كلية العلوم و التقنيات فاس
ⵜⴰⵖⵙⵓⵍⵜ ⵏ ⵜⵇⴰⵏⵉⵜⵉⵏ ⴷ ⵜⵙⵙⵜⵉⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵜ ⵙⵉⴷⵉ ⵎⵓⵃⴰⵎⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ
Université Sidi Mohamed Ben Abdellah

DÉPARTEMENT D'INFORMATIQUE

PROJET DE FIN D'ÉTUDES

MASTER SCIENCES ET TECHNIQUES
SYSTÈMES INTELLIGENTS & RÉSEAUX

ANALYSE DES SENTIMENTS DES TWEETS MAROCAINS PAR EXTRACTION DE TEXTE

LIEU DU STAGE : Laboratoire Systèmes Intelligents et Applications

L.S.I.A de la Faculté des Sciences et Techniques

Réalisé par :

- GAROUANI Moncef
- CHRITA Hanae

Encadré par :

- Pr. KHARROUBI Jamal

Soutenu le 18.06.2019 devant le jury composé de :

- Pr. F. Mrabti	Faculté des Sciences et Techniques de Fès	(Président)
- Pr. A. Zarghili	Faculté des Sciences et Techniques de Fès	(Examinateur)
- Pr. A. Benabbou	Faculté des Sciences et Techniques de Fès	(Examinateur)
-Pr. J. Kharroubi	Faculté des Sciences et Techniques de Fès	(Encadrant)

Année Universitaire 2018 – 2019

Dédicaces

Je dédie ce travail :

A ma chère mère

Qui m'a soutenu et encouragé durant ces années d'études. Sa prière et sa bénédiction m'ont été d'un grand secours pour mener à bien mes études. Qu'elle trouve ici le témoignage de ma profonde reconnaissance.

A mon cher père

Qui a été à mes côtés pour me soutenir et m'encourager. Il *a su m'inculquer le sens de la responsabilité, de l'optimisme et de la confiance en soi face aux difficultés de la vie. Ses conseils ont toujours guidé mes pas vers la réussite. Que ce travail traduit ma gratitude et mon affection.*

A mes frères, mes grands-parents et ceux qui ont partagé avec moi tous les moments d'émotion lors de la réalisation de ce travail. Ils m'ont chaleureusement supporté et encouragé tout au long de mon parcours.

A ma famille, mes proches et à ceux qui me donnent de l'amour et de la vivacité.

A tous mes amis Qui m'ont toujours encouragé, et à qui je souhaite plus de succès.

A tous ceux que j'aime.

GAROUANI Moncef

Dédicaces

Je dédie ce travail :

A mon très cher père

Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être. Ce travail est le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation. Merci d'avoir été toujours là pour moi.

A ma très chère mère

Aucune dédicace ne saurait être assez éloquente pour exprimer ce que tu mérites pour tous les sacrifices que tu n'as cessé de me donner depuis ma naissance, durant mon enfance et même à l'âge adulte.

A mes très chers frères et sœurs Leila, Yassine, Saad et Ilhame

Vous étiez toujours ma source d'inspiration. Je ne pourrais d'aucune manière exprimer ma profonde affection et mon immense gratitude pour votre aide et votre générosité extrêmes. Je vous dédie ce travail en témoignage de mon amour et mon attachement.

A mes chers amis

En souvenir des moments merveilleux que nous avons passés et aux liens solides qui nous unissent. Un grand merci pour votre soutien, vos encouragements, votre aide.

A toute ma grande famille, et mes fidèles amis de proche ou de loin.

CHRITA Hanae

Remerciements

Nous remercions tout d'abord Dieu tout puissant de nous avoir donné le courage, la force et la patience d'achever ce bon travail.

Nous tenons à exprimer notre profonde gratitude et nos sincères remerciements à notre encadrant Monsieur KHARROUBI Jamal, pour la confiance qu'il nous a accordé en acceptant d'encadrer ce mémoire avec un grand intérêt et une grande compétence, pour ses disponibilités, ses soutiens, ses conseils, pour la qualité de son suivi durant toute la période de ce travail, et pour les encouragements qui nous ont permis de mener à bien ce travail.

Nous remercions également les membres de jury d'avoir pris le temps d'évaluer notre travail.

Ces remerciements vont tout au corps professoral et administratif de la Faculté des Sciences et techniques, spécialement ceux du département Informatique pour la richesse et la qualité de leur enseignement et qui déploient de grands efforts pour assurer à leurs étudiants une formation actualisée.

Merci à tous ceux qui ont participé de près ou de loin pour la réalisation de ce travail.

Analyse des sentiments des tweets marocains par extraction de texte

Résumé

L'émergence de la technologie Web 2.0 a généré une énorme quantité de données brutes en permettant aux utilisateurs d'Internet de publier leurs opinions, avis et commentaires sur le Web. Le traitement de ces données brutes pour extraire des informations utiles peut être une tâche très difficile. Un exemple d'information importante pouvant être automatiquement extraite des publications et des commentaires de l'utilisateur est son opinion sur différents problèmes, événements, services, produits, etc. Ce problème d'analyse des sentiments a été largement étudié pour la langue anglaise. Les solutions proposées sont largement dominées par l'utilisation de deux approches d'analyse basées sur les techniques de machine Learning et l'approche lexicale.

Ce document aborde les deux approches pour analyser les sentiments des tweets marocains écrit en arabe standard ou dialecte marocain. En raison du manque de ressources (bases de données, dictionnaires du lexique) pour la langue arabe, en particulier le dialecte marocain, ce travail commence par la construction d'un ensemble de données de 13 550 tweets valides sur la base de 36 114 tweets collectés et qui sont étiquetés manuellement comme positifs, négatifs, neutres ou mixtes et classés en Dialecte Marocain ou Arabe Standard. Puis décrit les étapes de la construction du Moroccan senti-lexicon, un dictionnaire de 30 000 mots étiquetés comme positif, négatif ou neutre.

Plusieurs expériences ont été réalisées pour évaluer les méthodes les plus utilisées dans le domaine d'analyse des sentiments tel que : N-grammes, Stemming, suppression des Stopwords. Les résultats expérimentaux montrent que les classifieurs à apprentissage profond (Deep Learning) utilisant le modèle de représentation Word Embedding sans retirer les Stopwords dépassent les classifieurs classiques avec le modèle de pondération TF-IDF. Les résultats de cette étude s'avèrent être supérieurs à ceux obtenus par d'autres travaux comparables.

Mots clés : Extraction de texte, Analyse des sentiments, Détection d'opinion, polarité, DM, AS, Approche Lexique, Approche ML Twitter, Langue arabe.

Sentiment analysis of Moroccan tweets using text mining

Abstract

The emergence of the Web 2.0 technology has generated a huge amount of raw data by enabling Internet users to post their opinions, reviews, comments on the web. Processing this raw data to extract useful information can be a very challenging task. An example of important information that can be automatically retrieved from the user's posts and comments is their opinions on different issues, events, services, products, etc. This problem of Sentiment Analysis has been widely studied on the English language. The proposed solutions are largely dominated by the use of two analysis approaches based on machine learning techniques and the lexical approach.

This paper discusses the two approaches to analyze the sentiments of Moroccan tweets written in Standard Arabic or Moroccan Dialect. Due to the lack of resources (databases, lexicon dictionaries) for the Arabic language, especially the Moroccan Dialect, this work starts with the construction of a dataset of 13,550 valid tweets based on 36,114 collected tweets that are manually tagged as positive, negative, neutral or mixed and classified as Moroccan Dialect or Standard Arabic. Then describes the steps of the construction of the Moroccan Senti-lexicon, a dictionary of 30,000 words labeled as positive, negative or neutral.

Several experiments have been carried out to evaluate the most used methods in the field of sentiment analysis (N-grams, Stemming, Stopwords removal). The experimental results show that deep learning classifiers using the Word Embedding without removing Stopwords outperformed classical classifiers with term frequency–inverse document frequency (TF-IDF) weighting scheme through N-grams feature. The results of this study prove to be superior to those obtained by other comparable works.

Keywords: Text mining, Sentiment Analysis, Opinion Detection, polarity, SA, MD, Lexicon-based, Corpus-based, Twitter, Arabic language.

تحليل مشاعر التغريدات المغربية عن طريق تحليل النصوص

ملخص

يقدم هذا المنشور نظاما لتحليل البيانات المستخرجة من شبكة التواصل الاجتماعي تويتر. في هذه الدراسة قمنا بمعالجة التغريدات المغربية المحررة باللغة العربية وباللهجة المغربية الدارجة. هذا النظام يسمح بتجميع، معالجة وإظهار المشاعر المعبر عنها بتغريدات المستعملين المغاربة على شبكة التواصل الاجتماعي تويتر. تتوفر قاعدة البيانات التي قمنا بتجميعها على 13550 تغريدة ذات معنى تم استخراجها من قاعدة بيانات عامة تقدر ب 36114 تغريدة تم تجميعها عبر شبكة التواصل تويتر والتي تم تصنيفها على أساسين: أحدهما نوع اللغة المستعملة (عربية فصحى أو عربية دارجة) والآخر على أساس قطبيتها (إيجابية – سلبية - محايدة أو مختلطة.

لقد أجرينا العديد من التجارب لتقييم استخدام نماذج التمثيل المختلفة والتقنيات والسيناريوهات من حيث طريقة التقطيع، تجديع الكلمات وتوحيد الإملاء. النتائج التجريبية تقدم أحسن السيناريوهات لكل تصنيف وتبين الأساس التي تم عليه هذا التصنيف. (التجديع – التقطيع – نماذج التمثيل) وتجدر الإشارة إلى أن النتائج المحصلة تفوقت على نتائج أعمال مماثلة لما قمنا به نحن.

