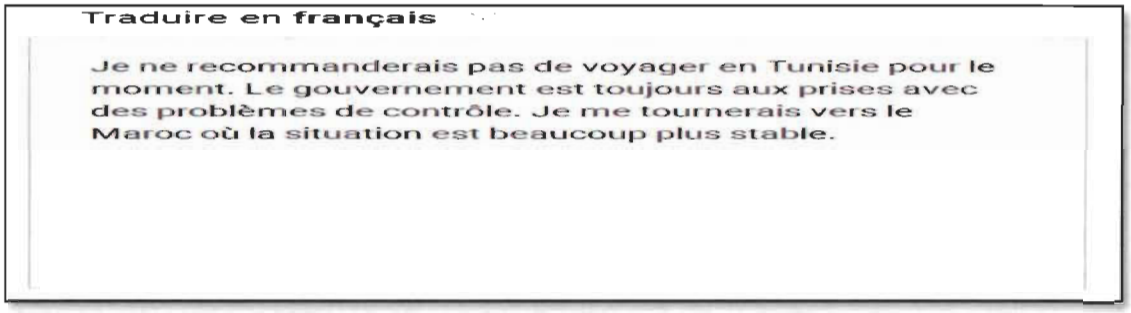

Figure 44 : Capture d'écran d'un commentaire traduit à l'aide de Deepl

- Un commentaire rédigé en anglais

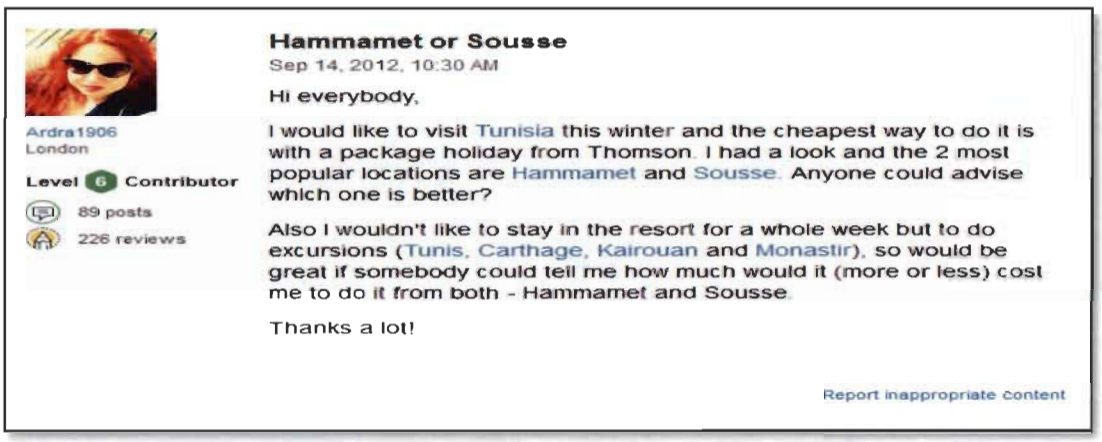

Figure 45 : Capture d'écran d'un commentaire en anglais

- La traduction avec Deepl

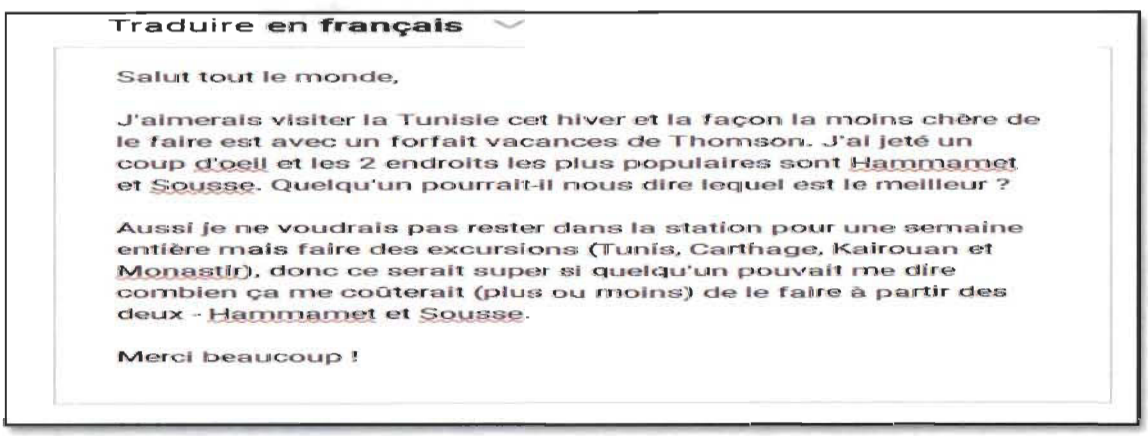

Figure 46 : Capture d'écran d'un commentaire traduit à l'aide de Deepl

Tel que précisé précédemment, encore une fois l'étape préalable nous permet de garantir une meilleure analyse des textes en unifiant la langue. L'étape de la traduction a permis de bénéficier des commentaires écrits dans une langue autre que le français et enrichir par conséquent les résultats de notre analyse. À titre d'Exemple, notre corpus contenait 6 pages en anglais qu'on aurait pu les perdre si l'exercice de traduction n'a pas été fait. .

Nom	Taille
Pages	6
Mots	1366
Caractères (espace non compris)	6063
Caractères (espace compris)	7364
paragraphes	122
ligne	167

Tableau 17 : La taille des commentaires anglais dans le corpus

À la fin de la phase préalable où nous avons éliminé les textes non pertinents et nous avons unifié la langue, nous avons appliqué le logiciel Iramuteq sur notre corpus afin de déterminer le nuage des mots clés de notre corpus. L'image ci-après décrit le résultat de cette action qui représente une sorte de condensé sémantique d'un document présentant l'occurrence des mots-clés contenus dans notre corpus.

Figure 47 : Nuage de mots-clés de corpus avant l'étape de prétraitement

À la lumière de ce résultat, nous avons constaté que notre corpus contient plusieurs mots vides tels que (de, les, et, la, des, du…). Ces mots vides seront éliminés dans la phase de prétraitement du processus habituel d'analyse des données.

c) Processus habituel d'analyse des données

Cette étape comporte trois phases, à savoir le prétraitement, l'analyse avec le logiciel tropes et l'interprétation des résultats.

➢ **Le prétraitement :**

Le prétraitement implique les quatre étapes suivantes : la tokenisation, la suppression des mots vides, la normalisation lexicale et la racinisation. Dans la démarche proposée précédemment, nous avons déterminé trois algorithmes pour les étapes de tokenisation, de suppression des mots vides et de racinisation.

Dans notre cas, nous avons appliqué le logiciel Iramuteq qui a exécuté conjointement les trois étapes sus-indiquées afin de nettoyer les commentaires. L'application de ce logiciel se résume comme suit :

- Transformer les lettres majuscules en minuscule pour éviter que le logiciel considère le même mot qui commence une fois par une majuscule et parfois par une minuscule comme étant deux formes distinctes. En effet, le logiciel Iramuteq offre l'option de transformer toutes les lettres majuscules en minuscules.
- Éliminer les caractères en dehors d'une liste prédéfinie : Dans notre cas, nous avons défini une liste des lettres en français faisant partie de l'alphabet français ainsi que leur variation. À titre d'exemple pour la lettre « a », nous avons déterminé la liste suivante « a, à, â, ä ». Pour la lettre « e », nous avons déterminé la liste suivante « e, é, è, ê ». Ainsi, notre liste finale compte 39 caractères et se présente comme suit : « a, à, â, ä, b, c, d, e, è, é, ê, f, g, h, i, î, ï, j, k , l ,m, n, o, ô, ö, œ, p, q, r, s, t, u, û, ù, v, w, x, y, z ».
- Remplacer les apostrophes et les tirets par des espaces.

La figure ci-après décrit les différentes options qu'offre le logiciel pour nettoyer notre corpus.

