الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليـــــم العالـــي والبحـث العلـمـي
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
جامعة محمد البشير الإبراهيمي - بـرج بـوعـريـريـج -
Université Mohamed El Bachir El Ibrahimi de Bordj Bou Arreridj

Faculté des Mathématiques et d'Informatique
Département d'Informatique

Mémoire

Présenté en vue de l'obtention du diplôme
Master en Informatique
Spécialité : Réseaux et multimédia

THEME

Analyse des sentiments des tweets liés au Hirak

Présentée par :

- Djerrad Maissa.
- Zidoune sarah.

Encadrée par :

- Dr. Laifa Meriem.

Les membres de jurys :

Président :	Mr.Maza Sofiane	Université de BBA
Examinateur :	Mr.Saha Adel	Université de BBA

Année universitaire : 2020 - 2021

Remerciement

A l'issue de travail nous remercions Allah qui nous aide donne la patience et le courage durant ces longues années d'études.

Nous tenons à saisir cette occasion et adresser nos profonds remerciements et nos profondes reconnaissances à :

Notre encadreur Mme : Laifa Meriem pour ses précieux conseils et son aide durant toute la période de travail.

Nos remerciements vont également à nos parents.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils sont eus porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

Nous tenons à remercier toute personne qui a participé de près ou de loin à l'exécution de ce travail.

Dédicace

Je dédie sincèrement et affectueusement mon humble travail :

A mes chères parents ma mère et mon père pour leur patience, leur amour, leur soutient et leurs encouragements.

A ma grande mère, A mes amies et mes camarades sans oublier tous les professeurs.

.

Djerrad Maissa

Je dédie ce mémoire

A mes chères parents ma mère et mon père pour leur patience, leur amour, leur soutient et leurs encouragements, A mes frères et A mes amies et mes camarades sans oublier tous les professeurs.

Zidoune Sarah

Résumé

Depuis le 22 février 2019, des millions d'Algériens sont descendus dans les rues de toutes les grandes villes du pays pour exprimer leur rejet d'un cinquième mandat d'Abdelaziz Bouteflika. Ce mouvement social (appelé Hirak) a été diffusé dans divers médias sociaux tels que Twitter, où les internautes ont exprimé leurs opinions et sentiments qui différaient entre positifs, négatifs ou neutres. Le but de ce projet était d'analyser les sentiments et les tweets liés au mouvement algérien à travers une application des Algorithmes de classification tels que naïve bayésienne, machine a support vectorielle, arbre de décision et algorithme de régression logistique avec des différentes méthodes d'extraction des attributs qui sont « sac de mots » et « TF-IDF » sur une base de données qui contienne 10000 tweets divisé en 5846 avis positifs, 1741 avis négatifs et 2184 avis neutres. Le résultat expérimental a montré que le bon classificateur est le svm avec une précision raisonnable égal 67%.

Mots clés : *Hirak ; Analyse des sentiments ; Le traitement automatique du langage naturel ; Les algorithmes de classification ; Annotation manuelle ; L'apprentissage automatique ; Dialecte Algérien*.

Abstract

Since February 22, 2019, millions of Algerians have taken to the streets of all major cities in the country to express their rejection of a fifth term of Abdelaziz Bouteflika. This social movement (called Hirak) was broadcast on various social media such as Twitter, where netizens expressed their opinions and feelings which differed between positive, negative or neutral. The goal of this project was to analyze the sentiments and the tweets linked to the Algerian movement through an application of classification algorithms such as Bayesian naive, vector-supported machine, decision tree and logistic regression algorithm with different methods of extraction of the attributes which are « BOW » and « TF -IDF » on a dataset that contains 10,000 tweets divided into 5,846 positive reviews, 1,741 negative reviews and 2,184 neutral reviews. The experimental result showed that the correct classifier is the svm with reasonable precision equal to 67%.

Key words : *Hirak ; Sentiment analysis ; Natural language processing ; Classification algorithms ; Manual annotation ; Machine Learning ; Algerian dialect*.

الملخص

منذ 22 فبراير 2019، نزل ملايين الجزائريين إلى شوارع جميع المدن الكبرى في البلاد للتعبير عن رفضهم لولاية خامسة لعبد العزيز بوتفليقة. حيث تم بث هذه الحركة الاجتماعية (المسماة الحراك) عبر وسائل التواصل الاجتماعي المختلفة مثل تويتر، حيث عبر مستخدمي الإنترنت عن آرائهم ومشاعرهم التي اختلفت بين الإيجابية والسلبية والحيادية. كان الهدف من هذا المشروع هو تحليل المشاعر والتغريدات المرتبطة بالحركة الجزائرية من خلال تطبيق خوارزميات التصنيف مثل تصنيف بايزي ساذج، الة المتجهات الداعمة، شجرة القرار وخوارزمية الانحدار اللوجستي. بواسطة مختلف طرق لاستخراج السمات والتي هي عبارة عن "حقيبة الكلمات" و" تردد المصطلح- تردد المستند العكس". "في قاعدة بيانات تحتوي على 10000 تغريدة مقسمة إلى 5846 تعليقًا إيجابيًا ،1741 تعليقًا سلبيًا و2184 تعليقا محايدا. حيث أظهرت النتائج التجريبية أن المصنف الصحيح هو الة المتجهات الداعمة وبدقة معقولة تساوي 67%.

الكلمات الرئيسية: *الحراك؛ تحليل المشاعر؛ معالجة اللغة الطبيعية؛ خوارزميات التصنيف؛ وسم يدوي؛ تعلم الالة؛ اللهجة الجزائرية.*

Table des Matières

Remerciement ……i

Dédicace ……...ii

Résumé ……iii

Liste des figures ……viii

Liste des tableaux ……ix

Liste des équations ……x

Introduction Générale

1. Contexte 2

2. Problématique et objective.....2

3. Plan de mémoire 2

Chapitre II : Etat de l'art

II.1. Introduction 5

II.2. Traitement automatique du langage naturel.....5

II.2.1. Définition 5

II.2.2. Le processus du traitement automatique du langage naturel 5

II.2.3. Domaine d'application le NLP 6

II.2.3.1. Traduction automatique 6

II.2.3.2. Catégorisation de texte.....6

II.2.3.3. Filtrage des spams.....7

II.2.3.4. Résumé.....7

II.2.3.5. Système de dialogue.....7

II.2.3.6. Médecine.....8

II.2.3.7. Extraction informations.....8

II.3. Analyse des sentiments 8

II.3.1. Définitions d'analyse des sentiments.....8

II.3.1.1. Définition d'opinion :.....9

II.3.1.2. Définition du sentiment :.....9

II.3.2. Le processus d'analyse des sentiments : 9
II.3.2.1. La collection des données 10
II.3.2.2. Le pré-traitement 10
II.3.2.3. Le filtrage 11
II.3.2.4. La classification 11
II.3.2.5. La visualisation des résultats d'analyse des sentiments.....11
II.3.3. Domaine d'application d'analyse des sentiments 11
II.3.3.1. Application en Sociologie, Psychologie et Analyse des Sentiments Politiques.....12
II.3.3.2. Application en marketing.....12
II.3.3.3. Applications en finance.....13
II.3.3.4. Application en soins de santé.....13
II.3.4. Les travaux connexes à notre projet.....13
II.4. Conclusion.....16

Chapitre III : Conception et Modélisation

III.1. Introduction.....18
III.2. Description détaillée de l'objective 18
III.3. La conception générale du système 18
III.4. La conception détaillée du système 19
III.4.1. La collection des données 19
III.4.2. La division de la base de données.....20
III.4.3. Annotation 21
III.4.4. Application du processus d'analyse des sentiments 22
III.4.4.1. Le pré-traitement.....22
III.4.4.1.1 La conversation des données textuelles en minuscules 23
III.4.4.1.2 Le nettoyage des données 23
III.4.4.1.3. Suppressions des mots d'arrêt 26
III.4.4.1.4. La tokenisation.....27
III.4.4.1.5. Le mot du nuage 27

III.4.4.2. Feature extraction 28
III.4.4.2.1. La méthode Sac des Mots 28
III.4.4.2.2 La méthode TF-IDF 28
III.4.4.3. La classification 28
III.4.4.3.1. Naïve Bayes 29
III.4.4.3.2. Machines à Support Vectorielle 29
III.4.4.3.3. Arbre de décision 30
III.4.4.3.4. Régression logistique 30
III.4.4.4. Évaluation 31
III.4.4.4.1. Précision 31
III.4.4.4.2. Rappel 32
III.4.4.4.3. Exactitudes 32
III.4.4.4.4. Le score F1 32
III.5. Conclusion 32
Chapitre IV : Implémentation et Résultats
IV.1. Introduction 34
IV.2. L'environnement de travail et les outils utilisés 34
IV.2.1. L'environnement Matériel 34
IV.2.2. L'environnement Logiciel 34
IV.3. Présentation des données 35
IV.4. Prétraitement 36
IV.5. Classification 37
IV.6. Comparaison 38
IV.7. Conclusion 39
Conclusion Générale 41
Les références 43

Liste des figures

Figure 1. Processus d'analyse des sentiments. 9
Figure 2. Conception générale du notre système. 18
Figure 3. La collection des données.....19
Figure 4. La base qui contient des tweets en Arabe. 20
Figure 5. La base qui contient des tweets en Français.....21
Figure 6. Annotations des tweets. 22
Figure 7. La conversion de majuscule en minuscule. 23
Figure 8. La suppression des numéros, des mentions, des URL, des emojis, des caractères spéciaux et des ponctuations.....24
Figure 9. Suppression des hashtags. 25
Figure 10. Suppression des mots d'arrêt.....26
Figure 11. La tokenisation. 27
Figure 12. La répartition des avis positifs, négatifs et neutres. 36
Figure 13. Le nuage de mot. 37

Liste des tableaux

Tableau 1. La suppression des tweets répétés. 25
Tableau 2. Matrice de confusion. 31
Tableau 3. Les différents attributs dans la base de données.....35
Tableau 4. Description des attributs du base de données utilisé. 35
Tableau 5. Le résultat d'exactitude des classificateurs avec le TF-IDF..37
Tableau 6. Le résultat d'exactitude des classificateurs avec le BOW.....38
Tableau 7. Comparaison des performances selon les classificateurs utilisés. 39

Liste des équations

(1).....31
(2).....32
(3).....32
(4).....32

Introduction Générale

1. Contexte

Depuis l'apparition de la notion de Web 2.0 et l'émergence des sites communautaires, l'internet devient le moyen le plus sophistique qui donne la possibilité de communiquer, de s'exprimer leurs opinions à propos de différents sujets, à travers les réseaux sociaux, les blogs, les forums, les sites de e-commerce et les sites d'actuelles …etc. à une échelle mondiale. La grande partie de ces informations dont leur taille est en pleine expansion, qui en une sorte ou autres décrivent des sentiments qui sont devenus un objet d'étude dans plusieurs domaines de recherche tel que « l'analyse des sentiments et la détections des opinions ».

L'analyse de sentiment est l'un des nouveaux défis apparu en traitement automatique des langues qui consiste à analyser les opinions des internautes sur un sujet donné pour découvrir sa polarité et de les classer comme positive, négative ou neutre [1].

2. Problématique et Objective

Le 22 février 2019, l'Algérie a vu l'émergence d'un mouvement pacifique pour la démocratie, qui a fait descendre les gens ordinaires dans la rue à une échelle sans précédent. Connu sous le nom de Hirak, des marches hebdomadaires de millions de personnes ont conduit à la démission du président Bouteflika, et à l'arrestation et au procès de hauts responsables politiques et d'élites des affaires. Largement inédit en Occident, cette mobilisation phénoménale s'est poursuivie jusqu'en 2020. Ce mouvement social a été diffusé dans divers médias sociaux tels que Twitter et Facebook, où les internautes ont exprimé leurs opinions et leurs sentiments où les opinions différaient entre positifs, négatifs ou neutres. Le but de notre projet est l'analyse des données en utilisant les différentes techniques de l'apprentissage automatique tels que le traitement automatique de langage naturel et l'analyse des sentiments.