الكلمات المفتاحية : استخراج النص، تحليل المشاعر، اكتشاف الرأي، تويتر ، **AS، DM، TF-IDF.**

Table des Matières

Liste des Tableaux

Liste des Figures

Liste des Acronymes et Abréviations

AS : Arabe Standard

API : Application Programming Interface

BOW : Bag of Words

CNN : Convolutional Neural Network

DM : Dialecte Marocain

DL : Deep Learning

IDE : Application Programming Interface

LSIA : Laboratoire Systèmes Intelligents et Applications

LSTM : Long Short-Term Memory

LR : Logistic Regression

ML: Machine learning

NB : Naïve Bayes

RNN : Recurrent Neural Network

SO: Semantic Orientation

SVM : Support Vector Machine

TALN : Le traitement automatique du langage naturel

TF-IDF : Term Frequency-Inverse Document Frequency

SW : Stop Words

INTRODUCTION GENERALE

Avec l'avènement du Web 2.0, nous nous sommes trouvés face à plusieurs plates-formes qui permettent aux utilisateurs d'exprimer leurs sentiments et leurs opinions sur un sujet particulier (politique, commercial ou individuel). Facebook, Twitter, Instagram, LinkedIn, ces plateformes sociales font désormais partie du quotidien. L'aspect des données de ces réseaux sociaux, telles que les messages Twitter, génèrent une riche mine de données sur les personnes impliquées dans la communication.

Ces données jouent un rôle important dans la prise de décision pour de nombreuses personnes et organisations. Par exemple, il est important qu'un gouvernement connaisse les points de vue et les préoccupations de ces citoyens et de refléter leurs humeurs et leurs opinions par rapport à l'actualité. De même, l'exemple des entreprises qui souhaitent connaître les réactions des consommateurs vis-à-vis de leurs produits. Cela pourrait les aider à changer de stratégie pour améliorer la qualité de leurs produits et celle de leurs services. Pour cette raison, il est utile d'analyser le contenu des réseaux sociaux qui ont acquis une grande popularité dans le monde entier.

L'analyse des sentiments (Sentiment Analysis ou Opinion Mining) forme une méthode de traitement automatique du langage naturel (Naturel Langage Processing) qui tente de repérer la présence des sentiments ou d'émotions exprimés dans un texte, ou dans une phrase.

Dans ce domaine, la plupart des travaux de recherche ont été menés dans certaines langues européennes, notamment l'anglais et le français. Néanmoins, les recherches effectuées sur l'analyse du sentiment arabe sont considérées comme très limitées, en particulier le dialecte marocain. Cela peut s'expliquer par deux facteurs majeurs. Tout d'abord, le manque de ressources supplémentaires pour ce type de langues. Deuxièmement, la complexité du traitement de cette langue, à savoir le problème d'interférences syntaxiques entre l'Arabe Standard et le Dialecte Marocain, et de la variété d'arabe dialectal parlées au Maroc selon les régions.

Le manque de ressources est considéré comme un problème grave qui entrave, de façon décisive, le développement d'outils de traitement du langage naturel pour les dialectes arabes en général et en particulier pour le Dialecte Marocain. Cela pourrait être considéré comme la raison principale qui a motivé la création d'un corpus d'opinion pour l'arabe standard et le dialecte marocain dans notre travail.

En effet, notre étude vise à développer un système d'analyse des données extraites du réseau social Twitter. Ce système va permettre de collecter, traiter, classer la polarité des tweets marocains selon la catégorie (positifs, négatifs, neutres, mixtes) et selon la classe (arabe standard ou dialecte marocain) auxquelles ils appartiennent, et de localiser les statistiques des sentiments qui dominent la communication des utilisateurs marocains sur la carte du Maroc.

Dans notre travail, nous menons une étude pour découvrir les performances et l'efficacité des dernières techniques de classement automatique du texte à savoir : machines à support vectoriel (SVM), logistique Regression (LR), réseaux de neurones convolutifs (CNN), mémoire à court et long terme (LSTM), et la combinaison (CNN-LSTM), ainsi que l'évaluation de leur pertinence à l'aide de différentes stratégies de prétraitement telles que Stop Words Removing, Stemming, N-grammes et différents schémas de pondération et représentation (TF-IDF, Word Embedding), ainsi que les techniques de l'approche lexicale, pour l'analyse automatique des tweets écrits en arabe standard ou dialecte marocain.

Ce présent rapport se compose de quatre chapitres dont le premier donne un aperçu sur le laboratoire S.I.A ainsi que la problématique et l'objectif de notre recherche. Le deuxième chapitre introduit les concepts de base utilisés ainsi que les caractéristiques et les particularités des microblogs, et particulièrement Twitter, et présente les difficultés de la fouille d'opinions et d'analyse des sentiments (notamment dans les tweets marocains), puis décrit l'évolution de l'analyse des sentiments avec Twitter. Le chapitre trois décrit la structure générale du système proposé et la présentation des expérimentations et des résultats obtenus. Dans le dernier chapitre nous procédons à la description de l'application ARASENTIPEDIA.

Chapitre 1
Cadre général du projet

1 Introduction

Dans ce chapitre sont présentés le laboratoire d'accueil, la problématique, l'objectif du projet et les solutions proposées.

2 Présentation du laboratoire

Le laboratoire SIA, crée en 2011, est une unité de Recherche du Centre d'Etudes Doctorales en Sciences et Techniques de l'Ingénieur domicilié à la Faculté des Sciences et Techniques de Fès et regroupant des laboratoires de recherche tous accrédités par l'Université Sidi Mohamed Ben Abdellah de Fès, et domiciliés à la Facultés des Sciences et Techniques, l'Ecole Supérieure de Technologie, la Faculté Polydisciplinaire de Taza, l'Ecole Nationale des Sciences Appliquées de Fès et l'ENS de Fès.

Le LSIA est composé de 16 enseignants-chercheurs du département d'Informatique de la FST de Fès et de plusieurs doctorants. Cette imbrication étroite entre enseignement et recherche, est un élément essentiel de la dynamique du laboratoire.

Les thématiques de recherche se situent au cœur des Sciences et Technologies de l'Information et de la Communication et s'articulent essentiellement autour des thématiques de recherche des enseignants chercheurs du laboratoire et assure une large couverture thématique présentant un atout très important pour le laboratoire.

3 Problématique

Malgré que l'arabe fasse partie des 10 langues les plus couramment utilisées sur Internet (d'après la classification établie par Internet World State en 2018[1]), et est parlé par des centaines de millions de personnes, les recherches effectuées sur l'analyse du sentiment en langue arabe sont très limitées, en particulier le dialecte marocain par rapport à d'autres langues comme l'anglais. Cela peut s'expliquer par deux facteurs :

Premièrement, le manque de ressources pour ce type de langues. Pour l'analyse du sentiment marocain il n'existe aucun corpus annotés (étiquetés).

Deuxièmement, la complexité du traitement de cette langue : Le problème majeur auquel nous nous sommes confrontés lors du traitement des données Twitter du Maroc c'est que d'une part, le dialecte marocain n'est associé à aucune forme d'écriture normalisée et contient du bruit, des fautes d'orthographe, des abréviations, des répétitions, et des mots qui ne suivent aucune règle grammaticale. Néanmoins, grâce aux nouveaux moyens de communication, les marocains, qui disposent surtout de claviers latins, utilisent l'alphabet latin associé à des chiffres qui ressemblent à des lettres arabes pour s'exprimer.

[1] https://www.internetworldstats.com/

D'autre part darija, est une langue-toit rassemblant plusieurs variétés d'arabe dialectal parlées au Maroc [1]. L'arabe marocain a de nombreux dialectes et accents régionaux, ce qui rend la tâche d'analyse des tweets très difficile.

Le dernier problème dans l'analyse des sentiments c'est le problème du classement de la polarité (positive, négative, neutre ou mixte) à partir de données textuelles à l'échelle Web qui est une tâche très difficile et coûteuse en raison de la grande quantité de données bruitées.

4 Objectif du projet et solutions proposées

Dans ce travail de recherche, notre objectif consiste à étudier et analyser les sentiments dans les tweets marocains en utilisant les techniques les plus récentes pour le classement automatique du texte : les machines à support de vecteurs (SVM), Les réseaux de neurones convolutifs (CNN), les réseaux de neurones récurrents (RNN), Logistique Regression (LR) et les réseaux récurrents à mémoire court et long terme (LSTM) et par les techniques de l'approche lexicale. Pour cela nous étions amenés à construire notre propre corpus des tweets marocain afin de déterminer la polarité des sentiments exprimé là-dedans.

En outre, nous avons développé une application web, qui sera utilisée après, pour construire de nouveaux corpus et de les analyser. Pour cela, nous avons ajouté les modèles d'apprentissage mentionnés comme outils disponibles dans cette application.

5 Conclusion

Dans ce premier chapitre, nous avons donné une présentation du Laboratoire SIA. Ensuite, nous avons défini la problématique, la solution proposée et les objectifs du projet.

Chapitre 2
Analyse des sentiments et Revue de la littérature

1 Généralités

Dans cette section, nous définissons quelques concepts de base importants utilisés dans ce qui suit. Nous présentons aussi les difficultés de la fouille d'opinions et d'analyse des sentiments en générale et dans les tweets marocains en particulier. Ainsi qu'une revue de l'état d'art de l'analyse des sentiments avec twitter et les différentes tâches effectuées par les chercheurs dans ce domaine.

1.1 Réseaux sociaux

Pendant les dernières années, l'internet a connu encore un plus grand élan grâce au développement des médias sociaux. Basés sur des techniques de communication faciles et accessibles pour tous, ces médias favorisent les interactions sociales à travers internet. Les médias sociaux se distinguent des médias traditionnels tels que les journaux, la télévision et la radio par leur utilisation qui est peu coûteuse et libre, de façon à permettre à tout le monde d'accéder à l'information ou de la publier. Ils répondent aux besoins des individus d'exprimer leurs opinions, de demander des conseils et de communiquer de façon rapide et facile. Les possibilités sont nombreuses : envoyer des messages et des photos, réaliser des vidéo-conférences, télécharger des documents, etc… Offrant un accès libre et gratuit, les médias sociaux ont, considérablement, favorisé la communication de masse et ils ont relancé le débat public sur Internet.