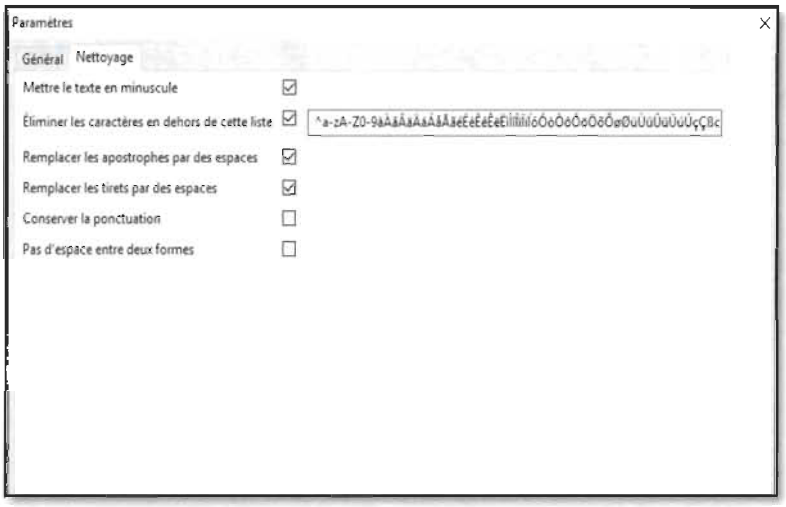

Figure 48 : Paramètres d'actions du logiciel

- Raciner le corpus : Le logiciel offre une étape de lemmatisation qui consiste à transformer les verbes à l'infinitif, les adjectifs au masculin singulier et les noms au singulier.
- Identifier les clés d'analyses : le logiciel offre une variété de clés d'analyses. Avec les différentes barres de défilement, nous avons choisi l'ensemble de nos critères qui répondent à nos besoins d'analyse.
- Choisir le dictionnaire : le logiciel nous offre de choisir la langue d'analyse. Dans notre cas, c'est le français.

Les figures ci-après présentent les étapes suivies lors de l'utilisation du logiciel.

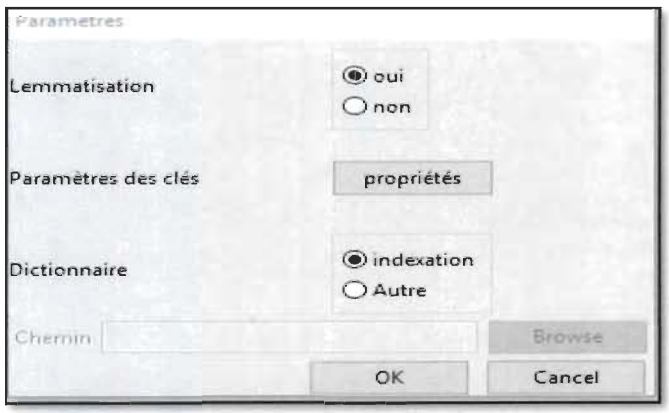

Figure 49 : Nettoyage des données à l'aide de logiciel IRaMuTeQ

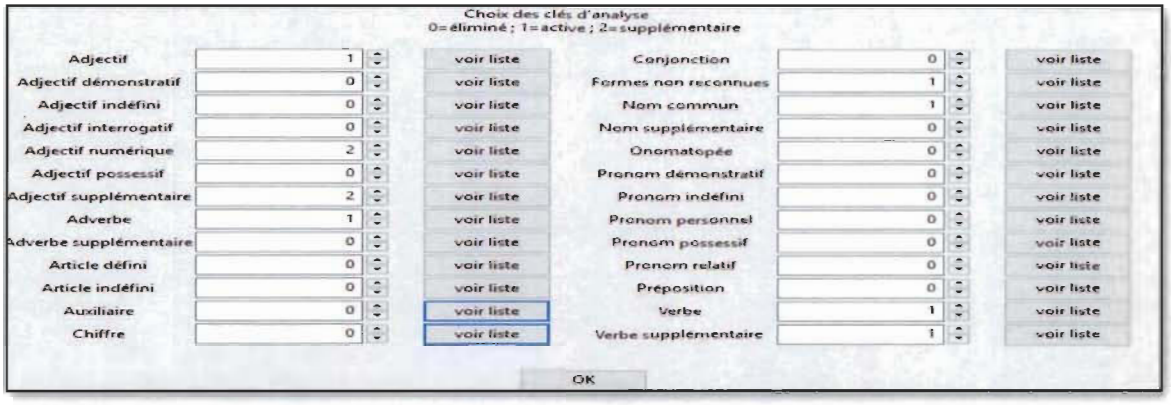

Figure 50 : Nettoyage des données à l'aide de logiciel IRaMuTeQ

À l'issue de cette étape, notre corpus est plus réduit en nombre de pages, il est à l'état brut (sans mots vides, sans chiffres, sans apostrophes, sans tirets, verbe à l'infinitif, adjectifs au masculin singulier, noms au singulier). Ci-après une image de l'état de notre corpus après l'étape de prétraitement.

Figure 51 : Nuage de mots-clés de corpus après l'étape de prétraitement

- **L'analyse du corpus**

- Les références les plus fréquentes dans le corpus

On commence par l'extraction de la liste des substantifs proposée par Tropes dans notre corpus. Dans cette liste, le terme « Tunisie » apparaît comme étant le plus fréquent. Le tableau ci-dessous, nous permet de répondre à la question « Sur quoi porte notre corpus ? ». Il présente l'information par rapport à l'occurrence des mots. On remarque aussi que ces

termes sont liés à notre but de recherche qu'on a défini dès le début qui est de déterminer les caractéristiques et les éléments clés du tourisme tunisien sur lesquelles il faut agir pour ramener ce secteur à son niveau avant la révolution.

Terme	Occurrence
Tunisie	435
Vacance	66
Tourisme	65
Séjour	52
Remerciement	48
Hôtel	47
Sécurité	33
Changement	23
Problème	19
Otage	14

Tableau 18 : Les références les plus fréquentes dans le corpus

- L'apparition des pronoms personnels dans le corpus

Pour distinguer les différents intervenants dans un texte et dégager leur point de vue, le repérage des pronoms personnels et de leur fonction représente une technique efficace pour déterminer qui parle à qui et qui dit quoi. Tropes permet d'extraire les pronoms personnels et leur pourcentage d'apparition.

Pronoms	Appariation
Je	34,1 % (426)
Tu	3,2 % (40)
Nous	12,8 % (160)
Vous	18,2 % (227)
On	8,3% (104)

Tableau 19 : L'apparition des pronoms personnels dans le corpus

Ce tableau nous permet de constater que l'emploi du pronom « je » est le plus dominant. Le pronom « nous » arrive en troisième position. L'emploi de ces deux pronoms personnels

(je, nous) caractérisent un corpus argumentatif qui discute, qui compare et qui critique, ce qui montre que le narrateur devient un personnage qui prend parti à l'histoire et il révèle un regard intérieur (subjectif).

- Les verbes, les adjectifs et les noms utilisés dans le corpus

Généralement, les voyageurs recourent à l'utilisation des verbes, des adjectifs et des noms pour exprimer leur état de satisfaction, leur sentiment ou leur opinion personnelle. Les tableaux ci-après présentent un extrait de certains verbes et adjectifs ainsi que le nombre de leur apparition.

Verbe	Apparition
Falloir	60
Décevoir	14
Éviter	13
Aimer	12
Changer	10
Craindre	10
Signaler	9
Remercier	8
Apprécier	8
Déconseiller	6
Inquiéter	5

Tableau 20 : Liste de certains verbes utilisés dans le corpus

Adjectif	Apparition
Accueillant	10
Mauvais	10
Sale	7
Désagréable	5

Tableau 21 : Liste de certains adjectifs utilisés dans le corpus

On peut remarquer que les verbes et les adjectifs d'incertitude, d'hésitation et de mécontentement sont les plus dominants dans le corps. Ce qui caractérise l'état du tourisme

tunisien après la révolution. À titre d'exemple, ci-après deux figures représentants les occurrences du verbe décevoir et de l'adjectif mauvais.