3. Plan de mémoire

Le contenu de ce mémoire est structuré comme suit :

- **Le premier chapitre :** Introduction générale

  Nous avons commencé notre mémoire avec une introduction générale.

- **Le deuxième chapitre :** l’état de l’art

Nous avons fait une étude sur les généralités de l’analyse des sentiments et de traitement automatique du langage naturel et nous avons mis en lumière les travaux qui sont effectués dans le cadre de l’analyse des sentiments.

- **Le troisième chapitre** : Conception et modélisation

Nous avons parlé en détail de l'objectif du notre projet. En plus, nous allons présenter Une explication détaillée du processus d’analyse des sentiments. On a aussi mentionné les méthodes de classification utilisée avec précision.

- **Le quatrième chapitre :** Implémentation et résultats

Dans ce dernier chapitre, nous avons fait une présentation de l’environnement de travail et les outils utilisés puis nous avons fait une analyse exploratoire des données.

- **Le cinquième chapitres** : Conclusion générale

Nous avons clôturé ce mémoire par une conclusion générale.

Chapitre II

Etat de l’art

II.1. Introduction

Le traitement de langage naturel est une application avancée de l'Intelligence artificielle et de l'apprentissage automatique utilisée pour comprendre le langage humain et pour extraire des informations sémantiques de toutes les sources des données qu'elles soient textuelles, audio ou vidéo. Dans ce chapitre, nous allons aborder deux notions nécessaires pour notre projet qui sont le traitement automatique du langage naturel et l'analyse des sentiments. Nous définirons également le contexte de notre étude.

II.2. Traitement automatique du langage naturel

II.2.1. Définition

Le traitement automatique du langage naturel (Natural Language processing ou NLP en Anglais) est défini comme un domaine spécialisé de l'informatique, de l'ingénierie et de l'intelligence artificielle avec des racines dans la linguistique informatique. Le NLP est utilisé pour la construction des applications et des systèmes permettant l'interaction entre les machines et les langages naturels crée par les humains. Ce dernier utilise des techniques permettant aux ordinateurs de traiter et de comprendre le langage naturel humain et l'utilisation davantage pour fournir des résultats utiles [2].

II.2.2. Le processus du traitement automatique du langage naturel

Le processus du traitement automatique du langage naturel est une discipline qui peut être divisé en deux parties principales. La première partie est la compréhension du langage naturel et la deuxième partie est la génération du langage naturel. Sachant que le type des données en entrées et de la sortie peuvent être du texte ou du la parole [3].

La compréhension du langage naturel :

- Définir l'entrée donnée en langage naturel en une représentation utile.
- Analyser les différents aspects de la langue.

La génération du langage naturel :

- La planification de texte : cela comprendre la récupération du contenu pertinent de la base de connaissances.
- Planification de la phrase : cela comprendre le mot requis et la formation des phrases significatives.

II.2.3. Domaine d'application le NLP

Le traitement du langage naturel peut être appliqué dans divers domaines comme [4] :

II.2.3.1. Traduction automatique

Comme la majeure partie du monde est en ligne, la tâche de rendre les données accessibles et disponibles pour tous est un défi. Ce dernier est la barrière de la langue. Il existe une multitude de langues avec une structure de phrase et une grammaire différente. La traduction automatique traduit généralement des phrases une langue à une autre à l'aide d'un moteur statistique comme Google Translate. Le défi avec les technologies de traduction automatique n'est pas de traduire directement les mots, mais de conserver le sens des phrases intact ainsi que la grammaire et les temps.

II.2.3.2. Catégorisation de texte

Les systèmes de catégorisation entrent un grand flux des données telles que des documents officiels, des rapports de pertes militaires, des données de marché, des fils de presse, etc. et les attribuent à des catégories ou indices prédéfinis. Par exemple, le système Construit du Carnegie Group, saisit les articles de Reuters et fait gagner beaucoup de temps en effectuant le travail qui doit être effectué par le personnel ou homme indexeurs. Certaines entreprises ont utilisé des systèmes de catégorisation pour catégoriser les tickets incidents ou les demandes de réclamation et les acheminer vers les bureaux appropriés.

II.2.3.3. Filtrage des spams

Le filtrage des spams fonctionne en utilisant la catégorisation de texte. Ces derniers temps, diverses techniques apprentissage automatique ont été appliquées à la catégorisation de texte ou au filtrage anti-spam comme Rule Learning, Naïve Bayes, Machines à vecteurs de support, Arbres de décision etc. L'utilisation de ces approches est la meilleure car le classificateur est appris à partir des données apprentissage plutôt que de le faire à la main.

II.2.3.4. Résumé

La surcharge des informations est la réalité première numérique, et déjà notre portée et notre accès aux connaissances et à l'information qui dépasse notre capacité à les comprendre. Cette tendance ne ralentit pas, donc une capacité à résumer les données tout en gardant le sens intact est hautement obligatoire. Ceci est important, non seulement pour nous permettre de reconnaître et de comprendre les informations importantes pour un grand nombre de données, mais également pour comprendre les significations émotionnelles plus profondes ; Par exemple, une entreprise détermine le sentiment général sur les médias sociaux et utilise sur sa dernière offre des produits. Cette application est utile comme un atout marketing précieux.

II.2.3.5. Système de dialogue

Peut-être application la plus souhaitable du futur, dans les systèmes envisagés par les grands fournisseurs applications pour utilisateurs finaux, les systèmes de dialogue, qui se concentrent sur des applications étroitement définies (comme les réfrigérateurs ou les systèmes de cinéma maison) utilisent actuellement les niveaux phonétiques et lexicaux du langage. On pense que ces systèmes de dialogue lorsqu' ils utilisent tous les niveaux de traitement du langage offrent un potentiel pour des systèmes de dialogue entièrement automatisés. Que ce soit par SMS ou par voix. Cela pourrait conduire à produire des systèmes permettant aux robots interagir avec les humains dans les langues naturelles. Des exemples tels que l'assistant de Google, Windows Cortana, Siri d'Apple et Alexa, Amazon sont les logiciels et les appareils qui suivent les systèmes Dialogue.

II.2.3.6. Médecine

Le traitement automatique du langage naturel est également appliqué dans le domaine de la médecine. Tel que Le Linguistic String Project-Medical Language Processor, ce dernier est l'un des projets à grande échelle de Le NLP dans le domaine de la médecine. Le LSP-MLP permet aux médecins extraire et de résumer des informations sur tout signe ou symptôme, la posologie du médicament et les données de réponse dans le but identifier les effets secondaires possibles de tout médicament tout en mettant en évidence ou en signalant les éléments des données.

II.2.3.7. Extraction informations

L'extraction d'informations concerne identification des phrases d'intérêt des données textuelles. Pour de nombreuses applications, l'extraction d'entités telles que les noms, les lieux, les événements, les dates, les heures et les prix est un moyen puissant de résumer les informations pertinentes aux besoins d'un utilisateur. Dans le cas d'un moteur de recherche spécifique à un domaine, l'identification automatique d'informations importantes peut augmenter la précision et l'efficacité d'une recherche dirigée. Des modèles de Markov cachés (HMM) sont utilisés pour extraire les domaines pertinents des articles de recherche. Ces segments de texte extraits sont utilisés pour permettre la recherche dans des domaines spécifiques et pour fournir une présentation efficace des résultats de la recherche et pour faire correspondre les références aux articles. Par exemple, remarquer les publicités contextuelles sur tous les sites Web affichant les articles récents que vous avez peut-être consultés sur une boutique en ligne avec des remises.

II.3. Analyse des sentiments

II.3.1. Définitions d'analyse des sentiments

Le terme d'analyse des sentiments lui-même raconte qu'il s'agit d'une analyse de divers sentiments exprimés par des humains sur internet, ou des opinions données par les clients à diverses organisations commerciales. De manière plus générale, l'analyse des sentiments ou l'exploration d'opinions utilise des techniques d'exploration des données et de traitement du langage naturel (NLP) pour découvrir, récupérer et distiller

des informations et des opinions à partir de vastes informations textuelles du World Wide Web. L'analyse des sentiments nous permet de suivre les attitudes et les sentiments sur le Web, comme les articles de blog, les commentaires, les critiques et les tweets sur toutes sortes de différents sujets pour déterminer s'ils sont vus positive, négative ou neutre sur le web [5].

II.3.1.1. Définition d'opinion :

L'opinion est un avis, un jugement personnel que l'on s'est forgé sur une question ou un sujet en discussions qui ne relève pas de la connaissance rationnelle. L'opinion est aussi une manière de penser, un ensemble d'idées ou une doctrine [6].

II.3.1.2. Définition du sentiment :

Le sentiment est comme le jugement que porte un individu sur un objet ou un sujet, ce jugement étant caractérisé par une polarité et une intensité. Tel qu'une polarité est soit positive, soit négative ou bien un mélange de ces deux valeurs, tandis que l'intensité montre le degré de positivité ou de négativité, et varie de faible à forte [7].

II.3.2. Le processus d'analyse des sentiments :

Le processus d'analyse des sentiments des données peut être effectué en passant par cinq étapes principales illustrées dans la **Figure 1** [8] :

Figure 1. Processus d'analyse des sentiments.

Ce processus commence par l'identification des mots-clés pour analyser ce que les gens pensent d'eux. Après cela vient l'étape de collecte et il existe différentes manières de colleter des tweets ciblés. Ces tweets agrégés seront en suit stockés dans un ensemble de données. Une fois les tweets sont colletés, l'étape suivante est le pré-traitement des tweets qui supprimera tout contenu non pertinent et obtiendra uniquement un texte brut. Puis l'étape de filtrage qui comprend la suppression de tous les mots qui n'affectent pas

le sens du texte. L'étape suivante consiste à classer le contenu en positive, négative ou neutre. La dernière étape consiste à avoir une idée générale de tous les tweets collectés.

II.3.2.1. La collection des données

La première étape du processus d'analyse des sentiments consiste à collecter des tweets en spécifiant un mot-clé pour récupérer tous les tweets liés à ce mot-clé. Les tweets peuvent être collectés à partir de différentes sources. L'un des moyens est le robot d'exploration de tweets qui collecte une collection de tweets liés en interrogeant le service Web Twitter. D'une autre façon en peut d'utiliser l'interface de programme d'application Twitter (API). Cette API fournie par Twitter et qui donne aux développeurs la possibilité d'utiliser les fonctions de Twitter telles que la récupération de tweets avec le mot-clé et la langue sélectionné. L'outil NODEXL de Microsoft est un autre outil utilisé pour collecter des tweets. C'est un outil qui prend en charge plusieurs fournisseurs des données de réseaux sociaux qui importent des données graphiques dans la feuille de calcul Excel. Les tweets collectés seront stockés dans une base de données afin de le classer.

II.3.2.2. Le pré-traitement

Le prétraitement est une technique utilisée pour nettoyer le texte à partir de contenus non sentimentaux, tels que des noms d'utilisateurs, des images, des hashtags, des URL et tous les mots non arabes ou parfois en bâillonnant ces contenus avec un nom unifié. Ce processus appelé étiquetage. L'étiquetage est un processus de marquage du contenu non sentimental dans un tweet qui n'a aucun impact sur le sentiment du tweet. Ceux-ci seront différents en type et en nombre. Par exemple, un lien URL peut être remplacé par une balise URL, le nom d'utilisateur qui est un mot qui apparaît après le symbole "@" dans Twitter sera balisé avec le nom d'utilisateur et le mot qui apparaît après un dièse "#" et ne se rapporte pas au sujet sera celui étiqueté avec hashtag. De plus, étant donné que les utilisateurs de Twitter utilisent des symboles tels que " (:" et "😃" pour exprimer leur opinions ces émoticônes expriment des informations précieuses au sentiment.