Les médias sociaux qui ont récemment pu bénéficier d'un considérable essor sont les réseaux sociaux tels que Facebook, LinkedIn et Twitter. Ce sont des sites web qui rassemblent des identités sociales telles que des individus, des entreprises et des organisations qui peuvent échanger de l'information à travers des interactions sociales. Grâce à leur caractère maniable et leur accès libre, les réseaux sociaux ont connu un succès croissant auprès du grand public.

Les réseaux sociaux font désormais partie intégrante du quotidien des gens. Ils peuvent être utilisés à diverses fins, notamment la publicité, la diffusion d'opinions politiques, l'obtention des commentaires des utilisateurs sur les produits et la diffusion d'informations. Ils créent des liens virtuels entre les utilisateurs, dans lesquels les personnes s'expriment et développent des relations à travers des publications, des commentaires, des messages, etc... Les médias sociaux permettent aux gens de partager leurs pensées, leurs sentiments et leurs opinions avec d'autres personnes instantanément et facilement.

La popularité des nouveaux médias est d'autant plus grande, que la demande d'informations est devenue plus importante dans notre société. En général, les gens aiment consulter les avis des autres avant de passer à l'action ou de se faire une opinion.

1.2 Le Traitement automatique des langues naturelles

Dans la littérature, Le traitement automatique du langage naturel (TALN), ou traitement automatique de la langue naturelle, ou encore traitement automatique des langues (TAL) sont utilisés indifféremment.

Pierrette bouillon [2] définit le TALN comme suit : TALN a pour objet la création de programmes informatiques capables de traiter automatiquement les langues naturelles.

Ela Kumar [3], par contre définit le TALN selon une vision de l'intelligence artificielle et programmation : Le TALN est un domaine significatif de l'intelligence artificielle parce qu'un ordinateur serait considéré comme intelligent s'il peut comprendre la commande donnée en langage naturel au lieu de C, Fortran ou Pascal. Par conséquent et avec la capacité d'ordinateur à comprendre le langage naturel, il devient beaucoup plus facile de communiquer avec les ordinateurs. Par ailleurs le TALN peut être appliqué comme outil de productivité dans des applications allant du résumé des informations jusqu'à la traduction d'une langue à une autre.

Une définition plus technique est donnée par JeanVeronis [4]. On regroupe sous le vocable TALN l'ensemble des recherches et développements visant à modéliser et à reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication.

2 Analyse des Sentiments et Domaines d'Applications

2.1 Définition de l'analyse des sentiments

Dans la littérature, Sentiment Analysis, Opinion Mining, Opinion Extraction, Sentiment Mining, Subjectivity Analysis, Affect Analysis, Emotion Analysis, Review Mining, Appraisal extraction, sont des termes utilisés pour désigner des technologies d'analyse automatique des discours, écrits ou parlés, afin d'en extraire des informations subjectives comme des jugements, des évaluations ou des émotions.

L'origine de la discipline d'analyse des sentiments se réfère aux sciences de la psychologie, la sociologie et l'anthropologie [5]. Le terme Analyse Sentimentale se réfère à l'extraction automatique de texte évaluative qui aide à produire des résultats prédictifs.

Bing Liu [6] a présenté une définition d'analyse des sentiments comportant les domaines d'application ainsi que sa relation avec le TALN : l'analyse des sentiments est le domaine d'étude qui analyse les opinions, les sentiments, les évaluations, les attitudes et les émotions des gens vers des entités telles que des produits, des services, des organisations, des particuliers, des problèmes, des événements, des sujets, et leurs attributs.

Il représente un grand espace de recherche. L'analyse des sentiments est un domaine de recherche extrêmement actif en traitement automatique des langues.

En effet, avant l'apparition du Web et d'Internet, les gens avaient intérêt à connaître les opinions de leurs amis ou de leur famille. Il leur était demandé de faire savoir quel parti politique recevrait leur voix lors des prochaines élections. Grâce à l'essor considérable qu'ont connu le Web et l'Internet à partir des années quatre-vingt-dix, il est devenu possible pour tous de consulter l'opinion d'un vaste groupe de personnes à travers le Web. Donc l'échange d'opinion est la phase principale qui permet d'effectuer une analyse de sentiment sur un sujet donné [7].

La révolution de l'information et précisément l'explosion des plates-formes Web 2.0 telles que les forums de discussion, les blogs et les réseaux sociaux a permis aux utilisateurs de partager des idées et des opinions, d'exprimer leurs sentiments et bien plus encore. Cette révolution entraîne l'accumulation d'une énorme quantité de données pouvant contenir de nombreuses informations précieuses.

Tous les développements récents dans le domaine d'échange d'informations et d'opinions ont motivé la réalisation des applications informatiques conçues pour l'analyse et la détection des sentiments exprimés sur internet. Présentée dans la littérature sous le nom de « Opinion Mining » ou « Sentiment Analysis », l'analyse des sentiments s'utilise, entre autres, pour l'extraction d'opinions sur des sites web et des réseaux sociaux, l'éclaircissement du comportement des consommateurs, la recommandation de produits et l'explication des résultats des élections. Elle consiste à rechercher des textes évaluatifs sur Internet tels que des critiques et des recommandations à analyser de façon automatique ou manuelle et les sentiments qui y sont exprimés afin de mieux comprendre l'opinion publique.

Il a déjà été démontré par des études antérieures que l'analyse des sentiments s'avère particulièrement intéressante pour ceux qui ont intérêt à connaître l'opinion publique, que ce soit pour des raisons personnelles, commerciales ou politiques. Ainsi, de nombreux systèmes autonomes ont, déjà, été développés pour l'analyse automatique des sentiments. Généralement, ces systèmes étaient entraînés aux textes évaluatifs traditionnels tels que les comptes rendus cinématographiques ou les critiques d'un livre. Toutefois, depuis quelques années se sont graduellement ajoutés à ces textes traditionnels les textes non traditionnels tels que les messages envoyés via les réseaux sociaux. Ceux-ci constituent une source précieuse d'opinions échangées parmi les multiples internautes.

Par conséquent, il est important de concevoir des systèmes automatiques capable de rechercher et d'analyser les sentiments qui sont exprimés sur les réseaux sociaux [2].

A cet effet, une grande partie de cette étude sera consacrée à l'analyse automatique des sentiments exprimés dans des tweets et à l'examen des possibilités qu'offre une telle analyse.

2.2 Approches de l'analyse des sentiments et la détection des opinions

2.2.1 Approche basée sur Machine Learning

- *Apprentissage supervisé*

Il est basé sur les données libellées et par conséquent, les étiquettes sont fournies au modèle au cours du processus d'apprentissage. Ces données libellées sont utilisées par l'algorithme d'apprentissage pour donner un modèle qui sera utilisé lors de la prise de décision. Certains modèles d'apprentissage automatique ont été formulés pour classer le texte. Les techniques d'apprentissage automatique comme Naïve Bayes (NB), l'entropie maximale (ME), les machines à vecteurs de support (SVM), CNN, LSTM et LR, ont donné un grand succès à l'analyse des sentiments.

- *Apprentissage non supervisé*

La classification, ou clustering, est un thème de recherche majeur en apprentissage automatique, en analyse et en fouille de données où l'objectif est, la répartition en classes d'un ensemble de données non étiquetées.

2.2.2 Approche lexicale

Méthode basée sur le lexique, elle utilise un dictionnaire des sentiments avec des mots d'opinion et les faire correspondre avec les données pour déterminer la polarité. Elle attribue les scores de sentiment aux mots d'opinion décrivant si les mots sont positifs, négatif ou neutre.

Les approches fondées sur le lexique reposent principalement sur un lexique de sentiment, à savoir, une collection de termes de sentiment connue et précompilée, des phrases et même des expressions idiomatiques, développés pour les genres traditionnels de communication, tels que le lexique Opinion Finder.

2.3 Domaines d'applications de l'analyse des sentiments

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

2.3.1 Le e-commerce

Le marketing a rapidement compris l'intérêt de l'analyse de sentiment : Des agences spécialisées traquent les moindres mots pointant l'image des entreprises et la qualité de leurs produits ou services et les vendent à celles-ci [7].

A travers l'analyse des sentiments, les entreprises peuvent connaître l'opinion des clients sur leurs produits ou leurs services dans une perspective d'améliorer leurs produits et d'augmenter leurs chiffres d'affaire [8].

L'analyse des sentiments fait partie également de la vie des internautes. Les sondages dans ce domaine montrent que la majorité des clients se réfèrent aux forums d'achat sur internet pour avoir le maximum d'informations sur un produit avant son acquisition [9].

2.3.2 La politique

Les acteurs politiques ont suivi la tendance de détection d'opinion. Avant de promulguer une nouvelle loi, les politiciens essayent de récolter l'avis des internautes sur cette loi. Il est intéressant de connaître aussi l'avis des internautes sur un homme politique pour une élection présidentielle [10].

3 Sources des données

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

3.1 Sites d'avis

Les opinions ont le rôle de décideur pour tout utilisateur durant la phase d'achat. Les avis générés par les utilisateurs sur les produits et les services sont largement disponibles sur internet.

Le classement des sentiments utilise les données de l'examinateur collectées à partir des sites Web tels que :

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

Ces sites accueillent des millions d'avis sur les produits par les consommateurs [11] [12].

3.2 Blogs

Un blog est un endroit où les personnes peuvent s'exprimer et partager avec d'autres personnes leurs avis et leurs expériences sur le même site. La simplicité de la création des postes blogs ainsi que leurs formes libres a donné accès libre à tout le monde. La blogosphère est un nom associé à l'univers de tous les blogs. Sur la blogosphère, nous trouvons un nombre important de messages relatif à une panoplie de sujets d'intérêt. Les blogs sont utilisés comme sources d'opinions dans la plupart des études relatives à l'analyse des sentiments [11] [13].