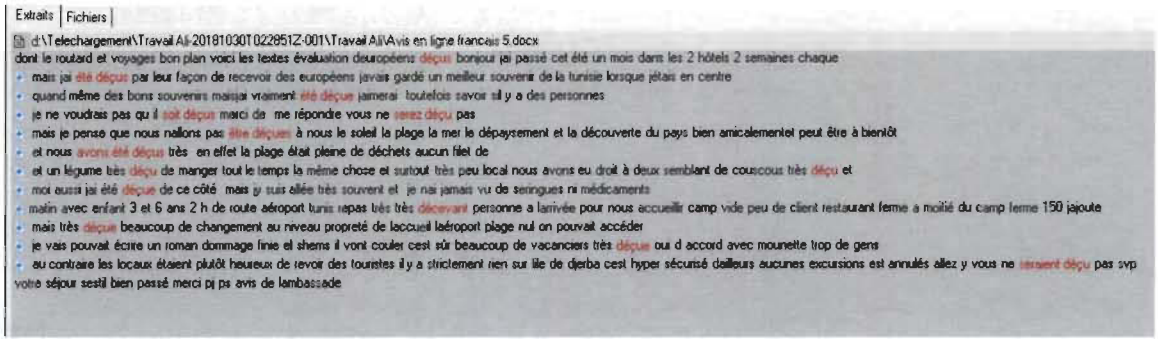

Figure 52 : Les occurrences du verbe décevoir

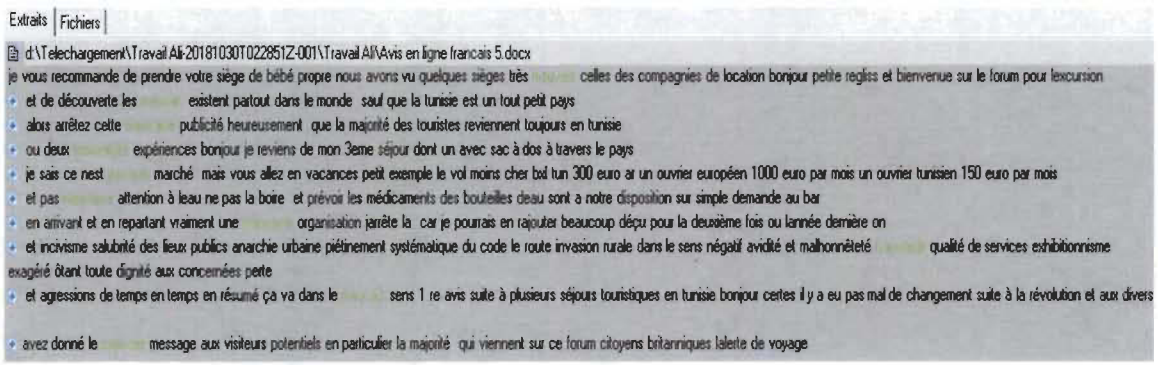

Figure 53 : Les occurrences du l'adjectif mauvais

Une analyse des noms pour les sentiments négatifs est très importante dans notre cas. Cette analyse prend en charge le substantif de sécurité et ses synonymes (peur, révolution, otage, risque, attentat…)

Substantif	Apparition
Sécurité	32
Peur	21
Révolution	17
Otage	14
Risque	8
Attentat	7

Figure 54 : Les substantifs de sécurité dans le corpus

On constate d'après le tableau et les graphes de répartition ci-dessous, les différents contextes dans lequel apparaissent les synonymes de sécurité dans notre corpus