II.3.2.3. Le filtrage

Après l'étape de prétraitement, le résultat ne sera que du texte. L'étape de filtrage comprend d'autres étapes nécessaires pour supprimer tous les mots qui n'affectent pas ou n'ont aucun rapport avec le sens. De plus, à cette étape, les fautes d'orthographe sont corrigées et les lettres répétées du texte sont supprimées.

II.3.2.4. La classification

Cette étape représente l'étape finale où chaque tweet sera classé comme positif, négatif et neutre par le classificateur. Ces tweets seront annotés manuellement pour être comparés aux résultats du classificateur afin d'examiner l'exactitude du classificateur. Les classifieurs en général sont classés sous deux approches supervisées et non supervisées.

- La première approche de la classification est l'approche supervisée ou à base de corpus, les classificateurs d'apprentissage automatique sont utilisés tels que Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (D-Tree), K-Nearest Neighbor (KNN).
- La deuxième approche de la classification est l'approche non-supervisée. Cette approche basée sur le lexique et qui utilise un dictionnaire de mot. Le classifieur classera le dataset directement à l'aide d'un dictionnaire de mots. Chaque mot a une polarité (+1, -1 ou 0 pour positif, négatif ou neutre, respectivement).

II.3.2.5. La visualisation des résultats d'analyse des sentiments

Cette étape est la dernière phase dans le processus d'analyse des sentiments qui contient les résultats de cette classification des données et que se représente sous forme des statistiques et avec des diagrammes explicatifs.

II.3.3. Domaine d'application d'analyse des sentiments

De nos jours, l'analyse des sentiments a gagné encore plus de valeur avec l'avènement des réseaux sociaux. Leur grande diffusion et leur rôle dans la société moderne représentent l'une des nouveautés les plus intéressantes de ces dernières années, captant l'intérêt des chercheurs, des journalistes, des entreprises et des gouvernements.

L'interconnexion dense qui surgit souvent parmi les utilisateurs actifs génère un espace de discussion capable de motiver et d'impliquer les individus d'un espace plus large, reliant les personnes avec des objectifs communs et facilitant diverses formes d'action collective.

Les réseaux sociaux créent dont une révolution numérique, permettant l'expression et la diffusion des émotions et des opinions à travers le réseau, ouvrant une fenêtre sur les mondes des autres et fouillant dans leur vie.

On peut dire que L'analyse des sentiments est régulièrement convoquée dans tout processus décisionnel, tel que :

II.3.3.1. Application en Sociologie, Psychologie et Analyse des Sentiments Politiques

Elle permet aux individus d'avoir une opinion sur quelque chose (les critiques) à l'échelle mondial comme les critiques de film, les opinions politiques, les opinions sur un problème mondial, l'identification de l'adéquation des vidéos aux enfants sur la base des commentaires, les indentifications des baises dans les actualités, etc. Le sentiment fait référence à l'émotion derrière une mention sur les réseaux sociaux. C'est un moyen de mesure le ton de la conversation : la personne est-elle heureuse, agacée ou en colère ? En politique, nous peuvent analyser les tendances, identifier les biais idéologiques, cibler les publicité/messages en conséquence, évaluer les opinions du public/des électeurs. En sociologie, la propagation d'idées à travers des groupes est un concept important (cf :Rogers, «Diffusion of innovations,» 1962), les opinions que les gens ont et leurs réactions aux idées sont pertinentes pour l'adaptation de nouvelles idées. En psychologie, l'analyse de sentiment fournit une plateforme pour augmenter la psychologie enquête / expérience avec des données extraites du texte NL (langue naturel). Exemple : analyse des sentiments de rêve [5].

II.3.3.2. Application en marketing

Actuellement, les médias sociaux sont devenus une plate-forme unique d'interaction avec les clients. L'utilisation de l'analyse de sentiment peut facilement amener le marketing a un tout autre niveau, là où les entreprises ont compris que les émotions des

médias sociaux façonnent l'image de leur marque. D'une autre cote les outils d'analyse des sentiments donnent aux spécialistes du marketing un moyen de mesurer leurs efficacités et aident les consommateurs qui tentent de rechercher un produit ou un service [9].

II.3.3.3. Applications en finance

L'analyse des sentiments peut également être utilisé dans le monde financier. Avec l'analyse des sentiments, les investisseurs peuvent facilement suivre leur entreprise préférée et surveillent leur donnée de sentiment en temps réel. L'analyse des sentiments aide les investisseurs à acquérir plus facilement des informations commerciales et agréger ces informations pour prendre de meilleures décisions financières [9].

II.3.3.4. Application en soins de santé

Les blogs médicaux sont partout sur internet ces jours-ci. Ces blogs contiennent que des problèmes de médicaux, et de soins de santé tel que les maladies, les traitements médicaux et les médicaments. En raison des expériences liées à la santé et des antécédents médicaux que ces pages web fournissent aux praticiens et aux patients, des outils d'analyse de sentiment ont de l'être développer pour l'utilisation dans les domaines médicaux [9].

II.3.4. Les travaux connexes à notre projet

Les réseaux sociaux jouent un rôle essentiel dans notre vie quotidienne active non pas en termes de vie sociale mais aussi en termes de commerce électronique, d'apprentissage en ligne et de politique. Dans cette section nous présentons certaines études qui sont étroitement liées à notre projet.

Ajay Band et Aziz Fellah.[10] ont effectué une recherche qui analyse les récents changements sociaux dans le mouvement #MeToo, Où ils ont développé le Socio-Analyzer. Dans cette étude, Ils ont utilisé leurs approches en quatre phases pour mettre en œuvre le Socio-Analyzer. La classification de ce Socio-Analyzer a identifié et classé les données en trois catégories soit positive, négative et neutre. Les résultats de cette étude montrent que l'opinion de la plupart des gens est neutre. D'après ces résultats, [10] ont validé que les 765 tweets de données #MeToo est généralisent les résultats aux

données météo. Les valeurs de précisions de Socio-Analyzer et TexteBlob sont 70,74 % et 72,92% respectivement, lorsqu'elles sont considérées comme positive pour les tweets neutres.

Matalon, Magdaci et al. [11] ont utilisé l'analyse de sentiment pour prédire l'inversion d'opinion dans les tweets de la communication politique entre Israël et la Palestine. Dans cette approche, les chercheurs ont identifié 7147 paires source-citation, en appliquant un modèle d'apprentissage qui est l'arbre aléatoire. Ce modèle basé sur les caractéristiques de traitement du langage naturel du texte source et les attributs utilisateurs, pour prédire si cette source connaitra l'inversion d'opinion de retweeter avec un ROC-AUC de 0,83. Le résultat de cette étude fait apparaitre qu'environ 80% des facteurs qui expliquent l'inversion d'opinion sont liés au sentiment des messages original envers le conflit. D'une autre cote, il y a environ 14% des pairs source-citation associes au sentiment similaire à la source. Cette étude prouve que la prédiction du l'inversion d'opinion joue un rôle important dans la communication politique sur les réseaux sociaux pour optimiser la propagation de contenu.

Opinion Corpus for Arabic (OCA) a été proposé par [12] pour les critiques arabes extraites de page web liées aux films. Ce corpus avait été traduit en langue Anglaise afin de générer l'EVOCA (English Version of Opinion corpus for Arabic), le corpus d'opinion pour l'anglais. Le système proposé a été testé sur de nombreux algorithmes d'apprentissage automatique comme les machines à vecteurs de support (SVM) et Naïve Bayes (NB), en utilisant une validation croisée 10 fois. Dans cette expérience, les résultats ont indiqué que EVOCA était pire que OCA mais EVOCA est toujours comparable aux méthodes anglaises [13].

Un Framework d'analyse des sentiments arabe proposé par [9] a pu analyser les commentaires ou les tweets afin de les classer en trois catégories de sentiment soit positive, soit négative ou bien neutre. La nouveauté du Framework proposé peut gérer les dialectes arabes, l'arabizi et les émoticônes. L'ensemble de données collectées contenait 350000 tweets, pour chaque tweet, les chercheurs ont appliqué deux choses le tag et le suffrage., Pour décider du tag final pour chaque tweet. Pour évaluer les performances du Framework, les chercheurs ont utilisé trois classificateurs comme Naïve Bayes (NB), K-Plus proche voisin (KNN) et les machines à vecteurs de support

(SVM). Le résultat expérimental a montré que le Framework a obtenu de bons résultats [13].

Rezvaner Rezapour [14] a fait une analyse sur les caractéristiques et les indices linguistiques, tels que l'individualisme contre le pluralisme, le sentiment et l'émotion. Pour examiner la relation entre le médium et le discours au fil du temps. Ce travail a mené dans un contexte applicatif spécifique, comme le mouvement « #Black Lives Matter » noté (BLM). A la fin de cette analyse, [14] a fait une comparaison entre les discussions liées à cet événement dans les médias sociaux et les articles de presse. Les résultats montrent que les utilisateurs de Twitter ont tendance à utiliser davantage « nous » et « notre » lorsqu'un incident majeur survient. Cela prouve que les gens dans une société se connectent et participent au moins virtuellement au mouvement. De plus, ce résultat a constaté que les émotions et les sentiments dans le langage étaient significativement fortement influencé par les événements majeurs de l'actualité et sur Twitter. Néanmoins, il est nécessaire d'approfondir les caractéristiques linguistiques et pragmatiques pour étudier les deux médiums.

Md Hassan Zamir [15] a fait une étude qui examine les pratiques de communication menées sur Twitter dans les trajectoires du mouvement Shahbag et des citoyens du Bangladesh. Cette étude a identifié deux motifs communs aux manifestants de Shahbag : les diffuseurs, qui retweetaient fréquemment et avaient un grand degré de sortie, et les récepteurs, dont les tweets étaient retweetés à un taux élevé et avaient un grand degré d'entrée. Où L'objectif principal de cette recherche est de donner un aperçu de la façon dont les manifestants en réseau partagent des informations et communiquent sur les réseaux sociaux lors des mouvements sociaux en ligne.

Gonzalo et al. [16] ont fait une analyse des sentiments sur les catastrophes naturelles ou des mouvements sociaux en langage Espagnol, tels que le tremblement de terre chilien de 2010 et le référendum sur l'indépendance de la Catalogne en 2017. Pour effectuer cette analyse [16] ont pris comme considération les classificateurs de réseaux bayésiens et d’utiliser l’approche de facteur de bays, pour produire des réseaux plus réalistes. Le résultat de cette analyse montre l'efficacité de l'utilisation de la mesure du facteur de Bayes ainsi que ses résultats prédictifs compétitifs par rapport aux machines à vecteurs de support et aux forêts aléatoires, compte tenu d'un nombre suffisant d'exemples de formation. De plus, les réseaux résultants permettent d'identifier les

relations entre les mots, offrant des informations qualitatives pour comprendre historiquement et socialement les principales caractéristiques de la dynamique de l'événement.

Tobailil, Fernandez1 et al. [17] ont fait une représentation sur le premier lexique d'analyse des sentiments pour le dialecte Libanaise Arabizi qui s'appelle SenZi. Pour créer ce lexique [17] a fait plusieurs étapes commençant par la construction, la traduction, l'annotation et la translation de diverses ressources pour atteindre un ensemble initial de mot de sentiments 2K. Pour cette étude, Ils ont étendu à 24,600 mots de sentiment. Le résultat de cette étude a été composé un nouveau lexique Arabizi composé de 11,3 K mos positifs, 13,3 K mot négatifs avec une évaluation de score F1 estimé à 0,72.

II.4. Conclusion

Dans ce chapitre, nous avons expliqué en détail les concepts de base de notre étude, ce qui sont le traitement automatique du langage naturel et l'analyse des sentiments. De plus, nous avons également présenté les différents domaines d'application de ces derniers. En outre, nous avons cité quelques travaux connexes qui sont reliés dans notre projet. Dans le chapitre suivant, nous allons introduire les concepts et les méthodes de l'apprentissage automatique.