3.3 Micro-blogs

Les micro-blogs sont parmi les outils de communication les plus populaires des d'internautes. Chaque jour, des millions de messages publiés dans des sites Web populaires pour les micro-bloggings tels que : Twitter, Tumblr, et Facebook. Les messages Twitter expriment des opinions qui sont utilisées comme source de données pour classifier le sentiment [11] [9].

4 Twitter

En Mars 2006, Twitter a vu le jour par le développeur Jack Dorsey, comme outil de communication, pour rester en contact avec les amis. Twitter est un service Web qui permet aux utilisateurs d'envoyer et de lire un message court [14].

4.1 Twitter et tweet

Twitter est un réseau social et un microblog qui permet aux utilisateurs de publier des messages en temps réel, appelés tweets. Les tweets sont des messages courts, limités à 280 caractères. En raison de la nature de ce service de microblogging (messages rapides et courts), les gens utilisent des acronymes, commettent des erreurs d'orthographe, utilisent des émoticônes et d'autres symboles qui expriment des significations particulières [15].

Twitter est actuellement l'une des plates-formes de micro-blogage les plus populaires. Son premier slogan était : « Que faites-vous ? » néanmoins, son utilisation a pris une autre piste où les utilisateurs s'échangent des avis et des informations, le slogan devient « Quoi de neuf ? ». Plusieurs célébrités utilisent Twitter, on y trouve même des chefs d'Etat.

Selon les derniers chiffres [2] :

- Twitter a plus que 645 millions d'utilisateurs inscrits.
- 58 millions de tweets envoyés chaque jour.

Dans le cadre de l'analyse des sentiments, la taille minime du message (280 caractères) formule l'hypothèse que ce dernier ne renferme, à priori, plus d'une seule idée, ce qui facilite l'identification de l'opinion exprimée. Certains tweets apparaissent comme des messages codés à cause de l'usage des hashtags, abréviations en tout genre, argot, et émoticons.
Les termes à connaître pour bien utiliser Twitter [16] :

- Followers : les personnes qui vous suivent.
- Followings : les personnes que vous suivez.
- Friends : les personnes que vous suivez et qui vous suivent.

[2] http://www.statisticbrain.com/twitter-statistics/

4.2 Caractéristiques d'un tweet

On peut se sentir un peu perdu dans le vocabulaire utilisé dans les tweets, notamment, à cause du langage et des symboles spécifiques à l'utilisation de Twitter. A quoi sert le @ et # ? C'est quoi RT ? Toutes ces abréviations peuvent paraître un peu floues. Dans une perspective de classement, un petit lexique des principaux mots et signes Twitter est présenté [17] [18] :

- **Mention @ :** se présente sous la forme @NomUtilisateur. Il cible un utilisateur de Twitter dans le tweet posté. Exemple : salut à vous de la part de @Hanae et @Moncef.
- **Hashtag # :** se présente sous la forme #mot-clé. Il identifie le mot-clé en question comme important et peut en faire un sujet populaire. Exemple : #gouvernement.
- **RT (ReTweet) :** se présente sous la forme RT Nom_Utilisatuer. Il permet de partager le tweet d'un utilisateur. Exemple : RT Hanae Excellente.
- **URL (Lien) :** se présente sous la forme https :// ou http ://www. Twitter, permet à l'utilisateur de rejoindre les liens dans son tweet. Exemple : https://www.fstf.com.
- **VIA :** s'utilise pour mentionner votre source d'information, dans votre tweet. Exemple : Via YouTube, Via Facebook.

4.3 L'analyse des sentiments avec Twitter

Parmi les réseaux sociaux, Twitter est l'un des plus importants services de microblogging. L'architecture de Twitter fait de la question « Que se passe-t-il ?» La pierre angulaire de l'échange d'informations. Cela a inspiré l'idée d'utiliser les utilisateurs de Twitter en tant que capteurs distribués [3].

Les plates-formes des médias sociaux telles que Twitter sont de plus en plus courantes et fournissent des informations précieuses (générées par les utilisateurs via la publication et le partage de contenus) [4].

Twitter est une plate-forme de communication basée sur le Web, qui permet à ses abonnés de diffuser des messages appelés « tweets » de 280 caractères maximum, leur permettant de partager des pensées, des liens ou des images. Par conséquent, Twitter est une source riche de données pour l'exploration d'opinion et l'analyse de sentiment. La simplicité d'utilisation et les services offerts par la plate-forme Twitter lui permettent d'être largement utilisée dans le monde arabe et en particulier au Maroc. Cette popularité nous donne accès à une mine riche d'informations qui peuvent servir comme base de données à l'analyse des tweets, qui nous fournissent des informations précieuses [5].

Twitter est une plate-forme multimédia qui permet de partager facilement des opinions en utilisant diverses formes de contenu, notamment du texte, des images et des liens contrairement à de nombreuses autres plates-formes de médias sociaux (tel que Instagram). De plus, en fournissant un accès en temps quasi réel aux publications publiques via l'API,

Twitter est une plate-forme appropriée pour l'exploration d'opinion à grande échelle en temps quasi réel.

5 Difficultés de la fouille d'opinions et d'analyse des sentiments

L'analyse de sentiments ou d'opinions consiste à déterminer la polarité d'une telle opinion. Ces sentiments peuvent être positifs, négatives, neutres, ou mixtes. Nous montrons ci-dessous quelques difficultés de cette fouille d'opinions [19] :

- Difficulté due à l'ambiguïté des mots. Par exemple le mot « petit » est un fait dans la phrase suivante «il est petit ». Par contre il exprime une opinion dans « c'est un petit ».
- Difficulté due à la structuration de la phrase. Par exemple on oppose deux parties d'une phrase avec la conjonction « mais », par exemple l'histoire du film est intéressante mais les acteurs étaient mauvais. Dans ce cas la polarité de la deuxième partie est opposée à la première.
- Difficulté due au vocabulaire utilisé pour exprimer une opinion. Il diffère d'une personne à une autre, selon la région, l'âge, le sexe, etc…
- Difficulté due à l'analyse de la phrase par « paquets de mots ». Les deux phrases suivantes contiennent les mêmes paquets de mots sans, pour autant, exprimer les mêmes sentiments. La première phrase contient un sentiment positif alors que la deuxième est négative : « Je l'ai apprécié pas seulement à cause de … », « Je ne l'ai pas apprécié seulement à cause de … » où se présente la gestion de la négation.

6 Difficultés de la fouille d'opinions et d'analyse des Sentiments dans les tweets marocains

Le dialecte marocain est considéré comme la langue informelle du Maroc et utilisé principalement dans la communication quotidienne, les médias et la publicité commerciale.

Alors que l'arabe standard moderne est rarement parlé dans la vie quotidienne et n'est utilisé qu'à des degrés divers dans des situations formelles telles que des sermons religieux, des livres, des journaux, des communications gouvernementales, des émissions de nouvelles ou des débats politiques. La darija est la langue commune parlée au Maroc.

Comme tous les autres dialectes arabes, le dialecte marocain possède son propre :

- *Transcription* (G, V et P que nous transcrivons respectivement avec les lettres, ك et ب, telles que جراج / garage /, سطوب / stop / et فلاج / village /).
- *Phonologie* (D'une part, les voyelles sont souvent omises dans le DM, surtout lorsqu'elles se trouvent à la fin d'une syllabe ouverte, ceci est probablement dû à une interférence avec la langue amazighe. Par exemple, le mot لغتهم / leur langue / est

prononcé dans l'arabe standard par « loratohom », tandis que dans darija, il est prononcé « lorthom ». Par contre, le dialecte marocain est caractérisé par la prononciation des consonnes G, V et P qui n'existent pas dans l'AS).

- *Morphologie* (le préfixe س dans l'AS est converti en غادي ou غا en darija. De plus, la conjugaison des verbes dialectaux marocains est possible à l'aide d'une liste de préfixes et de suffixes avec une légère modification du motif du mot), et vocabulaire (le vocabulaire de base du dialecte marocain est très influencé par le lexique de l'arabe standard. Cependant, le DM est grammaticalement plus simple et a un vocabulaire moins volumineux que l'AS. De plus, il emprunte ses mots dans d'autres langues telles que le français, l'espagnol et le tamazight). Par conséquent, il partage certaines caractéristiques avec l'AS. Cependant, il y a quelques différences dans certaines autres. De plus, le DM peut être divisé en un groupe de sous-dialectes selon les régions.

L'arabe marocain ou darija marocain (الدارجة au Maroc), a été fortement influencé principalement par les langues berbères (Muš or meš : chat (orig. Amouch), xizzu : carottes, šħal : combien), le français (forshita : fourchette, sbitar : hôpital (hôpital)…), l'espagnol (rwida : rueda (roue), kuzina : cocina (cuisine), skwila : escuela (école), simana : semana (week)) et finalement la langue arabe standard.

Mot	Origine	Equivalent en Darija
شربت	AS	شربت
أتمشى	AS	كنتمشى
فهمتني	AS	فهمتني
موش	Amazigh	مش
تمارة	Amazigh	تمارة
Autobus	French	طوبيس
Fromage	French	فرماج
Roueda	Spanish	رويدا
Cucina	Spanish	كوزينة

Tableau 1 : les origines de quelques mots en darija

Il existe des différences lexicales notables entre l'arabe marocain et la plupart des autres langues arabes. Certains mots sont essentiellement propres à la Darija : daba « maintenant ». Beaucoup d'autres, cependant, sont caractéristiques de l'arabe maghrébin dans son ensemble comme HBET » « descendre » du classique habata. D'autres sont partagés avec l'arabe algérien, tels que hdeṛ « parle », du hadhar classique, et temma « là-bas », du thamma classique.