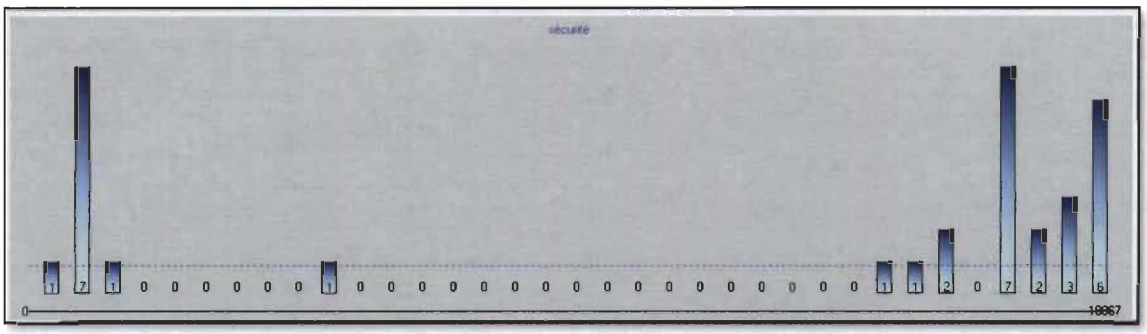

Figure 55 : L'apparition de substantif « sécurité » dans le corpus

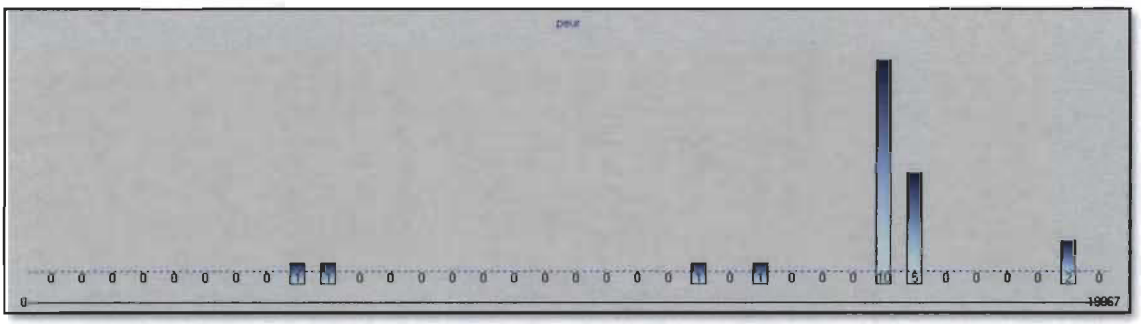

Figure 56 : L'apparition de substantif « peur » dans le corpus

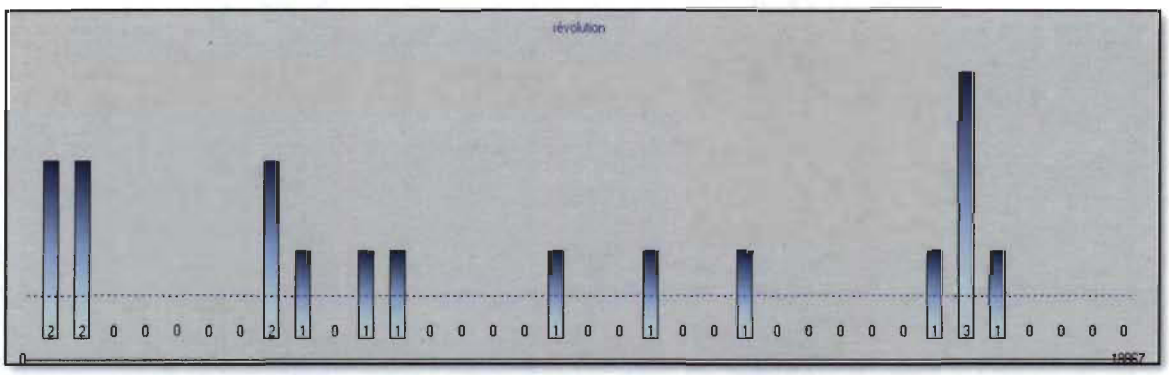

Figure 57 : L'apparition de substantif « révolution » dans le corpus

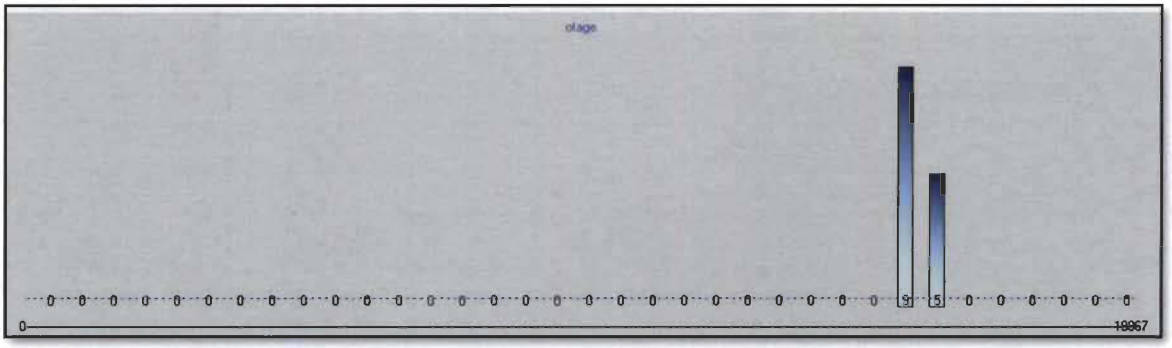

Figure 58 : L'apparition de substantif « otage » dans le corpus

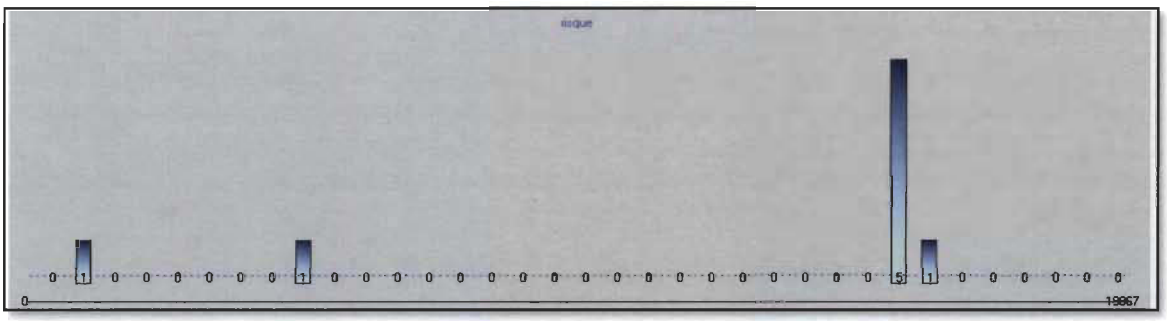

Figure 59 : L'apparition de substantif « risque » dans le corpus

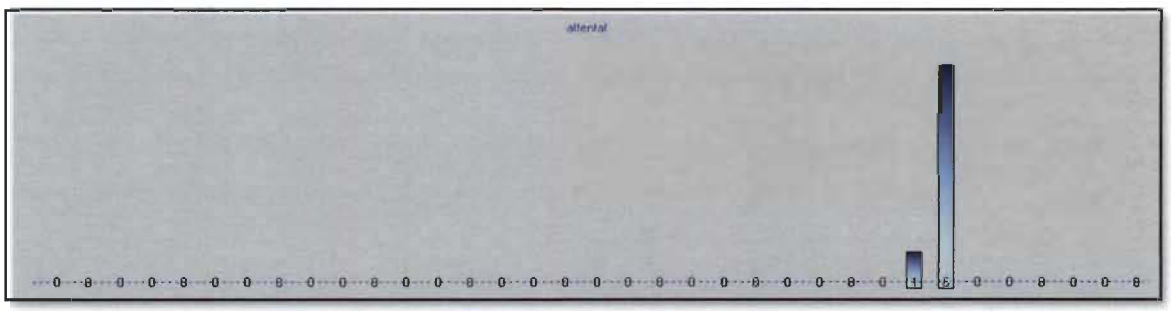

Figure 60 : L'apparition de substantif « attentat » dans le corpus

Ces graphes nous affichent un histogramme de répartition des substantifs (sécurité, peur, révolution, otage, risque, attentat) dans le corpus. Ils l'ont divisé en secteurs contenant un nombre égal de mots et en calculant la fréquence d'apparition de ces substantifs à l'intérieur de chaque secteur.

Les barres de l'histogramme affichent chaque secteur dans l'ordre chronologique, de gauche début du corpus jusqu'à droite (fin du corpus). La ligne en pointillés indique la taille moyenne des barres de l'histogramme. Effectivement, notre campus est divisé en trois périodes chronologiques :

- La première période : Début de la révolution

D'après le tableau ci-dessous, on peut constater une apparition importante de substantif de sécurité. Cette appariation est due au déclenchement de la révolution tunisienne.

Substantif	Apparition
Sécurité	9
Peur	0
Révolution	6
Otage	0
Risque	1
Attentat	0

Tableau 22 : L'apparition des substantifs de sécurité au début de corpus

- Deuxième période : Après la révolution

D'après le tableau ci-dessous, on remarque une faible apparition des substantifs de sécurité (11 sur 99).

Substantif	Apparition
Sécurité	1
Peur	3
Révolution	6
Otage	0
Risque	1
Attentat	0

Tableau 23 : L'apparition des substantifs de sécurité au milieu de corpus

- Troisième période : Après les attaques terroristes

D'après les figures précédentes des histogrammes et aussi le tableau ci-dessous, on constate une apparition très importante des substantifs de sécurité à la fin de corpus Elle représente

presque 72 %. Cette apparition est due à deux évènements qui ont influencé le tourisme de tunisien. Le premier évènement est arrivé le 18 mars 2015 où la Tunisie a été victime d'une attaque sanglante et de prise d'otages au musée Bardo. Il a eu 24 personnes décédées et 45 blessées. Le deuxième évènement, c'est l'attentat qui a lieu dans une station balnéaire à Sousse le 26 juin 2015.

Substantif	Apparition
Sécurité	22
Peur	18
Révolution	5
Otage	14
Risque	6
Attentat	7

Tableau 24 : L'apparition des substantifs de sécurité à la fin de corpus

- Analyse graphique des corrélations dans le corpus

L'analyse graphique est une bonne manière pour comprendre les différentes caractéristiques énumérées dans le corpus. Les graphiques tels que « graphe en sphère », « graphe en étoile » et « le graphe de ligne pointillée » sont des outils privilégiés. On a choisi les références les plus fréquentes dans le corpus (sécurité, inquiétude, peur et appréhension, terrorisme, otage) pour déterminer ses corrélations avec les autres références.

- Les graphes en sphères

Ces types de graphes serrent à analyser l'environnement d'une référence ou d'une catégorie pour déterminer la corrélation entre eux.