Chapitre III

Conception et Modélisation

III.1. Introduction

Ce travail porte sur la tâche d'analyse du sentiment des tweets écrites dans la langue Arabe. Pour atteindre cet objectif et pour obtenir la meilleure performance possible, nous avons suivi la méthodologie présentée au-dessous, présentant la conception de notre processus en commençant par sa conception générale puis sa conception détaillée.

III.2. Description détaillé de l'objective

Notre objectif principal est d'appliqué les algorithmes de classification pour analyser les données de Hirak Algérien. À l'aide d'un processus qui s'appelle le processus d'analyse des sentiments ce processus permet d'analyser les commentaires twitter ou « les tweets » comme ayant des sentiments positifs, négatifs ou neutre en utilisant l'apprentissage automatique. Le processus d'analyses des sentiments contient cinq étapes nécessaires on commence par la phase de la collection de données puis la phase de prétraitement par suite en fait une implémentation en utilisant quelques approches classiques comme naïve bayésien, arbre de décision et Machines à Support Vectorielle et ont conclu par une évaluation.

III.3. La conception générale du système

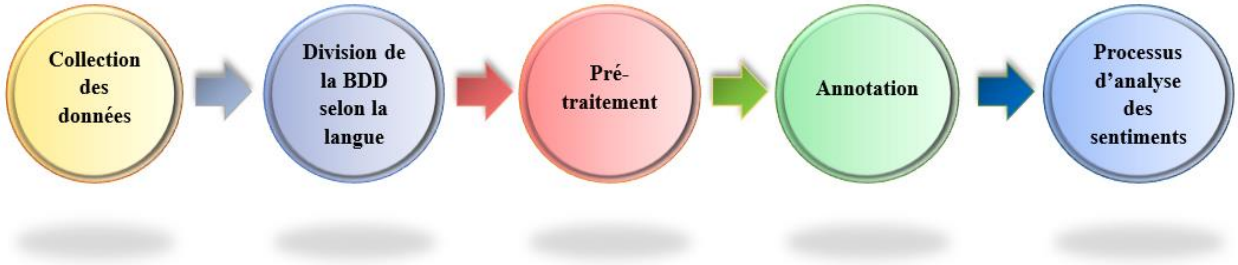

Figure 2. Conception générale du notre système.

III.4. La conception détaillée du système

III.4.1. La collection des données

La première étape du processus d'analyse des sentiments consiste à collecter des tweets en spécifiant un mot-clé pour récupérer tous les tweets liés à ce mot-clé. Les tweets peuvent être collectés à partir de différentes sources. L'un des moyens est le robot d'exploration de tweets qui collecte une collection de tweets liés en interrogeant le service Web Twitter. D'une autre façon en peut d'utiliser l'interface de programme d'application Twitter (API). Cette API fournie par Twitter et qui donne aux développeurs la possibilité d'utiliser les fonctions de Twitter telles que la récupération de tweets avec le mot-clé et la langue sélectionnés [8]. Les tweets collectés seront stockés dans une base de données afin de les classer. Dans notre cas les données sont déjà collectées et sauvegardées par les anciens étudiants dans un fichier CSV qui s'appelle " le 22-02-2019_labeled.csv". Cet ensemble de données compris 16087 tweets. La **Figure 3** représente la base de données collecté.

	Unnamed: 0	user	Date	Text	retweets	Target
0	0	nourmanpo	2019-02-22 23:59:46+00:00	@APS_DZ singulièrement parle des manifestation...	0.0	1
1	0	nourmanpo	2019-02-22 23:59:46+00:00	@APS_DZ singulièrement parle des manifestation...	0.0	1
2	0	newsemaratyah	2019-02-22 23:59:07+00:00	... تظاهرات في الجزائر احتجاجاً على ترشح بوتفليقة	0.0	1
3	2	karim_serier	2019-02-22 23:57:58+00:00	...حراك_22_فيفري #لا_للعهدة_الخامسة ماك هواري من#	2.0	1
4	2	karim_serier	2019-02-22 23:57:58+00:00	...حراك_22_فيفري #لا_للعهدة_الخامسة ماك هواري من#	2.0	1
...	...	...	...	...	...	...
16081	1772	Asmma_rk	2019-01-01 19:56:11+00:00	... لاله نعيمة صالحي تصرح بصحانية الوجه عيني عينك	0.0	0
16082	1773	pfinorgdz	2019-01-01 18:41:04+00:00	...بعد الفوز الكاسح لحزب جبهة التحرير الوطني في ا	0.0	0
16083	1774	Zakariabzd	2019-01-01 17:06:23+00:00	... بفضل مجهودات فخامة الرئيس المجاهد المعاصر Oui	0.0	0
16084	1775	MRezagui	2019-01-01 16:15:13+00:00	...اتمنى ان تكون السنة الميلادية 2019 عام خير وعا	0.0	0
16085	1776	Middleeast_code	2019-01-01 12:12:35+00:00	...الجزائر إلى مستقبل واعد وإقتصاد متنوع وغير نفط	0.0	0

16086 rows × 6 columns

Figure 3. La collection des données.

III.4.2. La division de la base de données

Dans cette phase, nous avons divisé la base de données en deux parties selon la langue. Une base de données qui contient la langue arabe et l'autres qui contient la langue française. Pour faire ça nous avons créé une fonction qui s'appelle « CheckLanguage ». Cette dernière a appliqué sur la colonne « Text » pour vérifier si ce tweet appartint à la langue arabe ou la langue français. Le résultat de cette phase est un deux fichiers de types csv qui sont appelées « Arabic.csv » et « French.csv ». La **Figure 4 et 5** représentent les BDD selon la langue.

	Unnamed: 0	Unnamed: 0.1	Unnamed: 0.1.1	user	Date	Text	retweets	Target	Type
0	1	2	0	newsemaratyah	2019-02-22 23:59:07+00:00	... تظاهرات في الجزائر احتجاجاً على ترشح بوتفليقة	0.0	1	A
1	2	3	2	karim_serier	2019-02-22 23:57:58+00:00	...حراك_22_فيفري #لا_للعهدة_الخامسة مات هواري من#	2.0	1	A
2	3	5	3	AbdellahBoudia	2019-02-22 23:57:47+00:00	...حراك_22_فيفري اليوم خرج الشارع للاحتجاج ضد ال#	0.0	1	A
3	4	6	4	hjlk_gf	2019-02-22 23:57:27+00:00	...تحية للشعب الجزائري و الامن لحمي تشوك #حراك_22	1.0	1	A
4	5	7	3	i_AhmedMu___	2019-02-22 23:57:20+00:00	...ممتاز ممتاز يا حبايبنا في الجزائر #سقط_بس #حر	0.0	1	A
...	...	...	...	...	...	...	...	...	...
12163	13293	16081	1772	Asmma_rk	2019-01-01 19:56:11+00:00	... لالة نعيمة صالحي تصرح بصحابية الوجه عيني عينك	0.0	0	A
12164	13294	16082	1773	pflnorgdz	2019-01-01 18:41:04+00:00	...بعد الفوز الكاسح لحزب جبهة التحرير الوطني في ا	0.0	0	A
12165	13295	16083	1774	Zakariabzd	2019-01-01 17:06:23+00:00	... بفضل مجهودات فخامة الرئيس المجاهد المعاصر Oui	0.0	0	A
12166	13296	16084	1775	MRezagui	2019-01-01 16:15:13+00:00	...اتمنى ان تكون السنة الميلادية 2019 عام خير وعا	0.0	0	A
12167	13297	16085	1776	Middleeast_code	2019-01-01 12:12:35+00:00	...الجزائر إلى مستقبل واعد وإقتصاد متنوع وغير نفط	0.0	0	A

12168 rows × 9 columns

Figure 4. La base qui contient des tweets en Arabe.

	Unnamed: 0	Unnamed: 0.1	Unnamed: 0.1.1	user	Date	Text	retweets	Target	Type
0	0	0	0	nourmanpo	2019-02-22 23:59:46+00:00	@APS_DZ singulièrement parle des manifestation...	0.0	1	F
1	11	15	6	MooeRaad	2019-02-22 23:53:43+00:00	Quand les manifestants trollent les flics qui ...	431.0	1	F
2	12	16	8	alidicabbio	2019-02-22 23:53:17+00:00	Next step الخطوة التالية prochain etape GREVE...	0.0	1	F
3	24	28	11	MinaMim32192553	2019-02-22 23:46:59+00:00	le seul héro depuis tjrs .. le peuple je ss vr...	0.0	1	F
4	26	31	18	alidicabbio	2019-02-22 23:46:28+00:00	RESPECT LES HOMMES #حراك_22_فيفري...	3.0	1	F
...	...	...	...	...	...	...	...	...	...
995	9680	12141	1826	lounes_dfn	2019-02-15 20:42:50+00:00	c'est drôle mais au fond ça fait vraiment mal ...	0.0	0	F
996	9719	12186	1838	zahrabmz	2019-02-15 19:53:37+00:00	Je continue à le faire, on sait jamais d'où vi...	0.0	0	F
997	9720	12187	1839	aimen_bouri	2019-02-15 19:52:31+00:00	#Rachid_Nakkaz visited our City Saida & ac...	4.0	0	F
998	9722	12189	3400	soumia_chebab	2019-02-15 19:51:57+00:00	Pourquoi ce n'est pas en tendance ??? #22_حراك...	0.0	0	F
999	9723	12190	1840	Inal_Boukhalfa	2019-02-15 19:51:09+00:00	Les supporteurs de #CRB .... #Algérie #لا_للعه...	58.0	0	F

Figure 5. La base qui contient des tweets en Français.

III.4.3. Annotation

L'objectif de l'étape d'annotation consiste à associer à chaque message une étiquette de polarité (positive, négative, neutre) qui représente son sentiment. Pour annoter les différentes commentaires collectes de la base de données, nous proposons d'utiliser la méthode manuelle. En fonction de la diversité de l'AlgD due aux différentes régions et accents, nous avons opté pour une annotation manuelle qui est le plus fiable que l'automatique. Pour accélérer le processus d'annotation, nous proposons d'utiliser une méthode de crowdsourcing basées sur une double annotation, ou deux personnes annotent les mêmes commentaires afin d'être sûr de l'étiquette donnée [18]. Finalement, on a obtenu un corpus qui contient 10000 tweets annotés. La **Figure 6** représente notre base annotée.

	Unnamed: 0	Unnamed: 0.1	Unnamed: 0.1.1	user	Date	Text	retweets	Target	Type	Annotation
0	1	2	0	newsemaratyah	2019-02-22 23:59:07+00:00	تظاهرات في الجزائر احتجاجاً على ترشح بوتفليقة ...	0.0	1	A	positive
1	2	3	2	karim_serier	2019-02-22 23:57:58+00:00	حراك_22_فيفري #لا_للعهدة_الخامسة مات# هواري من...	2.0	1	A	neutre
2	3	5	3	AbdellahBoudia	2019-02-22 23:57:47+00:00	حراك_22_فيفري اليوم خرج الشارع للاحتجاج# ضد ال...	0.0	1	A	neutre
3	4	6	4	hjlk_gf	2019-02-22 23:57:27+00:00	تحية للشعب الجزائري و الامن لحمي شوك #حراك_22...	1.0	1	A	positive
4	5	7	3	i_AhmedMu___	2019-02-22 23:57:20+00:00	ممتاز ممتاز يا حبائبنا في الجزائر #سقط_بس #حر...	0.0	1	A	positive
...	...	...	...	...	...	...	...	...	...	...
9994	11064	13702	563	GcYXHNgot5r3Un0	2019-02-12 16:41:06+00:00	لا اعتراض على ترشح بوتفليقه بس يظهر هو و يتكلم...	0.0	0	A	positive
9995	11065	13703	564	tsaarabi	2019-02-12 16:36:15+00:00	الجزائر – 13 حزبا يطلبون مساندة ترشح بوتفليقة ...	0.0	0	A	neutre
9996	11066	13704	565	aboutahahakim1	2019-02-12 16:34:52+00:00	وانا أقرأ خبر ترشح بوتفليقة لولاية 5 في الجزائ...	0.0	0	A	positive
9997	11067	13705	410	EslamSalihS	2019-02-12 16:34:32+00:00	عبد العزيز بوتفليقة	1.0	0	A	neutre
9998	11068	13706	566	NizarAburabie	2019-02-12 16:31:56+00:00	ترشح بوتفليقة ليس خيار لمن رشحوه بل لم يجدوا ل...	0.0	0	A	positive

Figure 6. Annotations des tweets.