Dans notre travail, nous nous sommes concentrés sur le DM standard parlé dans le centre du Maroc et compris par la majorité des Marocains.

7 Historique sur l'Analyse des Sentiments avec Twitter

Avant de se concentrer sur l'analyse des sentiments, il est important de définir le mot « Sentiment ». Le terme couvre plusieurs acceptions en fonction du domaine dans lequel il est appliqué. Dans la littérature, un sentiment est « un jugement fondé sur une appréciation subjective (et non sur un raisonnement logique) → avis, idée, point de vue ».

Suivant (Pang et Lee, 2008), l'opinion des autres a toujours été une pièce d'information très précieuse au moment de se faire une opinion ou de prendre une décision. En effet, avant l'apparition du Web et l'internet, les gens avaient intérêt à connaître les opinions de leurs amis ou de leur famille. Il leur était demandé de recommander un mécanicien automobile ou de faire savoir quel parti politique recevrait leur voix lors des prochaines élections. Grâce à l'essor considérable qu'ont connu le Web et l'internet à partir des années quatre-vingt-dix, il est devenu possible pour tous de consulter l'opinion d'un vaste groupe de personnes à travers le Web. L'opinion de ce groupe est d'autant plus précieuse que les personnes dedans ne sont ni des membres de la famille, ni des critiques professionnels intéressés à promouvoir un produit. Leur opinion est considérée comme plus objective et par conséquent, plus valable [7].

Il existe une demande croissante d'informations évaluatives qui a fait naître le phénomène des nouveaux médias ou médias sociaux. Les messages SMS, les forums en ligne, les blogs et les réseaux sociaux ne sont que quelques exemples de ces nouveaux médias. Ils connaissent une immense popularité grâce à la vitesse à laquelle les messages sont envoyés et à leur grande accessibilité (Pang et Lee, 2008).

A part l'intérêt que les individus ont à consulter les opinions d'autrui, il existe d'autres personnes et organisations qui veulent connaître les opinions échangées à leur égard sur le Web. En premier lieu, il existe l'intérêt commercial auquel pourrait servir un système autonome d'analyse de sentiments. Pang et Lee (2008) ont donné l'exemple d'un producteur qui aimerait savoir pourquoi son nouveau modèle de portatif (ordinateur portable) ne se vend pas bien. Il lui convient donc de connaître les jugements personnels des consommateurs à l'égard de ce portatif : Est-ce un problème de conception ? Le service client était mauvais ? Ce ne sont que quelques questions dont la réponse est très précieuse pour le producteur afin qu'il puisse adapter ses produits aux impératifs du marché.

L'année 2001 marque le début de la demande croissante de systèmes automatiques d'analyse des sentiments. Cette demande a émané d'une combinaison de développements : l'essor de méthodes d'apprentissage automatique, la disponibilité d'ensembles de données grâce à l'expansion du Web et la naissance de sites Web qui recueillent des critiques. Il existe dans la littérature plusieurs termes qui renvoient à l'analyse des sentiments : opinion mining, sentiment analysis, subjectivity analysis, review mining, appraisal extraction, etc. La différence entre les dénominations est insignifiante, étant donné que toutes font référence

à l'analyse des opinions et des sentiments exprimés dans des textes subjectifs. Toutefois, les termes opinion mining et sentiment analysis s'avèrent les plus univoques et s'utilisent par conséquent très fréquemment (Pang et Lee, 2008).

Pour présenter les recherches sur l'analyse des sentiments avec Twitter nous considérons trois catégories à savoir, Classement des sentiments, Prédiction des résultats, et détection des évènements.

7.1.1 Historique de classement des sentiments

En 2016, une recherche menée de Crannel et Al. Sur l'étude des comportements d'utilisation de Twitter chez des patients atteints d'un cancer a révélé que ces derniers décrivaient et expliquaient ouvertement et franchement leurs sentiments à propos de leur maladie sur Twitter [20].

D'autre part, Cheong et al., ont proposé un modèle d'analyse des sentiments qui fournit des visualisations graphiques utiles sur des scénarios de terrorisme potentiels, basées sur les données de l'opinion publique recueillies sur Twitter [21].

De même, Neppalli et al. Ont enquêté sur l'utilisation de l'analyse de sentiment sur Twitter pour extraire des informations en cas d'urgence. Ils ont analysé les sentiments des utilisateurs de Twitter dans les régions touchées par l'ouragan Sandy. Ils ont également présenté les changements de sentiment basés sur la localisation [22].

Une autre étude intéressante, publiée récemment, analyse les sentiments exprimés dans la messagerie de chat pour les jeux vidéo afin de mieux comprendre les utilisateurs (Thompson et al, 2017).

En 2018 ils ont fait une enquête sur les sentiments et l'opinion publique vis-à-vis de la crise des réfugiés syriens. Pour analyser les opinions du public sur ce sujet sur Twitter, ils ont collecté un total de 2 381 297 tweets pertinents en deux langues, dont le turc et l'anglais [23].

Sur la base de 17.573 tweets étiquetés avec quatre étiquettes pour le sentiment : positif, négatif, neutre et mixte, Al-Twairesh, Al-Khalifa et Al-Salman ont proposé un système d'analyse des sentiments concernant les tweets saoudiens écrites en arabe standard moderne et en dialecte saoudien. Leur modèle atteint une précision de 62% dans la prédiction des sentiments des tweets [24].

Shoukry et Rafea [25] ont collecté 1 000 tweets sur plusieurs sujets d'actualité. Ils ont construit deux listes : la première comprenait des mots de sentiments négatifs, tandis que la seconde contenait des sentiments positifs. Ils ont appliqué les caractéristiques unigrammes et bigrammes et extrait tous les mots vides arabes dans la base de données d'apprentissage. Dans leur expérience, la technique SVM a été appliqué à l'aide du logiciel 'Weka Suite' pour le processus du classement dans les deux approches utilisées (ML et SO). La précision du

classement par l'approche machine Learning était de 78,8% et celle de l'approche lexicale 75,9%. Cependant, leur ensemble de données était relativement limité.

Le travail rapporté par El-Halees a évalué la précision de l'analyse du sentiment arabe lors de l'utilisation d'approches basées sur le lexique et d'apprentissage automatique. Pour l'approche basée sur le lexique, il a construit un dictionnaire manuellement de mots subjectifs arabes, et dans l'apprentissage automatique, il a utilisé l'entropie maximale, les k-voisins les plus proches (KNN), Naïve Bayes et les algorithmes de machine à vecteurs de support (SVM). Ces classificateurs ont atteint une précision maximale de 63% avec une marge d'erreur de 57%, la meilleure précision obtenue étant avec KNN, à environ 63,58%.

Dans une étude récente [1], une équipe marocaine a travaillé sur une classification non supervisée (K-means) des tweets marocains en fonction du sentiment qui leur est exprimé : positif ou négatif. Ils ont collecté un total de 500 tweets. Ils ont construit un dictionnaire qui contient la transformation des mots écrit en dialecte marocaine ou en Berber en Arabe standard. En découvrant les sujets liés à chaque catégorie, puis de localiser sur une carte marocaine les zones d'où proviennent les tweets liés à ces sujets.

La même équipe a réalisé une autre étude [26] sur un classement supervisé (naïve bayes) de 700 tweets en positive ou négative en se basant sur les emojis, tout en découvrant les sujets liés à chaque catégorie.

Dans une autre recherche en dialecte jordanien [27], ils ont proposé un corpus étiqueté se compose de 1800 tweets (900 positives et 900 négatives) pour comparer les performances du SVM et naïve bayes en utilisant différentes méthodes de prétraitement.

7.1.2 Historique de prédiction des résultats

En 2012, une analyse des séries chronologiques est appliquée au sondage d'opinion publique politique aux messages Twitter qui ont mentionné le président Barack Obama. Les auteurs employaient le logiciel qui a mesuré le sentiment dans les messages de Twitter, pour comparer le sentiment public d'Obama aux sondages d'opinion publique collectées traditionnellement. Les auteurs ont conclu que Twitter est une mesure fiable de l'opinion publique [28].

En 2015, Wu et Shen ont proposé un modèle d'analyse des sentiments pour prédire la popularité des nouvelles sur Twitter (Wu et Shen, 2015). Ils ont étudié les caractéristiques de la propagation de l'information sur Twitter et ont découvert qu'il existait une corrélation entre la popularité de l'information et la fréquence d'interaction des retweeters avec la source d'information [29].

7.1.3 Historique de détection des évènements

Twitter constitue un excellent moyen pour diffuser des informations, pour discuter des évènements et pour donner des avis. A partir du message publié sur Twitter on peut détecter un évènement.

En l'année 2011 Weng et Lee [29] s'intéressent à la détection d'évènement sur Twitter en analysant le contenu des tweets publiés dans la plateforme. [19] ont introduit une structure nommée EDCoW (Event Detection with Clustering of Wavelet-based Signals). Dans EDCoW, le signal de chaque mot est calculé en appliquant l'analyse en ondelettes sur la fréquence des signaux bruts des mots. En considérant l'autocorrélation des signaux correspondants, les mots sans importance sont supprimés. Les mots restants sont ensuite regroupés pour construire des évènements avec une technique graphique. Sur la base de leur expérimentation, les auteurs affirment que EDCoW atteint une bonne performance dans l'étude.

Sakaki, Okazaki, et Matsuo [30] en 2010 ont essayé de détecter les tremblements de terre de l'information générée par les capteurs sociaux représentés par les utilisateurs de twitter. En utilisant le modèle à la fois temporelle et géo-spatiale, les auteurs ont démontré que les tweets pourraient être utilisé pour prédire la ligne des tremblements de terre qui vont se produire quelques instants après un tremblement dans une région spécifique. De même, les auteurs montrent qu'il est possible de prédire la trajectoire des ouragans en utilisant les tweets générés par la région affectée.