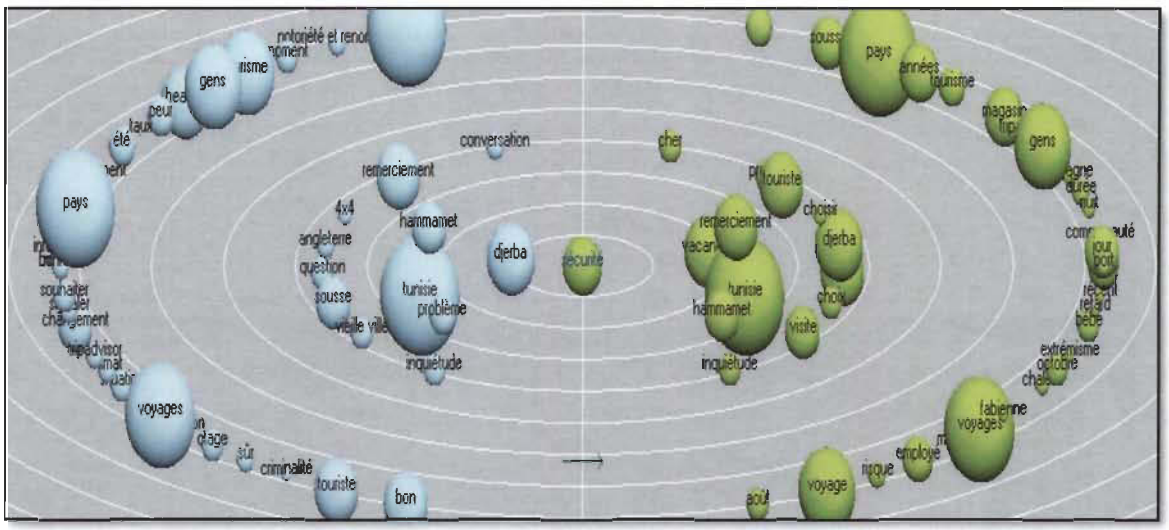

Figure 61 : Le graphe de sphère pour la référence Sécurité

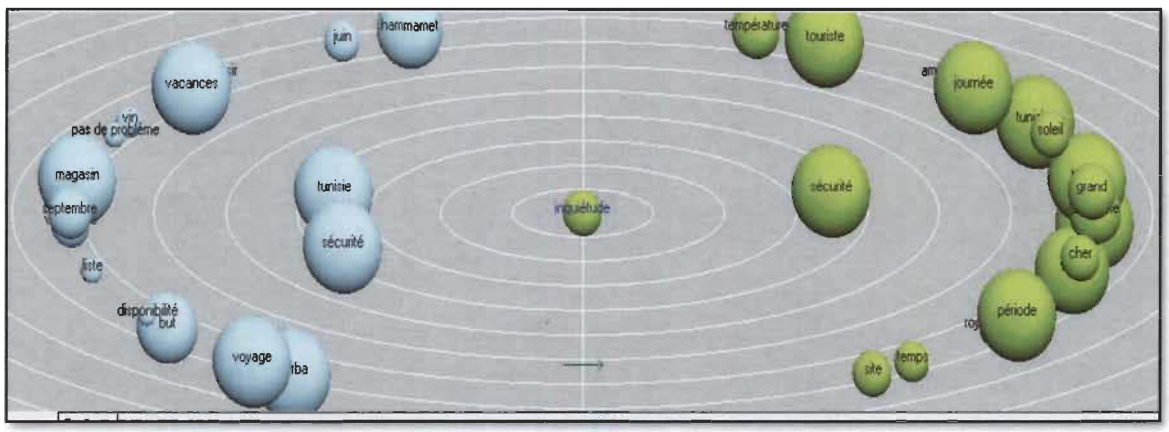

Figure 62 : Le graphe de sphère pour la référence Inquiétude

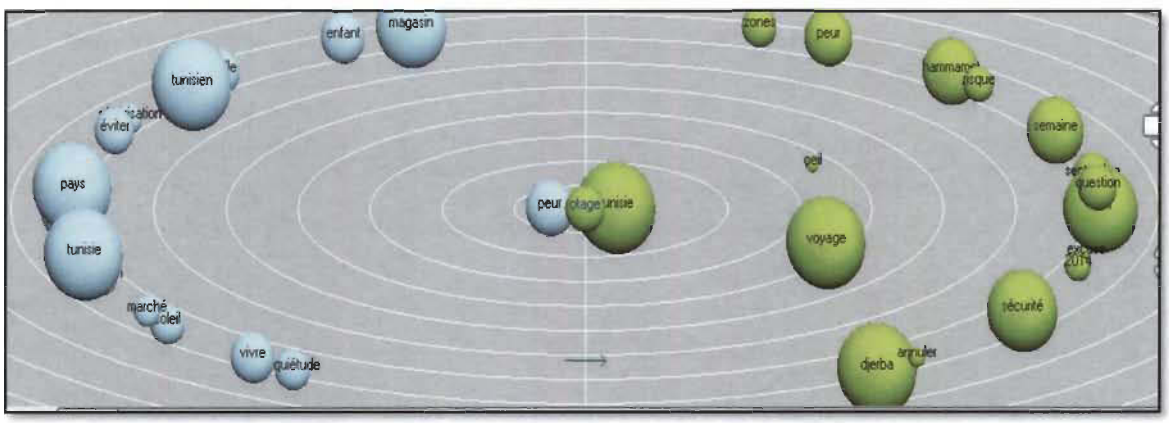

Figure 63 : Le graphe de sphère pour la référence Otage

Sur ces graphes, les références (sécurité, inquiétude et otage) sont représentées par une sphère au centre. La distance entre ces références et les autres est proportionnelle au nombre de relations qui les lient : autrement dit, lorsque deux références sont proches elles ont beaucoup de relations en commun, et lorsque qu'elles sont éloignées elles n'ont que peu de relations en commun. Ce qui se dégage de ces figures, il existe une forte corrélation entre la peur, l'otage, la sécurité et la Tunisie.

- Les graphes en étoile

Le graphe en étoile permet d'afficher les relations entre les références. La quantité de relations (fréquence de cooccurrence) qui existent entre les deux références sera indiquée par les nombres qui apparaissent sur le graphe.

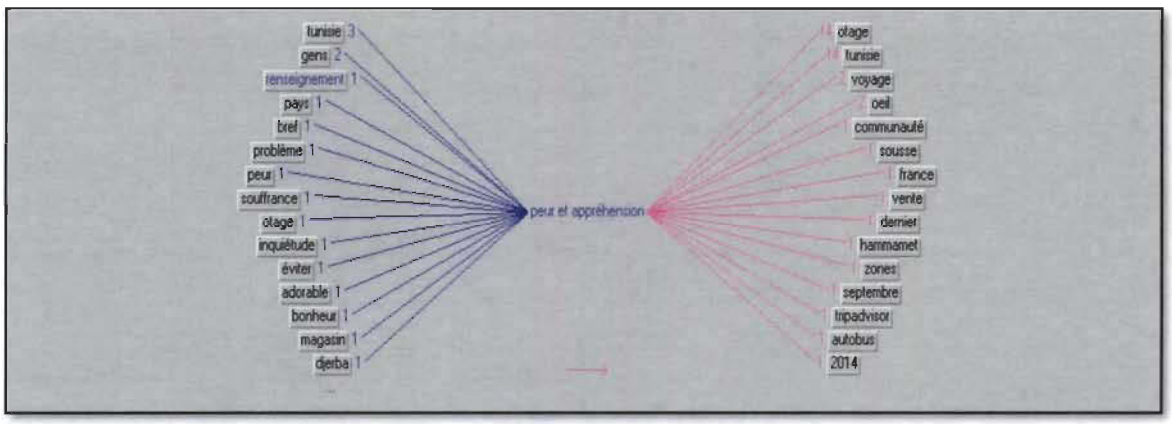

Figure 64 : Le graphe en étoile pour la référence Peur et appréhension

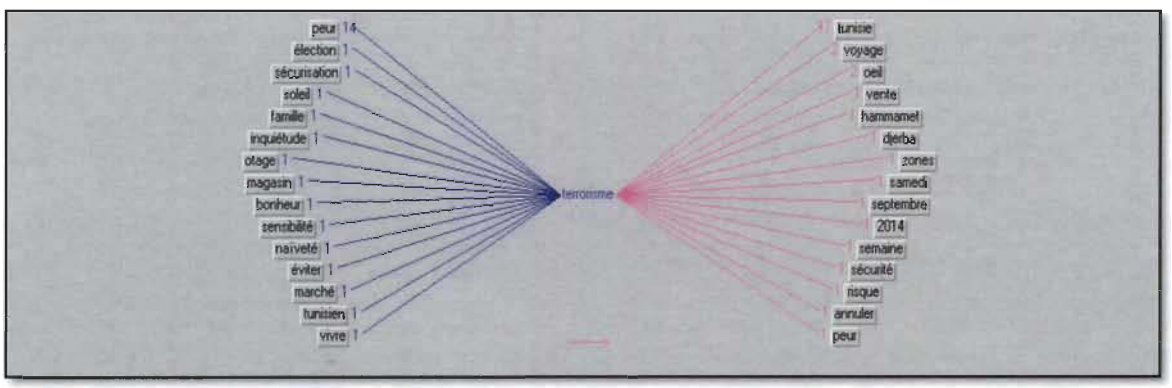

Figure 65 : Le graphe en étoile pour la référence Terrorisme

On peut constater d'après les deux dernières figures qu'il y a une forte corrélation entre (peur et appréhension), otage et Tunisie avec 14 apparitions. De plus, le terrorisme et la

Tunisie ont une corrélation forte avec 17 apparitions, de même avec le terrorisme et la peur avec 14 apparitions.

- Le graphe de ligne pointillée

Dans la figure ci-dessous, chaque référence est affichée sous la forme d'une ligne pointillée horizontale indiquant son étendue (longueur de la référence) et sa position par rapport au début du corpus. L'ordre chronologique de corpus est représenté sur l'axe horizontal du début (à gauche) jusqu'à la fin du corpus (à droite). Les références sont classées en fonction de leur ordre d'arrivée dans le texte.