III.4.4. Application du processus d'analyse des sentiments

III.4.4.1. Le pré-traitement

Cette étape est au cœur de notre approche d'analyse des sentiments. Avant de commencer la classification des messages en positifs, négatifs ou neutres, un ensemble de données propre doit être fourni au modèle d'apprentissage automatique afin d'obtenir un modèle de classification puissant avec un score plus élevé. Comme nous l'avons remarqué, la plupart des commentaires récupérés sur les réseaux sociaux ne sont pas préparés et non nets, avec des variations sémantiques et syntaxiques car ils sont écrits par différentes personnes de différents niveaux intellectuels. De plus, la richesse de cette base de base de données conduit à une variété de mots et à une variété de sens pour un même mot selon chaque région. Par ailleurs, nous avons identifié d'autres problèmes dans les commentaires comme : les fautes d'orthographe, la présence de : liens, hashtags, gifs, autocollants, caractères spéciaux, etc., qui doivent impérativement être supprimés et ces messages doivent être dédupliqués afin d'avoir un contenu minimal, unifié, corpus valide et propre prêt à être exploité. Notre objectif est de conserver un maximum de variantes de vocabulaire informatif afin d'enrichir notre corpus. Dans ce qui suit nous détaillerons les étapes de pré-traitement [18]:

III.4.4.1.1 La conversation des données textuelles en minuscules

Cette étape consiste de convertir tous les tweets en minuscule. Pour faire cette étape nous utilisons la fonction "**Lower()"** en Python . Cette dernière convertie tous les caractères majuscules. Le but de cette étape est d'avoir une base de données cohérent (les données français et anglais) cohérentes dans les minuscules à la fin de résoudre le problème de sensibilité à la case [19]. Le résultat est dans la **Figure 7**.

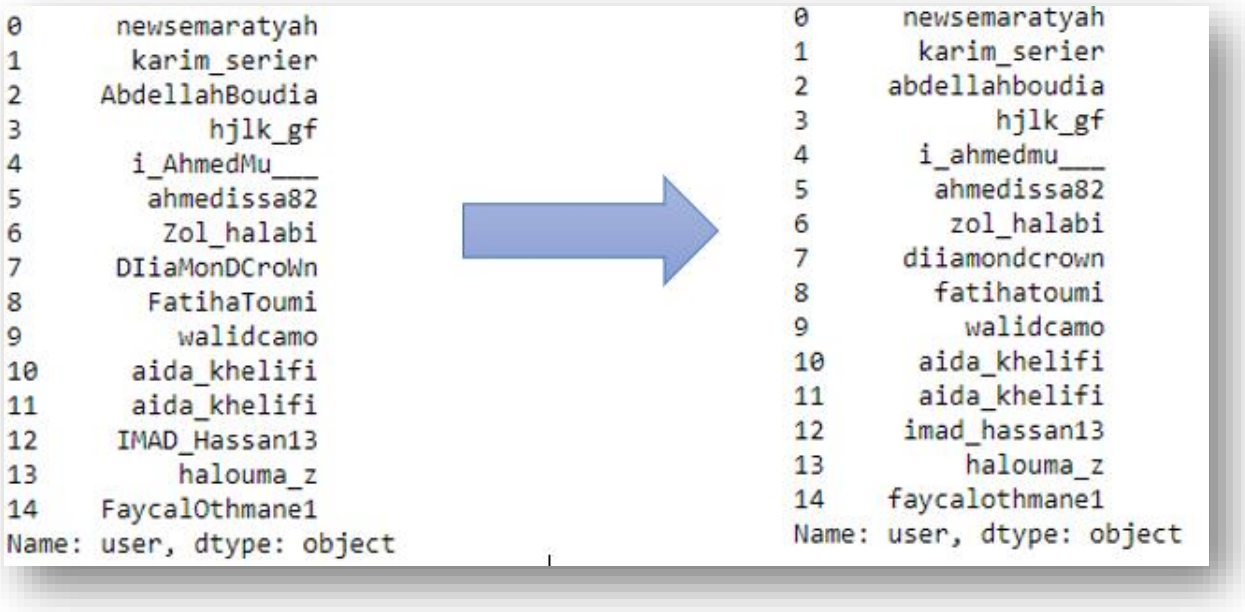

Figure 7. La conversion de majuscule en minuscule.

III.4.4.1.2 Le nettoyage des données

Le nettoyage de données joue un rôle essentiel dans la phase de pré-traitement, il consiste de supprimer et éliminer toutes les données non nécessaires dans le fichier csv. Ces données comme les hashtags, les émojis, les URL, les mentions, des tweets répètes, les dates, les ponctuations et les caractère spéciaux… etc. Pour faire l'étape de nettoyage en besoin d'importer la bibliothèque "**re**" en python pour obtenir une base de données son bruit et avec une taille plus réduite [19]. Le résultat obtenu est dans la **Figure 8, 9,** et le **Tableau 1**.

4.4.1.2.a. La suppression des numéros, des mentions, des URL, des emojis, des caractères spéciaux et des ponctuations

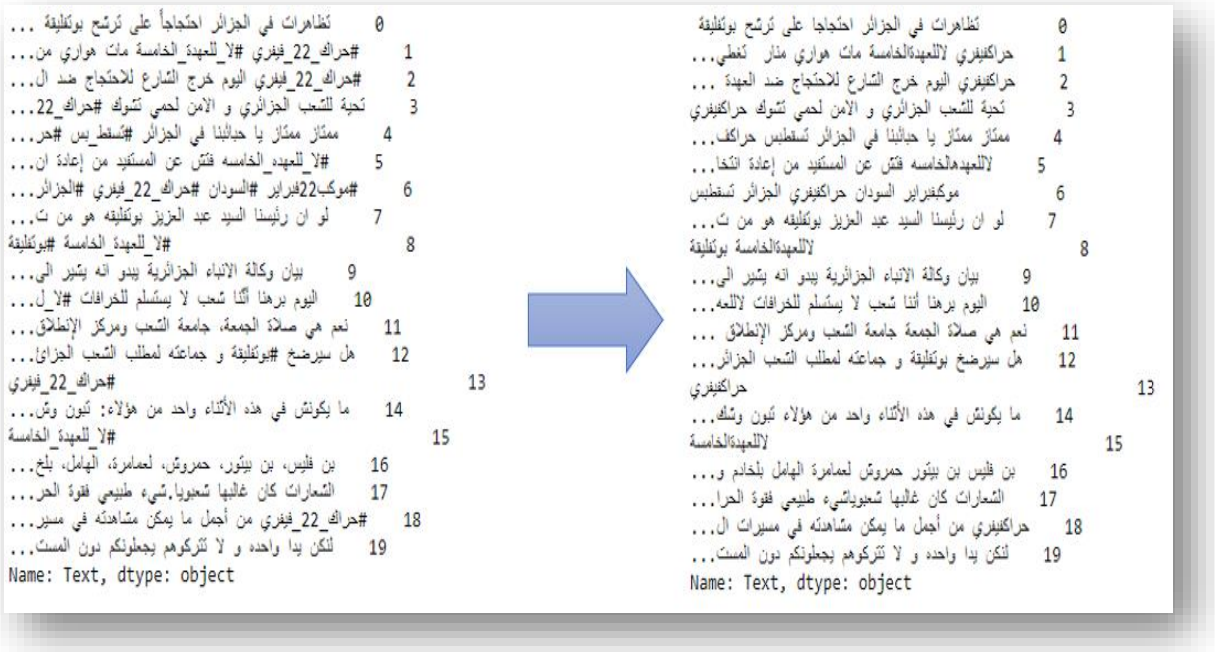

Figure 8. La suppression des numéros, des mentions, des URL, des emojis, des caractères spéciaux et des ponctuations.

4.4.1.2.b. La suppression des hashtags

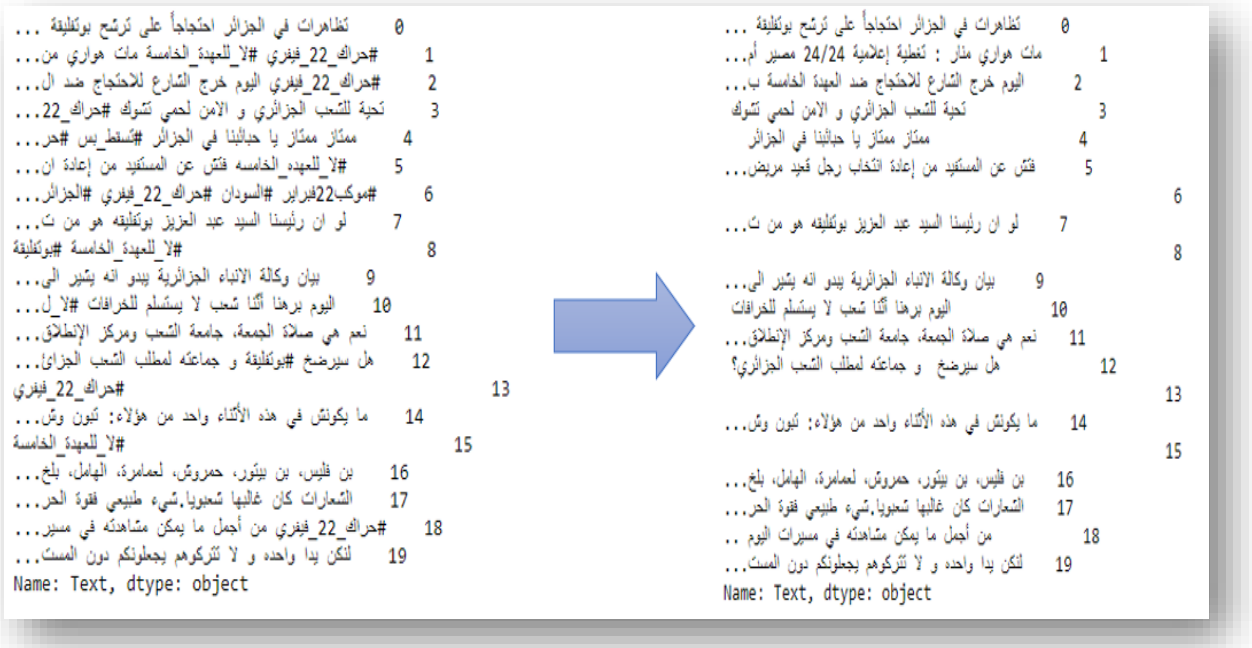

Figure 9. Suppression des hashtags.

4.5.1.2.c. La suppression des tweets répétés

Dans cette phase nous avons supprimé tous les tweets répétés écrit par l'utilisateurs plus d'une fois pour réduire la taille de fichier. Pour faire ça nous avons utilisé la fonction **Drop_duplicates**. Ce dernier est applique sur la colonne « user » et la colonne « Text » dans la base des données, le résultat obtenu est un fichier nommé « result of 22-02-2019.csv » avec une taille réduite estimé à environ 9683.le résultat de cette suppression est représenté dans le **Tableau 1** ci-dessous.

Tableau 1. La suppression des tweets répétés.

Evènements	***Nombre des tweets avant le prétraitement***	***Nombre des tweets après le prétraitement***
Le 22-02-2019	10000	9771

.