En 2012 et 2013, ils ont collecté 1,8 million de tweets sur « le changement climatique » et « le réchauffement de la planète » dans cinq langues principales (anglais, allemand, russe, portugais et espagnol). Ils discutent la géographie des tweets, des habitudes hebdomadaires et quotidiennes, des événements importants qui ont affecté le « tweeting » sur le changement climatique. Ils prévoient que l'exploitation de réseaux sociaux deviendra une source majeure de données dans le discours public sur le changement climatique.[31]

Un autre travail, ou ils présentent un système appelé PoliTwi, conçu pour détecter les sujets politiques émergents (Top Topics) sur Twitter plus tôt que d'autres canaux d'information standard. Les 27 meilleurs sujets reconnus sont partagés par différents canaux avec le grand public. Pour l'analyse, ils ont collecté environ 28 millions de tweets avant et pendant les élections législatives de 2013 en Allemagne, d'avril au 29 septembre 2013 [32].

8 Conclusion

Dans ce chapitre nous avons présenté les caractéristiques et les particularités des microblogs, notamment Twitter, ainsi que les difficultés de la fouille d'opinions et d'analyse des sentiments, et des travaux qui s'intéressaient à l'analyse des tweets. Ceux-ci nous ont apporté des idées pour le traitement des textes avec une taille réduite (tweets). La plupart de ces travaux traitent les tweets écrits en anglais ou en français. Dans notre

travail, nous nous intéressons au texte écrit en arabe marocain. Dans le chapitre suivant, nous décrirons le corpus que nous avons construit et les statistiques que nous avons réalisées, ainsi que le processus du prétraitement effectué par notre système et les résultats de classement.

Chapitre 3
Expérimentations
&
Résultats

1 Introduction

Nous présentons, dans ce chapitre, les corpus que nous avons collectés, les expérimentations que nous avons faites, et les résultats obtenus. Dans nos expérimentations, nous avons traité les tweets qui portent sur le Maroc et qui sont écrits en arabe standard et/ou dialectal.

2 Structure générale du système

Pour collecter les tweets, nous avons implémenté un programme Python qui a utilisé la bibliothèque Tweepy. Cette bibliothèque permet d'accéder aux données (tweets, les informations des utilisateurs, etc.) Twitter via son interface de programmation Twitter API.

Nous pouvons résumer notre approche du classement des sentiments dans les tweets marocains par le schéma suivant :

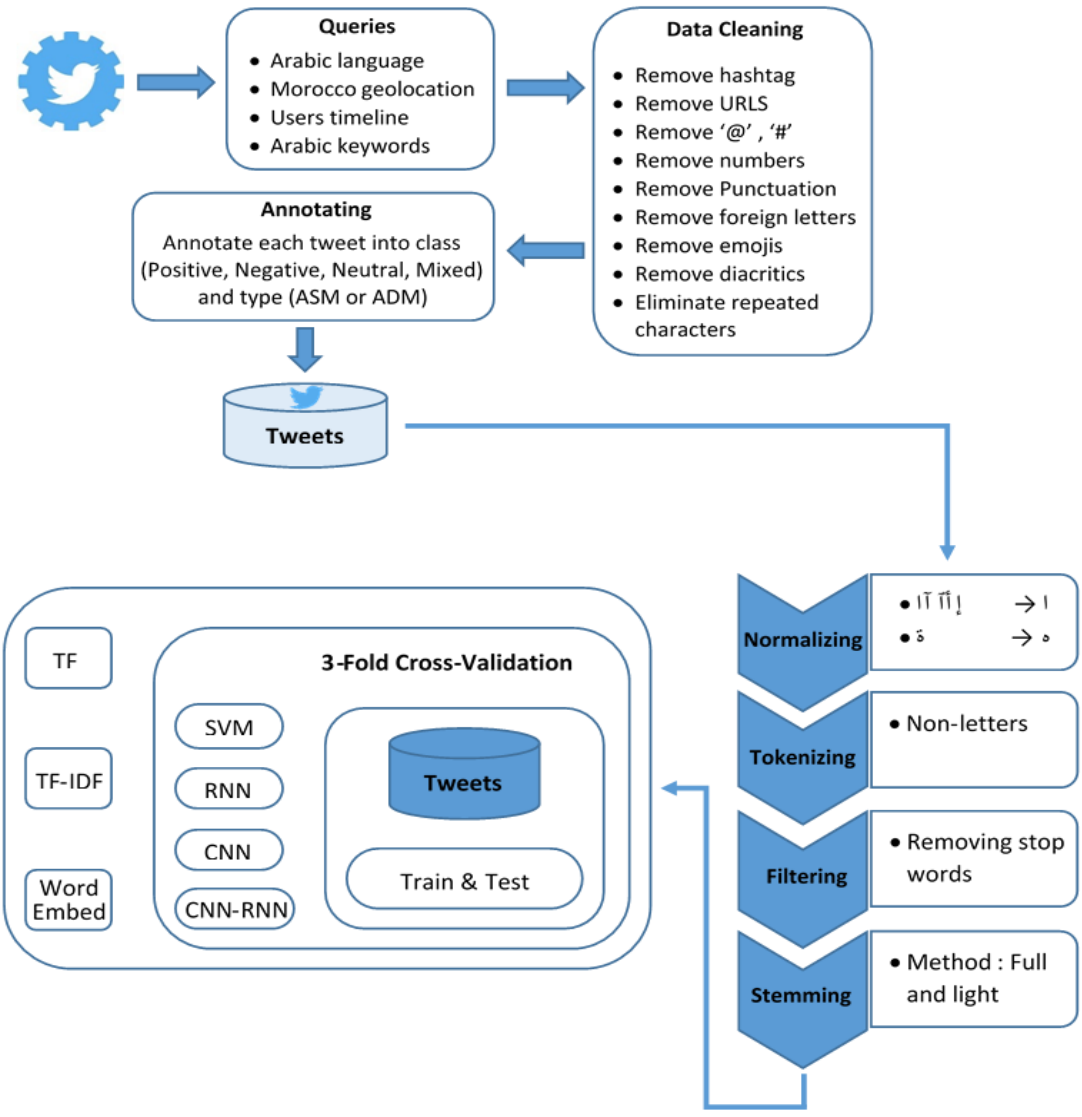

Figure 1: La méthodologie proposée pour l'analyse des sentiments sur les tweets marocains.

3 Extraction des données

3.1 Streaming API

Notre première étape dans ce travail consistait à nous familiariser avec *Twitter API*[1]. Au début, nous avons utilisé la *streaming API*[2] qui permet d'obtenir les tweets en temps réel, où on peut filtrer les tweets avec plusieurs mots-clés. Par exemple : 'الحكومة', 'المغرب', etc. On peut également filtrer les tweets selon leur positionnement géographique. Par exemple, pour récupérer les tweets des utilisateurs marocains, on doit utiliser les coordonnées latitudes/longitudes qui correspondent à 'Morocco'.

Cet API nous permet d'avoir un nombre plus important de tweets distincts, en spécifiant la langue des tweets. La figure 2 montre un extrait de tweet retourné par *streaming API*.

```
01 createdAt=Sun Apr 22 00:38:15 2019, id=193921636305604609,
02 text='نعمان بلعياشي يغني مقطع من أغنيته "جمالي" على بلاطو #بيناتنا
   http://t.co/vWM3t4aM',
03 Source='web'
04 InReplyToScreenName='null', geoLocation=null,
08 retweetedStatus=StatusJSONImpl {createdAt=Sat Apr 21 21:22:36 2019,
   id=193872400662802432,
10 place=null, geoLocation=null, annotations=null,
11 ...
12 geoLocation=null, place=null, retweetCount=0,
13 user=UserJSONImpl {id=259789350, name='hespress', screenName='hespress',
14 Location='Casablanca', description='',
27 ...
28 FriendsCount=291, createdAt=tue Apr 20 09 :39 :59 EST 2019, timeZone='Casablanca'
29 Lang='ar', statusesCount=2196,
```

Figure 2: Extrait d'un tweet retourné par streaming API (format Json).

Ligne 1 : informations sur le tweet (date, identifiant)
Ligne 2 : le texte du tweet
Ligne 14 : informations sur la location du tweet.

3.2 Search API

Par la suite, nous avons utilisé *Search API*[3] qui permet de retourner des tweets qui répondent à une requête q. Si q = 'المغرب', *Search API* retourne les tweets qui contiennent le terme 'المغرب'. On peut également filtrer les résultats selon plusieurs critères tels que :

- **Langue des tweets :** spécifier la langue avec laquelle le tweet est écrit.
- **La période :** trouver les tweets écrits entre une date *since* et une date *until*.
- **Type de résultats :** spécifier le type des tweets retournés les plus populaires (les plus retweetés), les plus récents ou mixtes.

[1] https://developer.twitter.com/en/docs

[2] https://developer.twitter.com/en/docs/tweets/filter-realtime

[3] https://developer.twitter.com/en/docs/tweets/search

La *Search API* a des limites :

- Le nombre de tweets retournés par requête ne peut pas dépasser 1500.
- Il ne peut pas trouver les tweets qui étaient envoyés il y a plus qu'une semaine.

La figure 3 montre un exemple d'un tweet retourné par *Search API*.

```
01 text=' تعددت أشكالها حسب العادات والتقاليد الموجودة بين القبائل
   المغرب #مراكش#',
02 fromUser='AnisKhez', fromUserId=121504252,
03 isoLanguageCode='ar',
04 source='<a
   href="http://twitter.com/#!/download/iphone"rel="nofollow">Twitter for
   iPhone</a>',
05 createdAt=Sat Apr 21 21:25:10 2019,
06 location='null',
07 place=null, geoLocation=null, annotations=null,
10 userMentionEntities=[],
11 hashtagEntities=[HashtagEntityJSONImpl{start=28, end=36, text='المغرب'},
   HashtagEntityJSONImpl{start=37, end=43, text='auMax'}],
12 mediaEntities=null
```

Figure 3: Extrait d'un tweet retourné par Search API (format Json).