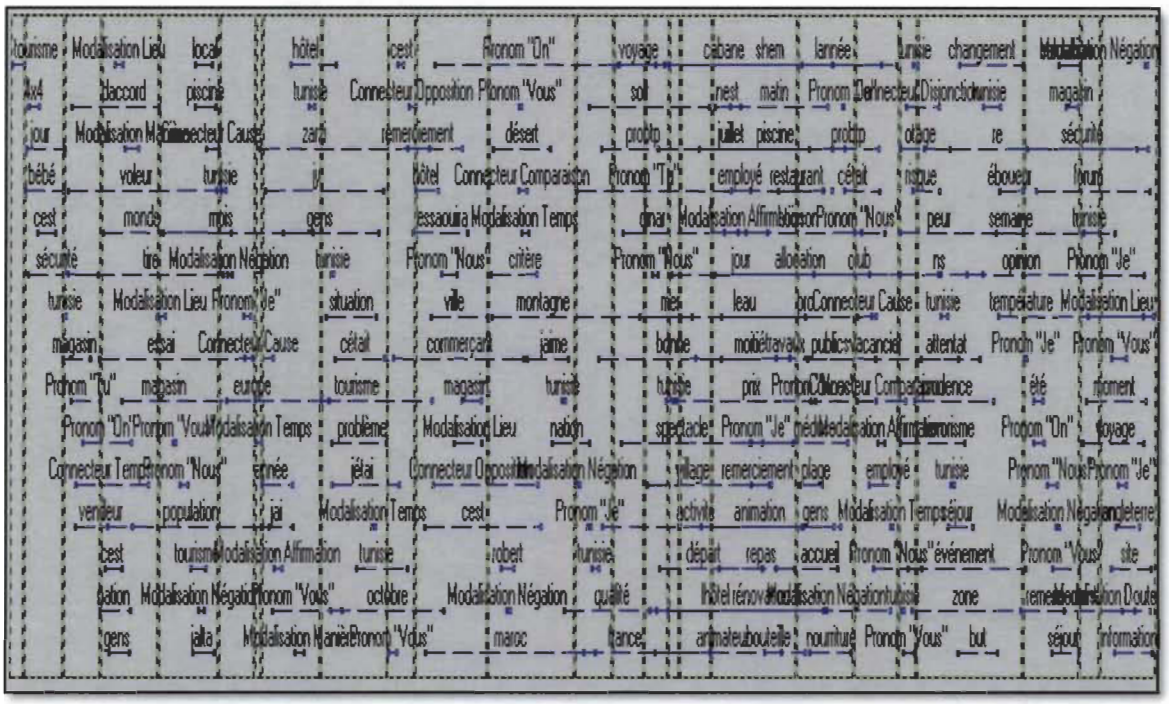

Figure 66 : Le graphe de ligne pointillée pour la référence sécurité

La référence de sécurité a été employée par les touristes durant tout le corpus, ce qui explique son importance comme un élément clé pour le tourisme en Tunisie. Toutefois, la référence sécurité a connu une forte apparition à la fin de corpus, ce qui conforme l'hypothèse dégagée de la première partie de l'analyse avec logiciel tropes où on précise que les deux graves évènements qu'a vécus la Tunisie ont une grande influence sur la sécurité et la peur pour les touristes.

d) Élaboration du tableau de bord

Tel que mentionné précédemment, la sélection de corpus est faite en tenant compte de la chronologie des événements. Ainsi trois périodes ont été identifiés, soient : le début de la révolution, l'après révolution et l'évènement marquant les attaques terroristes.

Notre analyse consiste à déterminer la constance du sentiment des voyageurs à travers ces trois évènements chronologiques pour identifier le plus dominant. Ainsi, on devrait déterminer le nombre de chronologie où le sentiment est apparu.

Substantif	**Apparition**			
	Début de la révolution	Après la révolution	Après les attaques terroristes	Nbr de chronologie où le sentiment est apparu
Sécurité	*	*	*	3
Peur	-	*	*	2
Révolution	*	*	*	3
Otage	-	-	*	1
Risque	*	*	*	3
Attentat	-	*	*	2
Total				14 (77,8 %)

Tableau 25 : Le nombre de chronologie où le sentiment est apparu

(*) : Le sentiment est apparu
(-) : Le sentiment n'est pas apparu

Pour la création du tableau, nous avons utilisé l'outil Edraw Max. C'est un outil de dessin simple, puissant et diversifié permettant aux chercheurs et aux professionnels de créer et publier de façon fiable des figures, diagrammes, tableaux, graphes en tout type pour représenter une information. Le choix des couleurs rouge, jaune et vert n'est pas arbitraire. On les a choisis pour mettre en valeur l'information.

Ci-après le résultat pratique de notre tableau de bord :

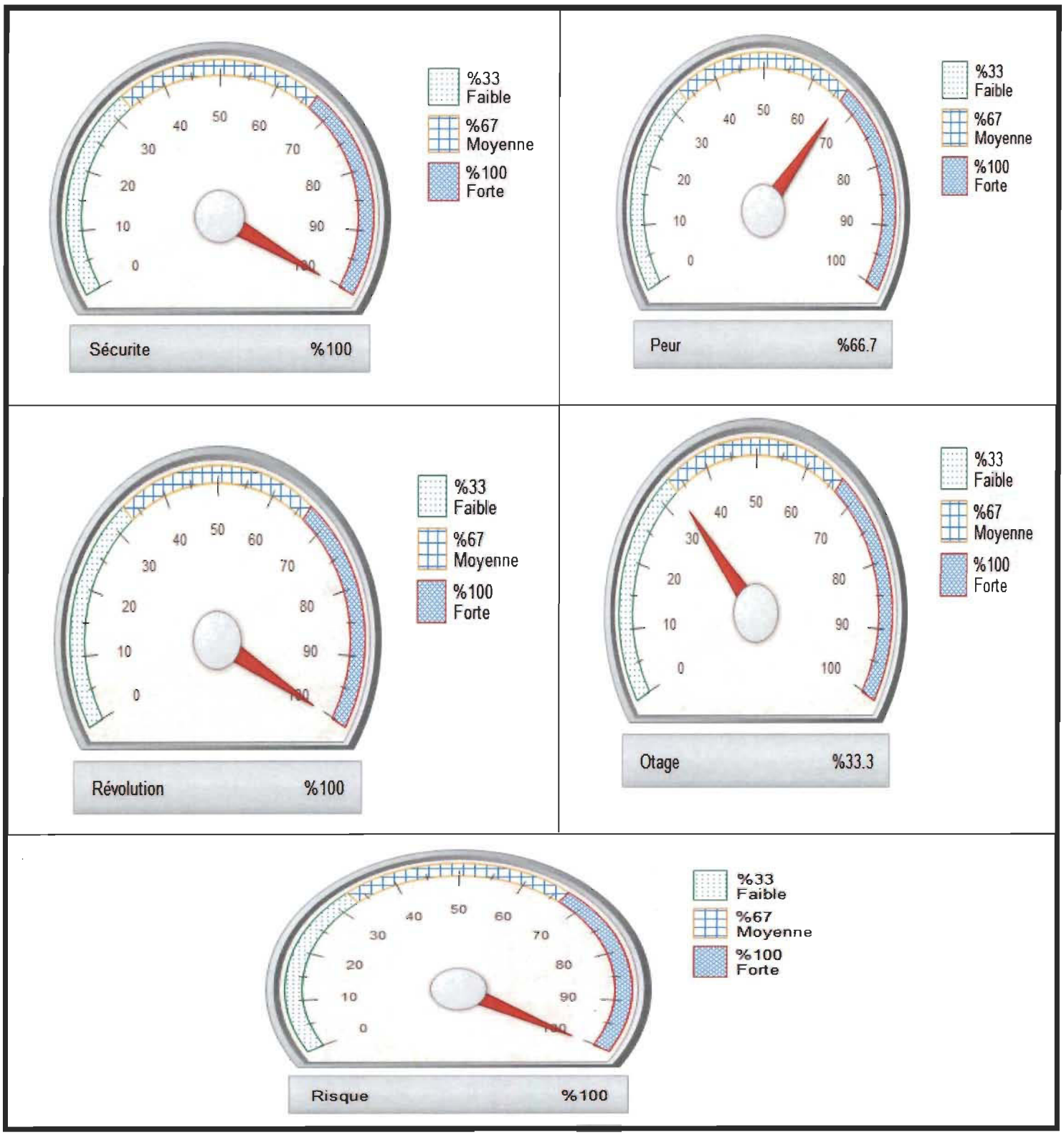

Figure 67 : Tableau du bord 1

On peut conclure que depuis la révolution, le terme le plus marquant du tourisme tunisien est la sécurité. Les résultats montrent que les touristes se préoccupent énormément de la sécurité au détriment de l'ambiance de leur voyage. Ce qui diminue conséquemment le plaisir de leur expérience. Ainsi, il est évident que si le gouvernement tunisien veut améliorer le secteur du tourisme en Tunisie, il va falloir investir dans la sécurité et mettre en place les mécanismes nécessaires afin de garantir un milieu propice au tourisme permettant de vivre une expérience agréable. Le schéma ci-après témoigne de l'importance des investissements en sécurité en Tunisie.