III.4.4.1.3. Suppressions des mots d’arrêt

Un mot d’arrêt est un mot inutile et non significatif apparaissant dans un texte, d'où la nécessité de l'éliminer de notre corpus. A cet effet, pour des langues telles que l'anglais, le français ou l'arabe standard moderne, il existe des listes de mots d’arrêt bien connus. Ces listes/outils sont disponibles gratuitement comme NLTK1. Néanmoins, il n'y a pas de ressource définie ou élaborée pour les mots vides d'AlgD à considérer d'où l'obligation de les créer. La difficulté réside dans l'identification de tous les mots d’arrêt ou sans signification ajoutée afin de les éliminer plus tard de notre corpus. En Algérie nous avons plusieurs accents, et cela nécessite de collaborer avec une ressource humaine conséquente et expérimentée en linguistique et dialectes pour englober tous ces mots vides. Pour faciliter notre travail, nous avons suivi une approche visant à créer une liste vide générale pour le dialecte et qui peut être utilisée comme source fiable [19] [18]. Le résultat est dans la **Figure 10**.

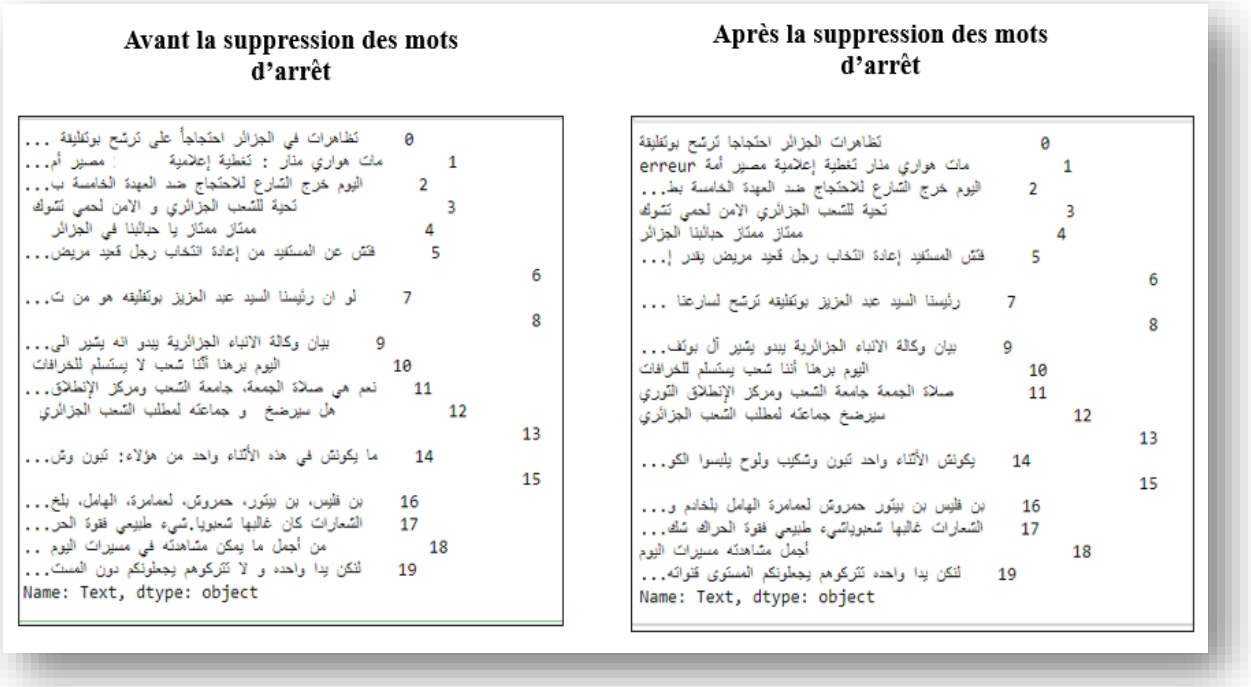

Figure 10. Suppression des mots d’arrêt.

III.4.4.1.4. La tokenisation

Cette étape permet de décomposer une chaîne de caractères (message ou commentaire) en mots appelés « Tokens ». La tokenisation est encore plus importante dans l'analyse des sentiments que dans d'autres domaines de la NLP, car les informations sur les sentiments sont souvent mal représentées [19]. La **Figure 11** représente la tokenisation.

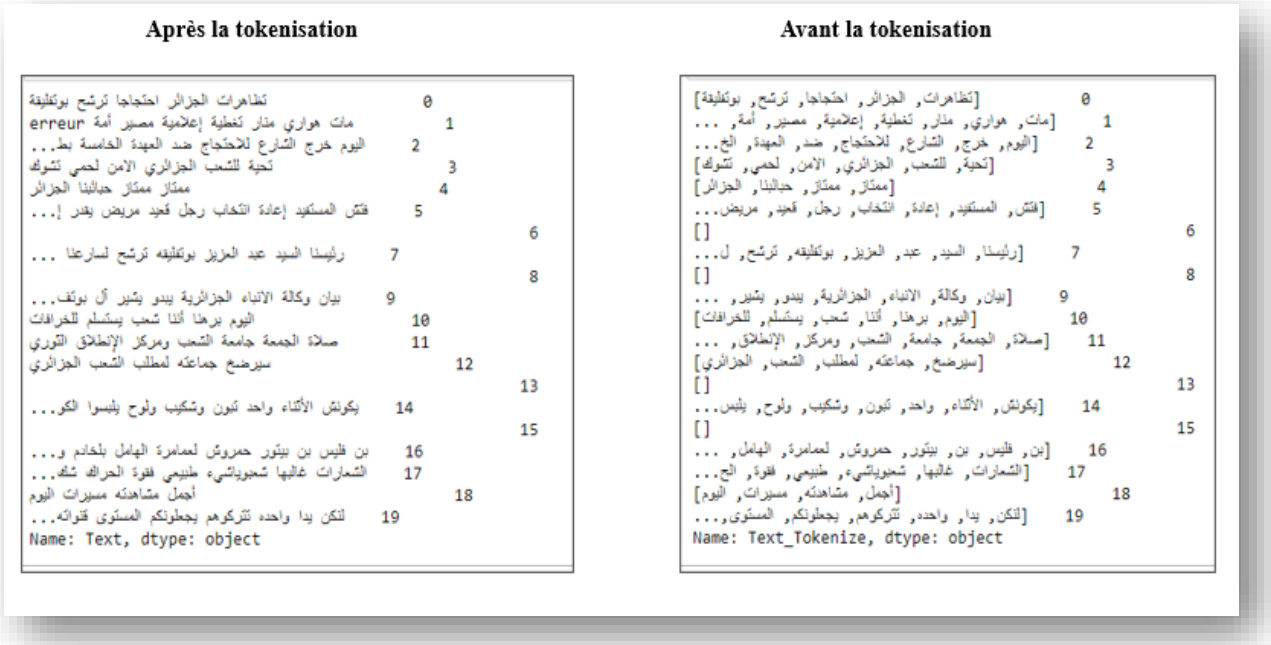

Figure 11. La tokenisation.

III.4.4.1.5. Le mot du nuage

Le nuage de mots est la représentation graphique des mots les plus fréquemment répétés représentant la taille du mot [19]. Pour créer notre nuage de mot, nous avons utilisé trois bibliothèques sont :

- La bibliothèque WordCloud.
- La bibliothèque arabic_reshaper
- La bibliothèque bidi.algorithm

III.4.4.2. Feature extraction

L'extraction de caractéristiques est une tâche concernant la transformation de données brutes en entrées appropriées (c'est-à-dire des caractéristiques) qui peuvent être consommées par un algorithme d'apprentissage automatique particulier. De manière expresse, les caractéristiques extraites doivent représenter le contenu textuel principal dans un format qui correspondra le mieux aux besoins de l'algorithme de classificateur sélectionné. Pour faire l'extraction des caractéristiques. Nous avons adopté deux méthodes sont [20] :

III.4.4.2.1. La méthode Sac des Mots

Sac des mots est une méthode utilisée dans le traitement du langage naturel et la recherche d'informations. Dans cette méthode chaque Test soit (phrase/ commentaire / tweet ou bien un document) est représenté sous la forme d'un vecteur numérique où chaque dimension est un mot spécifique du corpus et la valeur peut être une fréquence dans le document, une occurrence (notée 1 ou 0) ou même des valeurs pondérées [2].

III.4.4.2.2 La méthode TF-IDF

La méthode TF-IDF (Term Frequency-Inverse Document Frequency) est un processus appliqué dans l'extraction de texte et la recherche d'informations. Ce processus TF-IDF est centré sur un mot de la collection ou du corpus de documents. Il représente la durée pendant laquelle un mot apparaît dans un document, le numéro du document avec ce mot particulier et le rapport entre les documents avec ce terme par tous les documents. Cette méthode est utilisée dans la suppression des mots d'arrêt, des mots haute fréquence et basse fréquence. Il est également utilisé dans la synthèse et la classification de texte [21].

III.4.4.3. La classification

Dans cette phase, nous avons basées sur la classification supervisée, à partir du corpus annoté et traité. Pour accomplir cette tâche nous avons implémenté les algorithmes de classification supervisée les plus populaires comme Naïve Bayes, machine à support vectorielle, arbre de décision et l'algorithme de régression logistique.

L’objectif de la classification est de trouver la meilleure combinaison de paramètre donnant le meilleur score pour chaque algorithme.

III.4.4.3.1. Naïve Bayes

Naïve Bayes est une technique d'apprentissage automatique utilisée pour classer le texte dans des catégories prédéfinies en fonction de caractéristiques similaires. Le classificateur Naïve Bayes a été appliqué pour améliorer le traitement et la manipulation de textes ou d'informations provenant de différentes sources. Cet algorithme représente une méthode probabiliste. En d'autres termes, le classificateur Naïve Bayes suppose que l'absence de caractéristique de classe n'est pas liée à l'absence d'autres caractéristiques. Ce classificateur est couramment utilisé pour classer les documents en raison d'une bonne performance de classification, calcule la probabilité des documents liés à les classer dans différentes classes, puis les attribue à la classe spécifique avec la probabilité la plus élevée. Comme beaucoup d'autres modèles, le classificateur Naïve Bayes présente de nombreux avantages. Il est généralement considéré comme le modèle le plus puissant utilisé dans ce domaine. Ce classificateur est compréhensible et très simple à mettre en œuvre. Dans autre part, Ce classificateur souffre de certaines limitations telles que l'occurrence de la classe, car dépend de la probabilité, alors que la probabilité dépend généralement de la fréquence [22].

III.4.4.3.2. Machines à Support Vectorielle

Machines à Support Vectorielle est l'un des modèles d'apprentissage supervisé qui ont été appliqués pour classification du texte. Il classe les différents objets et documents dans un espace de dimension finie. La machines à support vectorielle est également utilisé pour analyser des données, des textes et des documents afin de calculer la similitude entre eux. Ce model montre différents aspects utiles en tant que modèle important utilisé en informatique. Premièrement, cette méthode se défend sur une algèbre linéaire, où elle ne contient aucune équation algébrique complexe. Parmi l’un des avantages de machines à support vectorielle est l'efficacité des poids attribués aux concepts ou aux termes. Ce modèle montre également une facilité particulière par rapport à d'autres méthodes. Cela permet à la machine de calculer la similitude entre les documents. D’une autre coté, la machines à support vectorielle contient certaines limitations qui empêchent certains chercheurs de l'utiliser. La difficulté d'utiliser des

synonymes en arabe représente un domaine très difficile, où la langue arabe a de nombreux synonymes pour chaque mot ou concept. D'autres limitations supposent que les termes sont statistiquement indépendants. Alors que la plupart des termes arabes ont une relation étroite avec d'autres termes [22].