Ligne 1 : texte du tweet
Ligne 3 : langue du texte (identifiée par Twitter).

3.3 User_timeline API

Vers la fin, nous avons utilisé le *User_timelineAPI*[4] qui permet d'obtenir une collection des tweets les plus récents publiés par l'utilisateur, indiqués par les paramètres Name ou User_id. Cette méthode ne peut renvoyer que 3 200 des tweets les plus récents d'un utilisateur avec une langue spécifiée. La figure 4 montre un exemple de tweet retourné par *user_timeline API*.

```
01 text=' 😀افضل اوقات الإبداع هو في آخر الليل', id=193873044287143936,
02 fromUser='imad', fromUserId=121504252,
03 isoLanguageCode='ar',
04 followers_count: 465425,
05 profileImageUrl='http://a0.twimg.com/profile_images
   2008493088/229604_10150324090608888_573338887_7473932_392125_n_normal.jp
   g'
06 createdAt=Sat Apr 21 21:25:10 2019,
07 location='null',
   …
13 userMentionEntities=["screen_name": "TwitterDev", "id": 2244994945]
   …
```

Figure 4: Extrait d'un tweet retourné par user_timeline API (format Json).

Ligne 1 : Le texte du tweet.
Ligne 2 : Le nom de l'utilisateur.
Ligne 3 : Le nombre des followers.

[4] https://developer.twitter.com/en/docs/tweets/timelines

4 Statistiques et interprétations

4.1 Expérience 1

Comme mentionné dans la section précédente, nous avons commencé par *Streaming API* pour extraire des tweets. Nous avons sélectionné des tweets écrits en arabe qui répondent à la condition du géolocalisation 'Morocco'.

Nous avons réussi à extraire 6 578 tweets valides à partir de 23 743 tweets collectés entre le 21 janvier 2019 et le 24 février 2019. La figure 5 montre la distribution des tweets collectés selon les dates.

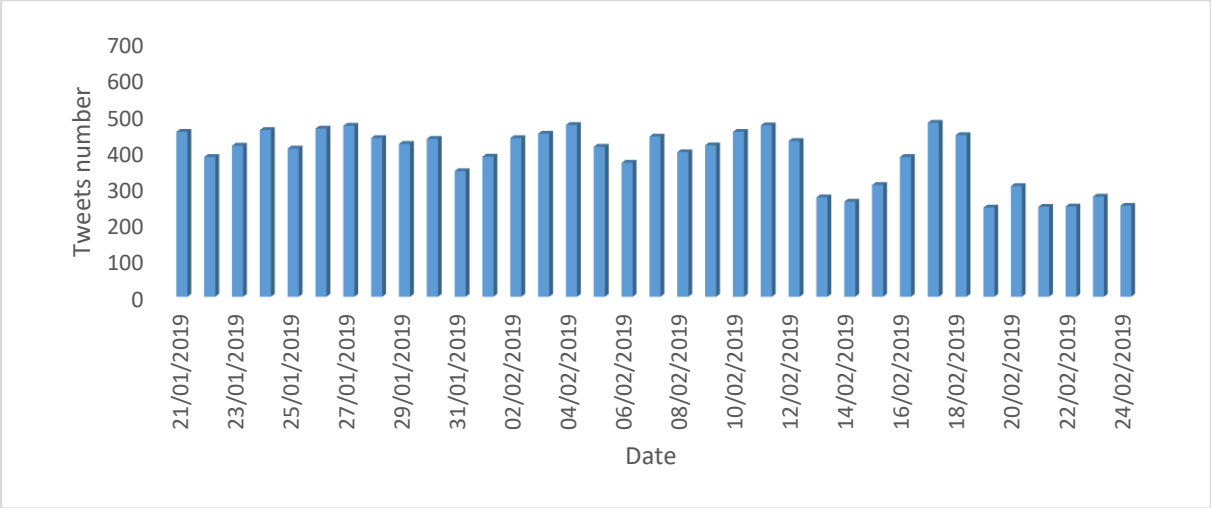

Figure 5: Nombre de tweets par jour (du 21 janvier au 24 février 2019).

Nombre de tweets	5 436
Nombre des retweets	1 142
Nombre d'utilisateurs distincts	2 354
Nombre de tweets qui contiennent au moins un hyperlien	5 867
Nombre de hashtags distincts	1 765

Tableau 2: Statistiques sur les tweets valides publiés entre le 21 janvier et le 24 février 2019.

Etant donné la difficulté de comprendre ce type de texte (courts, fautes d'écriture, convention d'écriture employée par les utilisateurs, etc.), nous avons essayé de nous baser sur d'autres éléments tels que : les hashtags, le comportement des utilisateurs, les relations entre les utilisateurs, etc.

Nous avons constaté que les utilisateurs utilisent souvent de hashtags, cet élément joue un rôle important pour avoir une idée sur les préoccupations des utilisateurs.

4.2 Expérience 2

Cette fois-ci, nous nous sommes intéressés aux tweets qui portent sur des personnes en particulier. Nous avons utilisé SocialBakers[5] pour déterminer les comptes Twitter marocains les plus actifs. Cela nous a donné une liste de 29 noms.

Pour recueillir des tweets, nous avons utilisé le *user_timeline API* sur une base des 29 utilisateurs, nous avons réussi à extraire 4 537 tweets valides à partir de 10 371 tweets collectés.

Pseudonyms	Nombre de tweets collectés	Nombre de retweets collectés	Nombre de tweets valides
Hespress	3 911	1557	946
Fekri_AKKOUH	1 934	1266	776
Klam_56	1 509	161	543
HaMaSa	3 407	3818	690
LeSheriiff	2 403	1497	435

Tableau 3: Informations sur les tops 5 utilisateurs, nombre de tweets collectés.

Nombre de tweets	3 124
Nombre des retweets	1 413
Nombre d'utilisateurs distincts	1 048
Nombre de tweets qui contiennent au moins un hyperlien	3 867
Nombre de hashtags distincts	927

Tableau 4:Statistiques sur les tweets de 29 utilisateurs marocains.

4.3 Expérience 3

Nous avons encore collecté des tweets à l'aide de *Search API*. Nous avons utilisé Trendsmap[6] pour déterminer les hashtags les plus en vogue au Maroc. Nous avons environ 950 balises de hashtags distinctes qui sont à nouveau utilisées pour télécharger les tweets. Nous avons réussi à extraire 2345 tweets valides après avoir filtré les tweets arabes.

4.4 Corpus final

Le corpus se compose des tweets marocains qui ont été collectés entre le 21 janvier et le 24 février 2019. Du total de 13 460 tweets valides à base de 36 114 tweets collectés, 8 750 ont été publiés par des hommes et 4 010 par des femmes. De 700 tweets n'a pas pu être identifié le sexe de l'auteur en raison d'un manque d'informations.

[5] https://www.socialbakers.com/statistics/twitter/profiles/morocco/

[6] https://www.trendsmap.com/local/morocco

Nombre de tweets collectés	36 114
Nombre de tweets valides	13 460
Nombre d'utilisateurs distincts	3 602

Tableau 5: Statistiques sur le corpus final.

Chaque entrée de notre ensemble de données est structurée comme suit :

- **Tweet id** : l'identifiant du tweet.
- **Tweet texte** : il contient le texte du tweet publié par l'utilisateur.
- **Tweet date** : date de publication du tweet.
- **Tweet author_id** : l'identifiant de l'auteur du tweet.
- **Tweet source** : informations sur la géolocalisation du tweet.

5 Prétraitement et Annotation

5.1 Prétraitement

Nous avons déjà abordé dans la Section II.4.2 les caractéristiques des tweets qui se résume en général dans les longueurs limitées et l'utilisation d'un langage informel. Ainsi, l'utilisateur de Twitter utilise des abréviations, des émoticons, et des argots pour exprimer ses opinions et ses sentiments. Par conséquence une étape de prétraitement est indispensable.

Dans ce qui suit nous allons présenter la procédure de prétraitement suivie dans notre travail, dont le but est de nettoyer les tweets et les rendre le plus proche possible d'un langage formel.

Nous avons procédé à un prétraitement qui suit les étapes suivantes :

1- Supprimer les chiffres, les ponctuations, les liens web : il faut les supprimer car ils ont aucun impact sur le classement.

2- Supprimer les émoticônes.

3- Normaliser les espaces : convertir plusieurs caractères d'espace en un seul caractère.

4- Supprimer les identifiants des utilisateurs : @user

5- Supprimer les Hashtags (TAG) : #hashtag.

6- Eliminer les commandes VIA, RT : Twitter possède son propre vocabulaire et fonctions, il y'a les commandes VIA et RT indique que le tweet a été rediffusé par un autre utilisateur, nous les avons éliminés à cause de son influence négligeable sur le classement.

7- Supprimer les voyelles courtes et autres symboles (harakat,الشكل) qui interfèrent avec les manipulations informatiques avec les textes arabes.

8- Eliminer les caractères répétés : nous avons éliminé les répétitions des caractères dans les mots comme (شكراااا : شكرا , جمييييل : جميل) que l'utilisateur l'utilise pour affirmer le sens.

9- Normalisation des caractères arabes : la normalisation est une étape de prétraitement courante dans le traitement de texte arabe. Dans cette étape, les lettres « آ - إ - أ » sont remplacées par « ا ».