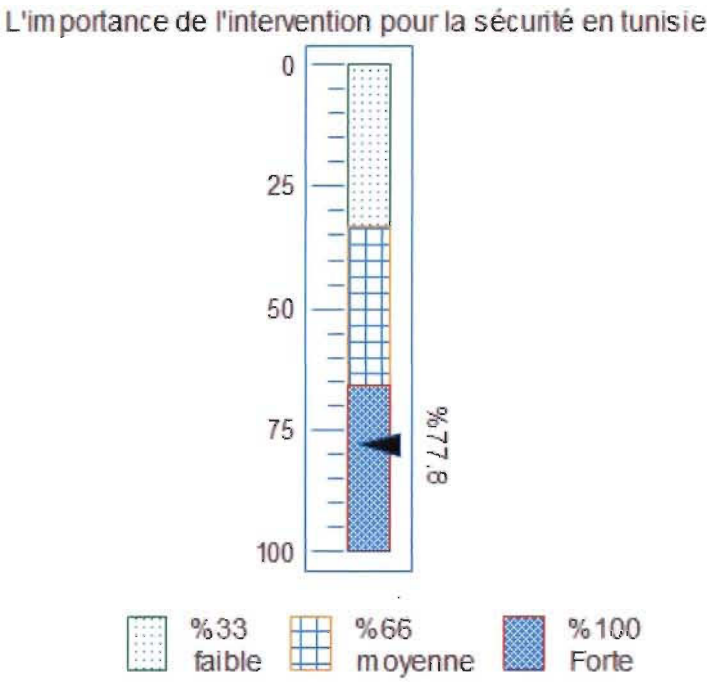

Figure 68: Tableau du bord 2

II. Conclusion

Nous avons consacré ce chapitre à la présentation des résultats obtenus pour tester notre approche. Les résultats donnent une grande satisfaction, soit pour la pertinence de l'information extraite en utilisant notre approche, soit par la présentation des résultats obtenus qui est devenue plus claire et permet de simplifier la communication des données.

CHAPITRE IX : CONCLUSION ET PERSPECTIVES

Avec l'émergence des technologies de l'information, on assiste de nos jours à des grands volumes de données qui transigent entre différents systèmes. Ces données constituent un actif stratégique et concurrentiel très important. Dans ce sens l'analyse de ce grand volume représente un défi pour les entreprises. Ce défi prend de l'ampleur avec l'analyses de textes provenant de forum de discussion à cause de la forme non structurée du langage utilisé et de la multitude des langues utilisées dans un même message.

Dans notre cas, nous étions confrontés à deux défis (forme non structurée et multilinguisme des messages). Les outils traditionnels ne permettent pas de résoudre ce problème. Ainsi, nous avons présenté une approche méthodologique (de haut en bas) intégrant plusieurs outils et méthodes dans le but d'identifier et présenter des indicateurs.

Notre approche repose sur 3 étapes :

1) **Étape préalable à l'analyse** : Cette étape consiste à :
    - Détecter et supprimer les textes non pertinents qui ne contiennent aucune information, ce qui permet de réduire le temps de prétraitement du processus habituel d'analyse des textes.
    - Unifier la langue utilisée à cause de l'aspect multilingue associé aux réseaux sociaux, cela permet d'améliorer la qualité et la précision des résultats d'analyse.
2) **Processus habituel d'analyse des textes** : Cette étape consiste à :
    - Appliquer le prétraitement pour nettoyer les données,
    - Utiliser le logiciel tropes pour :
        - Identifier le type de corpus en analysant les pronoms personnels,
        - Analyser les verbes, les adjectifs et les noms,
        - Analyser les graphes de types sphère, étoile et ligne pointillée.
3) **Élaboration du tableau de bord** : Il est parfois difficile de présenter des données de façon compréhensible à des personnes qui ne sont pas spécialisées. Nous avons opté de présenter les résultats dans un tableau de bord car une visualisation réussie permet de donner une valeur importante au résultat obtenu grâce à des éléments de couleur et de contexte.

Notre approche a été testée en utilisant comme corpus des commentaires en ligne des touristes de nationalités différentes décrivant leurs expériences vécues en Tunisie entre la période 2007 et 2017. Les commentaires sont extraits de la plateforme TripAdvisor. Le résultat le plus important qui se dégage de notre analyse est que le tourisme tunisien a souffert du manque de sécurité. Ce manque a créé un état de peur, d'incertitude et de crainte chez les touristes. De plus, ces différents sentiments se sont accentués après un évènement spécial et tragique, soit l'attentat du musée de Bardo. Effectivement, la peur et la crainte étaient le sentiment le plus dominant dans les opinions exprimées par les touristes sur le forum de voyage.

En conclusion, l'analyse textuelle des opinions exprimés par les touristes sur le forum des voyages confirme l'état problématique et alarmant du tourisme tunisien et montre aux décideurs qu'il faut investir sur la sécurité en premier lieu pour restaurer l'image pacifique de la Tunisie et réduire voire éliminer les sentiments de peur et de crainte qui règnent chez les touristes.

Tout système étant appelé à évoluer dans le temps, des améliorations peuvent être apportées à ce modeste travail afin de le rendre plus utile :

- Généraliser notre approche pour qu'elle couvre tous les réseaux sociaux et surtout Twitter, qui utilise d'autres formes de messages, soit les Hashtags;
- Généraliser l'utilisation de l'approche en incluant plusieurs autres langues comme la langue arabe, considérée comme l'une des langues les plus difficile en traitement automatique des langues à cause de ses propriétés morphologiques et syntaxiques;
- Codifier notre approche en utilisant les algorithmes proposés pour réaliser une application complète.

CHAPITRE X : REFERENCES

REFERENCES

[1] Camille Lafrance, Le tourisme en Tunisie deux ans après la révolution,3 août 2012.

[2] Kabane.ca, Les chiffres numériques du Canada en 2018, Février 2018.

[3] Michel Marcoccia, L'animation d'un espace numérique de discussion : l'exemple des forums usenet, 2001.

[4] Atefeh Farzindar, Mathieu Roche, Les défis du traitement automatique du langage pour l'analyse des réseaux sociaux, Mai 2015.

[5] Meishan et al, Résumé de blog orienté vers les commentaires par extraction de phrases,2007.

[6] Freddy Chongtat Chua, Résumé automatique des événements des médias sociaux, 2013.

[7] Mohamed Dermouche et Al, Analyse et visualisation d'opinions dans un cadre de veille sur le Web, 2015.

[8] Adeline et Al, Décodeur neuronal pour la transcription de documents manuscrits anciens, 2018.

[9] Nicolas Despres et Al, Apprentissage de Modèles de Langue Neuronaux pour la Recherche d'Information, 2016.

[10] Yves Bestgen, Construction automatique d'un lexique de n-grammes pour la fouille d'opinion, 2014.

[11] Jean Moscarola, Younès Boughzala, Analyser les corpus d'avis en ligne : Analyse lexicale exploratoire et/ou modélisation sémantique, juin 2016.

[12] Jean-Sébastien Vayre, Les tableaux de bord sur données massives, pour un nouveau management de l'innovation? janvier 2016.

[13] Fayyad et al, Knowledge discovery and data mining: Towards a unifying framework. in knowledge discovery and data mining. Pages 82–88, 1996.

[14] Grégoire de Lassence, Apport du Text Mining pour l'amélioration de la performance des modèles de Data Mining Grégoire de Lassence, 2006.