III.4.4.3.3. Arbre de décision

Un arbre de décision est une classification supervisée approche et construit à partir de nœuds qui représentent des cercles et les branches sont représentées par les segments qui relient les nœuds. Un arbre de décision commence à la racine, se déplace vers le bas et est généralement dessiné de gauche à droite. Le nœud à partir duquel l'arbre commence est appelé nœud racine. Le nœud où se termine la chaîne est appelé nœud « feuille ». Deux branches ou plus peuvent être étendues à partir de chaque nœud interne, c'est-à-dire un nœud qui n'est pas un nœud feuille. Un nœud représente une certaine caractéristique tandis que les branches représentent une plage de valeurs. Ces plages de valeurs agissent comme des points de partition pour l'ensemble de valeurs de la caractéristique donnée. Le regroupement des données dans l'arbre de décision est basé sur les valeurs des attributs des données. Cet arbre de décision est réalisé à partir des données pré-classifiées. D'une autre coté, la division en classes est décidée sur les caractéristiques qui divisent le mieux les données. Les éléments de données sont divisés en fonction des valeurs de ces caractéristiques. Ce processus est appliqué à chaque sous-ensemble fractionné des éléments de données de manière récursive. Le processus se termine car tous les éléments de données du sous-ensemble actuel appartiennent à la même classe [23].

III.4.4.3.4. Régression logistique

La régression logistique est l'un des classificateurs les plus célèbres dans les mondes de la statistique, de la science des données et de l'apprentissage automatique. Pour les données de faible dimension, la régression logistique est une approche standard pour la classification binaire. Cela est particulièrement vrai dans les domaines scientifiques tels que la médecine, la psychologie et les sciences sociales où l'accent est mis non seulement sur la prédiction mais aussi sur l'explication. Il existe également la version multinomiale de la régression logistique qui peut être utilisée pour modéliser des réponses non binaires (multi-catégories) [24].

III.4.4.4. Évaluation

Dans cette phase nous avons évalué les performances sur les différents modèles d'apprentissage automatiques qui nous avons utilisé dans notre projet. Pour faire cette phase nous avons utilisé les différentes mesures comme la précision, le rappel, L'exactitudes et le score f1. Ces derniers est basé sur une matrice qui s'appelle la matrice de confusion. La matrice des confusions [24] est également appelée table de contingence ou matrice d'erreur est utilisée pour présenter le résultat du classificateur pour la prédiction. C'est une table spéciale pour visualiser les performances du modèle, ce dernier est représenté dans le **Tableau 2**.

Tableau 2. Matrice de confusion.

Classe actuel	**Classe prédiction**	
	Positive	**Négative**
Positive	TP	FP
Négative	FN	TN

Vrai Positive (TP) : nombre de tweets positifs classés correctement.

Faux positive (FP) : nombre de tweets négatifs classés à tort comme positifs.

Vrai Négative (TN) : nombre de tweets négatifs classés correctement.

Faux Négative (FN) : nombre de tweets positifs classés à tort comme négatifs.

III.4.4.4.1. Précision

La précision [25] est également appelée valeur prédite positive, mesure la justesse du modèle. Une précision plus élevée indique moins de FP. Mathématiquement, il est défini comme :

$$\text{La précision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

III.4.4.4.2. Rappel

Le rappel [25] est également connu sous le nom de sensibilité, mesure les cas positifs correctement classés par le modèle, une valeur de rappel élevée signifie que peu de cas positifs sont mal classés comme négatifs. Le rappel peut être calculé à l'aide de la formule suivante.

$$\text{Le rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

III.4.4.4.3. Exactitudes

L'exactitudes utilisée comme mesure pour les techniques de catégorisation. Les valeurs de exactitudes, cependant, sont beaucoup moins réticentes aux variations du nombre de décisions correctes que la précision et le rappel. Cette technique est représentée sous la forme suivante [26]:

$$\text{L'exactitude} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

III.4.4.4.4. Le score F1

Le score F1 ou la mesure F1 [25] est la moyenne harmonique de la précision et du rappel. Le score F peut être calculé comme suit :

$$\text{score F1} = 2 * \frac{\text{La précision} \times \text{Le rappel}}{\text{La précision} + \text{Le rappel}} \tag{4}$$

III.5. Conclusion

Dans ce chapitre, nous avons introduit notre objective de projet, puis nous avons fait une étude détaille sur les différentes phases de processus d'analyse des sentiments. Nous avons commencé par une explication générale sur notre système puis nous avons appliqué ces phases dans notre base de données, on commence par le prétraitement de la base de données et nous terminons par évaluer chacun des algorithmes des classifications appliqués à la base de données. Le résultat obtenu sera discuté dans le chapitre suivante.

Chapitre IV

Implémentation et Résultats

IV.1. Introduction

Dans ce chapitre, nous allons présenter les outils matériels et logiciels que nous avons utilisé pour réaliser ce projet. De plus, nous allons introduit en détaille une analyse exploratoire de données sur les résultats que nous avons obtenue.

IV.2. L'environnement de travail et les outils utilisés

IV.2.1. L'environnement Matériel

Pour réaliser notre projet, nous avons utilisé deux PC marque Lenovo et Compaq, équipés d'un processeur multi-Core i3 et Intel, avec 4 et 2 Gi Octets de RAM respectivement.

IV.2.2. L'environnement Logiciel

Pour atteindre notre but, nous avons utilisé le langage de programmation Python, version 3.7. Python est relativement simple à prendre, open source, gratuit, interprété et le langage le plus employé par les informaticiens récemment. Ce dernier est développé depuis 1989 par Guido Var Rossum. Pour se focaliser sur notre projet et tirer profit des puissances du langage Python, nous avons utilisé les outils suivants :

Nous avons utilisé Jupyter notebook comme un éditeur et de divers packages comme :

- Package pandas.
- Package CSV.
- Package re.
- Package NLTK.
- Package Numpy.
- Package Matplot.
- Package Seabarn.
- Package Scklearn.

IV.3. Présentation des données

Pour effectuer l'analyse des sentiments, nous avons utilisé une base de données qui contient sept d'attributs avec de 10000 d'enregistrements des internautes sur le mouvement social en Algérie. Cette base de données est extraite par les anciens collègues. Pour faire cette étape, les anciens collègues ont utilisé une bibliothèque appelée **GetOldTweets3.**

À l'aide de **GetOldTweets3**, les anciens collègues pouvant extraire des tweets à l'aide de divers paramètres de recherche tels que les dates et les heures, le(s) nom(s) d'utilisateur, le nombre de retweets et le texte du tweet. Le **Tableau 3** représente les différents attributs dans notre base de données.

Tableau 3. Les différents attributs dans la base de données.

User	Date	Text	Retweet	Target

Pour notre projet, nous avons ajouté deux colonnes dans la base de données qui sont la colonne « type » et la colonne « Annotation ». Puis nous avons focalisé sur deux attributs qui sont : « Text » et « Annotation ». **Tableau 4** présente une description sur les attributs dans notre base de données.

Tableau 4. Description des attributs du base de données utilisé.

Attributs	*Description*
User	Il représente l'identificateur de chaque utilisateur.
Date	C'est la date et l'heure auxquelles le tweet a été tweeté par l'utilisateur.
Text	Il se compose des avis donnés par chaque utilisateur individuel.
Retweet	Il présente le nombre de fois que ce Tweet a été retweeté par l'utilisateur.
Target	Il indique les gens qui sont sortie pour les manifestations ou non.
Type	Il sépare le texte en arabe ou en français.
Annotation	Il contient des sentiments positifs, négatifs ou neutres.

Pour notre projet, l'attribut « Annotation » est divisé en trois classes selon les avis donnés par chaque utilisateur individuel. Le résultat de cette division est trois classes d'attributs de sentiments, à savoir positif, négatif ou neutre. La **Figure 12** montre la répartition des avis positifs, négatifs et neutres dans notre base de données. Il y a 5846 avis positifs, 1741 avis négatifs et 2184 avis neutres dans l'ensemble de données.

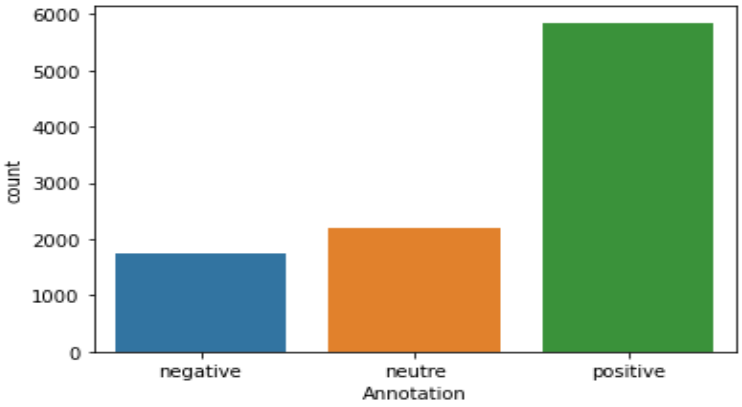

Figure 12. La répartition des avis positifs, négatifs et neutres.

IV.4. Prétraitement

Dans cette section, nous avons utilisé des différents package comme nous avons dit dans le chapitre 3 pour éliminer et supprimer tous les bruits comme les hashtags, mentions, les ponctuations. De l'autre part, nous avons éliminé aussi les mots d'arrêts pour obtenir un texte brut. Dans cette étape nous nous concentrerons sur le mot de Nuage (Word Cloud). Ce dernier s'agit d'une représentation visuelle des mots-clés les plus utilisés dans les tweets. En général, les mots sont affichés dans des tailles de police et des poids qui sont les plus significatifs lorsqu'ils sont utilisés ou populaires. La **Figure 13** représente notre mot du nuage.

Figure 13. Le nuage de mot.

IV.5. Classification

Avant de classer les sentiments en classe positive, négative ou neutre, nous avons fait une division train-test sur la base de données en 30% des données test, et en 70 % des données pour l'entrainement.

Les **Tableaux 5** et **6** ci-dessus contient les détails des résultats expérimentaux menés dans notre projet. Ces résultats ont obtenu à l'aide des classificateurs SVM, RL, DT et NB pour classer les sentiments en classes positive, négative et neutre. Pour chaque classificateur nous avons appliqués les deux méthodes d'extraction des attributs qui sont : le sac de mots (BOW) et le TF-IDF. Les **Tableaux 5** et **6** représentent les résultats d'exactitude des classificateurs utilisant le TF-IDF et le BOW respectivement.

Tableau 5. Le résultat d'exactitude des classificateurs avec le TF-IDF.

Classificateurs	Accuracy	Classe positive			Classe neutre			Classe négative		
		Précision	Recall	F1	Précision	Recall	F1	Précision	Recall	F1
SVM	0,67	0,72	0,87	0,79	0,48	0,34	0,40	0,61	0,42	0,50
NB	0,62	0,61	0,99	0,76	0,50	0,02	0,03	0,83	0,10	0,18
LR	0,67	0,67	0,95	0,79	0,53	0,20	0,29	0,74	0,29	0,42
DT	0,60	0,69	0,78	0,73	0,38	0,31	0,35	0,45	0,34	0,39

Pour les résultats qui nous avons obtenu avec le TF-IDF, nous notons que la précision est élevée dans la classe négative avec le classificateur NB par rapport les autres classes et les autres classificateurs avec une valeur égale à 0,83. D'une autre part nous notons que le recall est élevée dans la classe positive avec le même classificateur qui est le classificateur NB par rapport la classe négative et la classe neutre et les autres classificateurs avec une valeur égale à 0,99. Pour le F1-score, le F1-score est élevé dans la classe positive mais avec deux classificateurs qui sont SVM et LR par rapport les autres classes et les autre classificateurs avec une valeur égale à 0,79. Par contre le **Tableau 6** ci-dessous représente le résultats d'exactitude des classificateurs avec BOW.

Tableau 6. Le résultat d'exactitude des classificateurs avec le BOW.

Classificateurs	Accuracy	Classe positive			Classe neutre			Classe négative		
		Précision	Recall	F1	Précision	Recall	F1	Précision	Recall	F1
SVM	0,65	0,72	0,83	0,77	0,42	0,35	0,38	0,59	0,41	0,48
NB	0,66	0,69	0,92	0,79	0,51	0,23	0,31	0,60	0,36	0,45
LR	0,66	0,71	0,89	0,79	0,45	0,29	0,35	0,62	0,37	0,46
DT	0,59	0,69	0,78	0,73	0,35	0,33	0,34	0,50	0,33	0,40

Ce tableau représente le résultat qui nous avons obtenue avec le BOW, nous notons que chacune des Précision, Recall et F1-score est élevés dans la classe positive par l'utilisation de chaque classificateur comme SVM, NB et LR par rapport les autres classes et les classificateurs avec des valeurs égalent à 0.72,0.92 et 0.79 respectivement pour SVM, NB et LR.

IV.6. Comparaison

D'après les résultats qui nous avons obtenu, nous avons observé que la meilleure classification était avec TF-IDF. Pour coté performance, le **Tableau 7** résume les performances selon les classificateurs utilisés. Cependant, force est de constater que les résultats sont proches. De plus, les scores de précision sont relativement faibles, ce qui signifie que la classification n'était pas assez précise. Une explication possible est le fait que nos données ne sont pas équilibrées. Le nombre de tweets positifs était plus élevé que les tweets négatifs et neutres.

Une autre raison possible qui a causé le faible taux de précision est les défis auxquels nous avons été confrontés lors du prétraitement du dialecte algérien. La manipulation du dialecte algérien est encore un sujet de recherche récent et en cours de développement.

Tableau 7. Comparaison des performances selon les classificateurs utilisés.

Classificateurs	Accuracy	
	TF-IDF	BOW
SVN	0,67	0,65
NB	0,62	0,66
LR	0,67	0,66
DT	0,60	0,59

IV.7. Conclusion

L'analyse des sentiments est un sujet très intéressant et utile à notre époque. Cependant, peu de recherches ont été faites concernant les textes mixtes algériens qui, nous croyons, nécessitent des méthodes spéciales pour fonctionner parfaitement dans notre contexte. Des recherches supplémentaires sont absolument nécessaires pour améliorer les résultats de la classification des sentiments.

Conclusion Générale

Conclusion Générale

L'Algérie a connu beaucoup de changements depuis le début de son mouvement social de 2019. L'importance de ce mouvement social (appelé Hirak) nous a poussé à l'analyser à travers les réseaux sociaux. Plus précisément, L'objective de ce projet était l'analyse des sentiments des tweets reliés aux Hirak Algérien en appliquant des algorithmes de classifications. Le projet était réalisé en utilisant le langage de programmation Python.

Ce rapport a fourni des étapes détaillées du processus que nous avons suivi, y compris des projets de recherche connexes, des définitions de concepts de base liés au traitement du langage naturel et à l'analyse des sentiments.

Bien que la réalisation de ce projet fût intéressante et fructueuse, nous avons rencontré quelques difficultés, et limites. Premièrement, L'annotation manuelle des tweets a été un long processus difficile et, dans une certaine mesure, subjectif et biaisé par nos opinions. Cela pourrait avoir influencé le taux de précision des algorithmes de classification. D'autre part, les circonstances de travail de Covid 19 ont rendu la collaboration très difficile et limitée. Au moment de la rédaction de ce rapport, nous étions confrontés à la troisième vague du virus qui a eu un très mauvais impact sur notre état physique et psychologique, limitant notre capacité à travailler.

Pour améliorer ce travail, et surmonter les limitations que nous avons mentionnées ci-dessus, nous vous proposons les solutions possibles suivantes.

- Utiliser une approche d'annotation différente telle que l'annotation automatique ou les annotations de crowdsourcing.

- Implémenter d'autres algorithmes de classification plus adaptés pour gérer des langues et des dialectes complexes.

- Proposant une nouvelle procédure de prétraitement spécialement dédiée au dialecte algérien mélangé à d'autres langues telles que l'anglais et le français qui sont largement utilisées par la société algérienne.

Références

Les Références

[1] Sghaier, M. A., Abdellaoui, H., Ayadi, R., & Zrigui, M. Analyse de sentiments et extraction des opinions pour les sites e-commerce: application sur la langue arabe. (2014).

[2] Sarkar, D. Text analytics with Python: a practitioner's guide to natural language processing. Apress. (2019).

[3] Bathulapalli, C., Desai, D., & Kanhere, M. Use of Sanskrit for natural language processing. Int. J. Sanskrit Res, 2(6), 78-81. (2016).

[4] Khurana, D., Koli, A., Khatter, K., & Singh, S. Natural language processing: State of the art, current trends and challenges. arXiv preprint arXiv:1708.05148. (2017).

[5] Godsay, M. The process of sentiment analysis: a study. International Journal of Computer Applications, 126(7). (2015).

[6] «Définition : Opinion,» [En ligne]. Available: https://www.toupie.org/Dictionnaire/Opinion.htm. Consulté le 08 aout 2021.

[7] Pak, A. Automatic, adaptive, and applicative sentiment analysis (Doctoral dissertation, Université Paris Sud-Paris XI). (2012).

[8] Alhumoud, S. O., Altuwaijri, M. I., Albuhairi, T. M., & Alohaideb, W. M. Survey on arabic sentiment analysis in twitter. International Science Index, 9(1), 364-368. (2015).

[9] Duwairi, R. M., Marji, R., Sha'ban, N., & Rushaidat, S. Sentiment analysis in arabic tweets. In 2014 5th International Conference on Information and Communication Systems (ICICS) (pp. 1-6). IEEE. (2014, April).

[10] Bandi, A., & Fellah, A. Socio-Analyzer: A Sentiment Analysis Using Social Media Data. In Proceedings of 28th International Conference (Vol. 64, pp. 61-67). (2019).

[11] Matalon, Y., Magdaci, O., Almozlino, A., & Yamin, D. Using sentiment analysis to predict opinion inversion in Tweets of political communication. Scientific reports, 11(1), 1-9. (2021).

Références

[12] Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. OCA: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology, 62(10), 2045-2054. (2011).

[13] Altawaier, M. M., & Tiun, S. Comparison of machine learning approaches on arabic twitter sentiment analysis. International Journal on Advanced Science, Engineering and Information Technology, 6(6), 1067-1073. (2016).

[14] Rezapour, R. Using Linguistic Cues for Analyzing Social Movements. *arXiv preprint arXiv:1808.01742*. (2018).

[15] Zamir, M. H. Anatomy of a Social Media Movement: Diffusion, Sentiment, and Network Analysis (Doctoral dissertation, University of South Carolina). (2017)

[16] Ruz, G. A., Henríquez, P. A., & Mascareño, A. Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer Systems, 106, 92-104. (2020).

[17] Tobaili, T., Fernandez, M., Alani, H., Sharafeddine, S., Hajj, H., & Glavas, G. Senzi: A sentiment analysis lexicon for the latinised arabic (arabizi). In International Conference Recent Advances In Natural Language Processing 2019 Natural Language Processing in a Deep Learning World: Proceedings (pp. 1204-1212). (2019).

[18] Chader, A., Lanasri, D., Hamdad, L., Belkheir, M. C. E., & Hennoune, W. Sentiment Analysis for Arabizi: Application to Algerian Dialect. In KDIR (pp. 475-482). (2019).

[19] Kulkarni, A., & Shivananda, A. Natural language processing recipes. Apress. (2019).

Références

[20] El Kah, A., & Zeroual, I. The effects of pre-processing techniques on Arabic text classification. International Journal, 10(1). (2021).

[21] Jain, S., Jain, S. C., & Vishwakarma, S. K. Text mining methods and techniques-A survey. (2019).

[22] Al Sbou, A. M. A survey of arabic text classification models. *International Journal of Electrical and Computer Engineering (IJECE)*, *8*(6), 4352-4355. (2018).

[23] Ali, J., Khan, R., Ahmad, N., & Maqsood, I. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272. (2012).

[24] Mohamed, E., & Mostafa, S. A. Computing Happiness from Textual Data. Stats, 2(3), 347-370. (2019).

[25] Kundi, F. M., Khan, A., Ahmad, S., & Asghar, M. Z. Lexicon-based sentiment analysis in the social web. Journal of Basic and Applied Scientific Research, 4(6), 238-48. (2014).

[26] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974. (2005).

Résumé

Depuis le 22 février 2019, des millions d'Algériens sont descendus dans les rues de toutes les grandes villes du pays pour exprimer leur rejet d'un cinquième mandat d'Abdelaziz Bouteflika. Ce mouvement social (appelé Hirak) a été diffusé dans divers médias sociaux tels que Twitter, où les internautes ont exprimé leurs opinions et sentiments qui différaient entre positifs, négatifs ou neutres. Le but de ce projet était d'analyser les sentiments et les tweets liés au mouvement algérien à travers une application des Algorithmes de classification tels que naïve bayésienne, machine a support vectorielle, arbre de décision et algorithme de régression logistique avec des différentes méthodes d'extraction des attributs qui sont « sac de mots » et « TF-IDF » sur une base de données qui contienne 10000 tweets divisé en 5846 avis positifs, 1741 avis négatifs et 2184 avis neutres. Le résultat expérimental a montré que le bon classificateur est le svm avec une précision raisonnable égal 67%.

Mots clés : *Hirak ; Analyse des sentiments ; Le traitement automatique du langage naturel ; Les algorithmes de classification ; Annotation manuelle ; L'apprentissage automatique ; Dialecte Algérien*.

Abstract

Since February 22, 2019, millions of Algerians have taken to the streets of all major cities in the country to express their rejection of a fifth term of Abdelaziz Bouteflika. This social movement (called Hirak) was broadcast on various social media such as Twitter, where netizens expressed their opinions and feelings which differed between positive, negative or neutral. The goal of this project was to analyze the sentiments and the tweets linked to the Algerian movement through an application of classification algorithms such as Bayesian naive, vector-supported machine, decision tree and logistic regression algorithm with different methods of extraction of the attributes which are « BOW » and « TF -IDF » on a dataset that contains 10,000 tweets divided into 5,846 positive reviews, 1,741 negative reviews and 2,184 neutral reviews. The experimental result showed that the correct classifier is the svm with reasonable precision equal to 67%.

Key words : *Hirak ; Sentiment analysis ; Natural language processing ; Classification algorithms ; Manual annotation ; Machine Learning ; Algerian dialect*.

الملخص

منذ 22 فبراير 2019، نزل ملايين الجزائريين إلى شوارع جميع المدن الكبرى في البلاد للتعبير عن رفضهم لولاية خامسة لعبد العزيز بوتفليقة. حيث تم بث هذه الحركة الاجتماعية (المسماة الحراك) عبر وسائل التواصل الاجتماعي المختلفة مثل تويتر، حيث عبر مستخدمي الإنترنت عن آرائهم ومشاعرهم التي اختلفت بين الإيجابية والسلبية والحيادية. كان الهدف من هذا المشروع هو تحليل المشاعر والتغريدات المرتبطة بالحركة الجزائرية من خلال تطبيق خوارزميات التصنيف مثل تصنيف بايزي ساذج، الة المتجهات الداعمة، شجرة القرار وخوارزمية الانحدار اللوجستي. بواسطة مختلف طرق لاستخراج السمات والتي هي عبارة عن "حقيبة الكلمات" و" تردد المصطلح- تردد المستند العكس". "في قاعدة بيانات تحتوي على 10000 تغريدة مقسمة إلى 5846 تعليقًا إيجابيًا ،1741 تعليقًا سلبيًا و2184 تعليقا محايدا. حيث أظهرت النتائج التجريبية أن المصنف الصحيح هو الة المتجهات الداعمة وبدقة معقولة تساوي 67%.

الكلمات الرئيسية: *الحراك؛ تحليل المشاعر؛ معالجة اللغة الطبيعية؛ خوارزميات التصنيف؛ وسم يدوي؛ تعلم الالة؛ اللهجة الجزائرية.*