[15] Jean-Louis Monino, Soraya Sedkaoui, Big Data, Open Data et valorisation des données, 2016.

[16] Georges Gardarin, Fouille de Texte (Text Mining), 2009.

[17] Paul Robert, dictionnaire le petit robert, page 234.

[18] Younes Benzaki, Overfitting et Underfitting : Quand vos algorithmes de Machine Learning dérapent, 11 juillet 2017.

[19] Chris Manning, Hinrich Schütze, C.D. Manning & H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[20] Lefèvre, La recherche d'information - du texte intégral au thésaurus, 2000.

[21] Bensbaa Abdelaziz, Classification des textes arabes Basée sur une ontologie de domaine, 2018.

[22] Marti A. Hearst, Untangling Text Data Mining, 1999.

[23] Florence Herbulot, La Théorie interprétative ou Théorie du sens : point de vue d'une praticienne, 02 Juin 2004.

[24] Travaux Publics et Services Gouvernementaux Canada, 2104.

[25] Imane et Al, Étude comparative des approches de traduction automatique, juin 2104.

[26] Nina Khayat, Big Data pour augmenter la satisfaction client,16 Janv 2015.

[27] Greg Lessard, Introduction à la linguistique française, 24 décembre 2008.

[28] Martin Beaudoin,Phonologie : le système phonique d'une langue, Août 2002.

[29] Jukka Havu, Analyse linguistique, 2014.

[30] Ekué Wélédji Kpognon, Analyse syntaxique en JAVA, 30 Octobre 2015.

[31] Jo Katambwe, Kéren Genest et Béatrice Porco, Approches méthodologiques et objets d'induction organisationnels : la pertinence d'une stratégie de recherche multiétagée, Automne 2014.

[32] Aurelien Deixonne, La traduction automatique en 2018, révolution ou désillusion, 6 Mars 2018.

[33] Achraf Othman, Mohamed Jemni, La traduction automatique à base des statistiques au service de la langue des signes,13 mars 2017.

[34] Raphaël Dahl, L'intelligence artificielle au service de la traduction automatique de contenus, 09 Juin 2017.

[35] Tradonline, Traduction automatique ou traduction humaine ? Au cas par cas ! ,2 Août 2018.

[36] Thierry Poibeau, Traduire sans comprendre ? La place de la sémantique en traduction automatique, Mars 2016.

[37] Matthew Carrozo, L'engouement pour la Traduction Automatique Neuronale : « C'est très facile de battre Google qui propose des solutions toutes faites », 26 Octobre 2017.

[38] Marc Zaffagni, Google améliore son système de traduction instantanée grâce à l'intelligence artificielle (MAJ) ,30 septembre 2016.

[39] Eri, DeepL s'attaque à Google Traduction ,23 octobre 2017.

[40] Souhir Gahbiche-Braham, Amélioration des systèmes de traduction par analyse linguistique et thématique : application à la traduction depuis l'arabe, Octobre 2013.

[41] Karolak Magdalena, Social media in democratic transitions and consolidations: what can we learn from the case of Tunisia, 01 Juin 2018.

[42] Michel Courcelles,Un nouveau traducteur automatique: Deepl, 31 Aout 2017.

[43] Mokhtar Taffar,Support de Cours pour étudiants en Master en Intelligence Artificielle : Initiation à l'apprentissage Automatique, 2013.

[44] Frédéric Camps, Intelligence Artificielle, juin 2018.

[45] Jean-Francis Roy, Apprentissage automatique avec garanties de généralisation à l'aide de méthodes d'ensemble maximisant le désaccord, 2018.

[46] Vincent Dely, 5 conseils sur la classification des données, juin 2018.

[47] L'UC Berkeley,Q & R: L'avenir de l'intelligence artificielle.

[48] Claude Touzet, Les réseaux de neurones artificiels, introduction au connexionnisme, 27 juin 2016.

[49] Marc Guerrien, L'intérêt de l'analyse en composantes principales (ACP) pour la recherche en sciences sociales, 2003.

[50] Michel Bret, Les réseaux de Kohonen, juillet 2018.

[51] Christel Ruwet,2012, Efficacité de classification de la méthode des k-moyennes tronquées.

[52] Lance & Williams, A general theory of classificatory sorting strategies: I. Hierarchical systems. Computer Journal, 9, 373-380.

[53] Arnaud Revel, Apprentissage Semi-Supervise.

[54] Ievgen Redko, Younès Bennani, NMF multi-couches aléatoire pour l'apprentissage par transfert non-supervisé, Février 2018.

[55] In Principio, Machine learning et Intelligence Artificielle, 2017.

[56] Hubert Wassner, Approche Machine Learning non supervisée : quand l'I.A devient génératrice de business et d'emplois, 2017.

[57] Gilles R. Ducharme, Critères de qualité d'un classifieur généraliste, 21 Jun 2018.

[58] Pierre voyer, Tableaux de bord de gestion et indicateurs de performance 2e édition 1999, page 61.

[59] Le Conseil du trésor du québec, Guide sur les indicateurs, 2002.

[60] Bernard Lebelle, Construire un tableau de bord pertinent avec Excel 2e Édition, Édition d'Organisation, septembre 2013.

[61] Aurélien boutaud, Qu'est-ce qu'un indicateur, 30 novembre 2015.

[62] Salma Bougar, Ahmed Fath Allah Rahmouni, Badr Abouzaid, Pilotage stratégique par l'outil Balanced Scorecard, Mars 2018.

[63] Samuel Legault-Mercier Michèle St-Pierre, Mesure de la qualité ou appréciation chiffrée d'un phénomène pour offrir une certaine objectivation de la réalité étudiée,2011.

[64] M. Hammer, Carnet de route pour manager, Maxima, Paris, 2002.

[65] M. Gervais, Probablement Michel Gervais, Contrôle de gestion et planification de l'entreprise, 2000 (7e édition), page 598.

[66] H. Bouquin, Le Contrôle de gestion, Paris : Presses Universitaires de France, 8e édition, 2008.

[67] Claude Alazard et Sabine Sépari, Contrôle de gestion Ed. Dunod ; P. 591, 2010.

[68] Pierre voyer, Tableaux de bord de gestion et indicateurs de performance 2e édition 1999.

[69] Gabriel Dabi-Schwebel, Logiciel de tableau de bord (dashboard) 22 juillet 2015.

[70] L'équipe de Manager GO,Guide pour élaborer un tableau de bord, 17 juillet 2018.

[71] Michel Leroy, Le tableau de bord au service de l'entreprise, édition d'Organisation, 2001.

[72] Stephen Few, Information Dashboard Design: the effective visual communication of data, 2006.

[73] Direction générale de la planification de la performance et de la qualité, Guide de sélection et d'élaboration des indicateurs aux fins de l'évaluation de la performance du système public de santé et de services sociaux, Décembre 2012.

[74] Atefeh Farzindar, Mathieu Roche, Les défis du traitement automatique du langage pour l'analyse des réseaux sociaux, Mai 2015.

[75] Pierre Voyer, Tableaux de Bord de Gestion et Indicateurs de Performance, 2006.

[76] Ryan Saad, Quelles sont les plateformes d'avis client à privilégier? Juin 2017.

[77] Stéphane Goria, Évolution des systèmes de gestion des connaissances et d'intelligence économique, 2018.

[78] Chloé et al, 2018, Nouvelles modalités d'interaction pour des opérateurs de maintenance en milieu contraint : Contribution d'une approche conjointe FH et IHM dans le contexte d'un projet multipartenaire, 2018.

[79] Marli et al, Utilisation du logiciel IRAMUTEQ dans l'analyse de données en recherche qualitative, 2018.

[80] Daras et al, Automatisation du processus de prétraitement des données spatiales, 2017.

[81] Philipp Koehn, Traduction automatique statistique, 2010.

[82] Saint André, Quelle formation donner aux traducteurs-post éditeurs de demain, 2015.

[83] Support Microsoft, Aide et Support Microsoft en ligne, 2011

[84] Mathieu Feuilloy, Étude d'algorithmes d'apprentissage artificiel pour la prédiction de la syncope chez l'homme, mars 2010.

[85] Joseph Lark et al, CANÉPHORE : un corpus français pour la fouille d'opinion ciblée, 2015.