Le Tableau suivant donne quelques exemples de tweets avant et après le prétraitement

Tweets avant prétraitement	Tweets après prétraitement
@AnaChawki ❤❤ الله يوصلك على خيير	الله يوصلك على خير
@AdilRaquy 🍾🍡🍭🎂🌹 عيد ميلاد سعيييييد	عيد ميلاد سعيد
😉 !! ويبقى الحال على ماهوhttps://t.co/6aKc	ويبقى الحال على ماهو
👍👰😍 #wac هدف رائع للوداد في شباك الرجاء	هدف رائع للوداد في شباك الرجاء
تَوَقَّعْ الخَير و افْتَح صَبَاحَك بِالتَّفَاؤُل و الأمَل	توقع الخير و افتح صباحك بالتفاؤل و الامل

Tableau 6:Exemple de tweets avant et après le prétraitement.

5.2 Annotation

Etant donnée l'absence de ressources étiquetées pour des messages écrits par les arabophones marocains, tunisiens, algériens, syriens, etc., il est difficile d'appliquer des techniques de traitement de la langue naturelle afin de déterminer la polarité d'un tweet (Positive, Négative, Neutre, Mixte) et le type de la langue (Arabe Standard AS ou Dialecte Marocain DM). Pour cette raison, nous avons décidé de construire notre propre corpus d'apprentissage.

Dans cette tâche, nous avons préparé un corpus qui contient plus que 13 460 tweets publiés entre le 21 janvier et le 24 février 2019 que nous avons étiqueté selon la polarité et la langue des tweets.

L'étiquetage a été effectué à travers une application web[7] que nous avons développé pour réaliser cette tâche. La figure 6 montre une capture d'écran de la page qui nous permet d'étiquetage des tweets.

[7] https://annotation.pythonanywhere.com

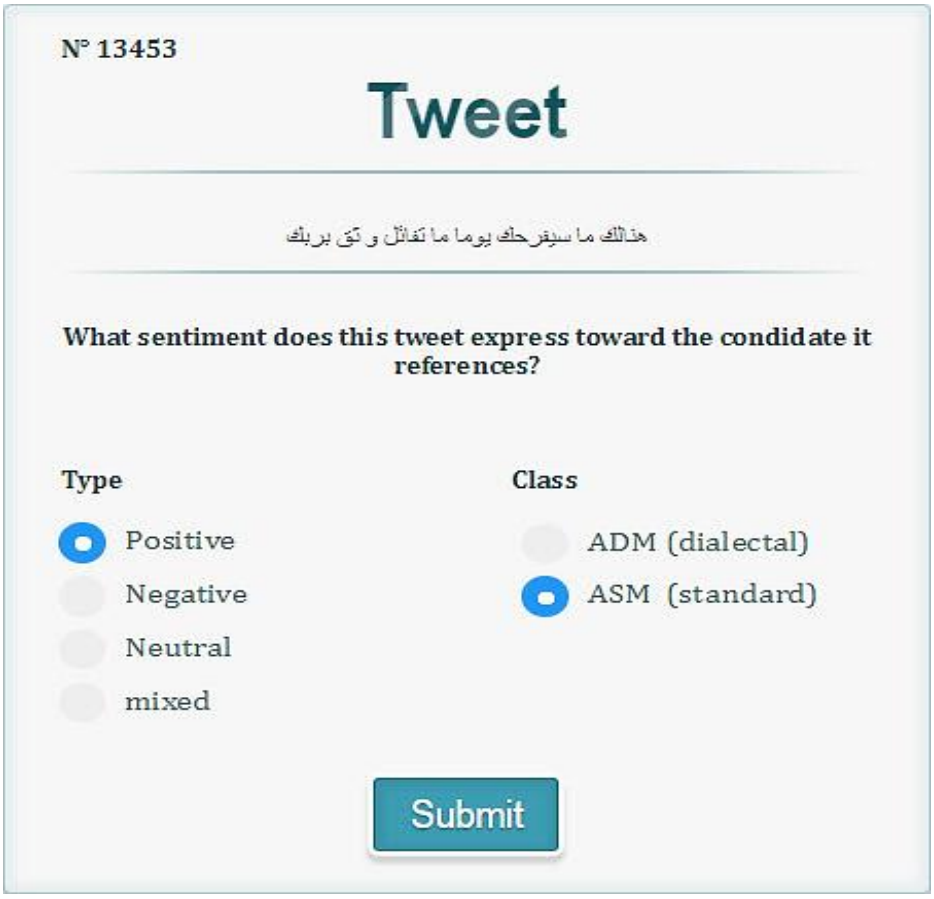

Figure 6: Capture d'écran de la page dédiée pour l'annotation des tweets.

Les tweets ont été étiquetés par deux annotateurs (nous-même). Une première étape de double étiquetage portant sur 1500 messages a été réalisée. Elle nous a permis de se familiariser avec la tâche d'étiquetage. La suite des tweets a été étiquetée en simple étiquetage. Nous avons calculé les accords inter-annotateurs sur les 1 500 messages étiquetés en double (AIA=0.93). Le tableau suivant montre un exemple de tweets étiquetés :

Tweet	Type	Classe
Ar : توقع الخير و افتح صباحك بالتفاؤل و الأمل صباح النور Fr : Attendez-vous aux bonnes choses et commencez votre journée avec optimisme et espoir	Positive	AS
Ar :من المؤسف ان هذا حالنا الذي نعيشه الآن Fr : Malheureusement, c'est notre situation actuelle	Négative	AS
Ar :تابعيني باش نقدر ندخلك Fr : Abonnez moi pour que je puisse vous ajoutez	Neutre	DM
Ar : رغم الصعوبات لي قاتلاني والمشاكل لي كنمر منها كنحاول نضحك ونقول الحمد لله Fr : Malgré les difficultés et les problèmes que j'ai j'essaie de rire et de remercier Dieu	Mixed	DM

Tableau 7: Exemples de tweets étiquetés.

5.2.1 Défis d'annotation du corpus

Nous résumons quelques défis auxquels nous étions confrontés en donnons des exemples.

Supplications : nous avons trouvé des difficultés à déterminer le sentiment des supplications, car ils peuvent contenir des mots positifs ou négatifs, mais il est difficile de savoir s'ils expriment un sentiment.

- Ex : اللهم اجعل لنا في القلب نور و في المال بركة و في الناس محبة و في الدنيا سعادة مساء الخير

  Traduction : Oh mon Dieu, donne-nous de la lumière dans nos cœurs et de la bénédiction en argent et de l'amour dans les gens et dans ce monde le bonheur bonsoir.

  Annotation : Ce tweet a été qualifié positif par les deux annotateurs.

- Ex : يا رب فرج هم كل من كان في ضيق.

  Traduction : O Seigneur, soulage ceux qui sont en difficultés.

  Annotation : Ce tweet a été étiqueté neutre par les deux annotateurs.

Conseil : il s'agit de conseils donnés comme s'il était bon ou mauvais de faire ou de ne pas faire quelque chose.

- Ex : أي شيئ مستور لا تحاول أن تكشفة

  Traduction : Tout ce qui est caché, n'essayez pas de l'exposer.

  Annotation : Ce tweet a été étiqueté neutre.

Citations : elles se présentent sous la forme de citations inspirantes qui véhiculent généralement une signification positive, mais ne constituent pas un avis explicite sur une cible spécifique. Nous suggérons que ce type de texte soit considéré comme neutre car il ne transmet pas de sentiments.

- Ex :على المرأ أن يحاول إعادة نفسه للحياة دون سعادة سواء جاءت أم لم تجئ -جورج إليوت

  Traduction : Que le bonheur vienne ou non, il faut essayer de se préparer à s'en passer.

  Annotation : Ce tweet a été étiqueté neutre.

Détermination du sujet du tweet : il était parfois difficile de déterminer le sujet du tweet et nous ne pouvions donc pas déterminer le sentiment exprimé. Ce défi est en corrélation avec la nature de la langue sur Twitter qui est informelle et courte.

6 Statistiques

L'ensemble de données comporte quatre classes des sentiments, à savoir Positive, Négative, Neutre et Mixte (positive + négative), et deux classes de langues utilisées DM et AS.

La répartition des données selon leurs classes et leurs sentiments est illustrée dans le tableau suivant :

AS	DM	Total
9 640	3 807	13 460

Tableau 8: Statistiques sur le corpus.

Dans nos expérimentations, nous avons divisé le corpus final en deux. Le premier corpus est le corpus AS et le deuxième le corpus DM. Dans ce qui suit, nous présentons des statistiques sur chacun des deux corpus.

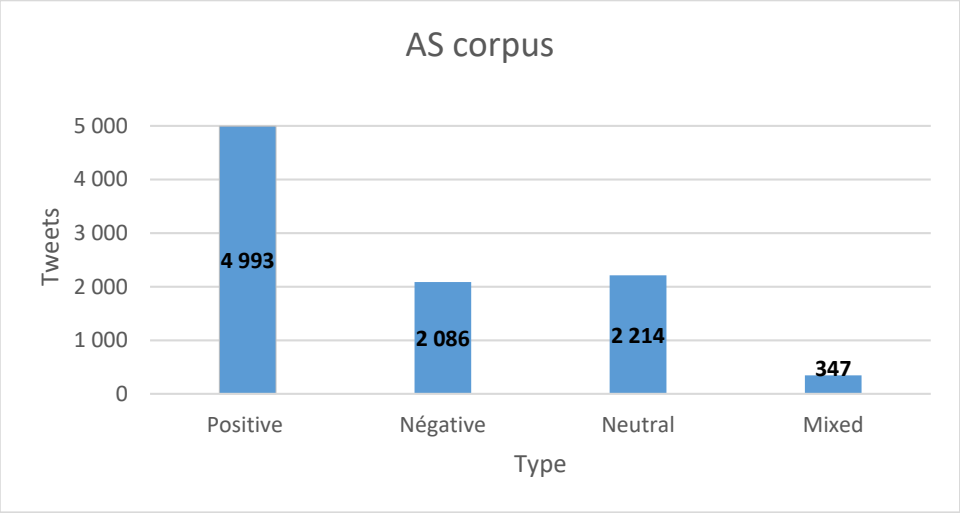

Figure 7: Distribution des sentiments exprimés dans le corpus AS.

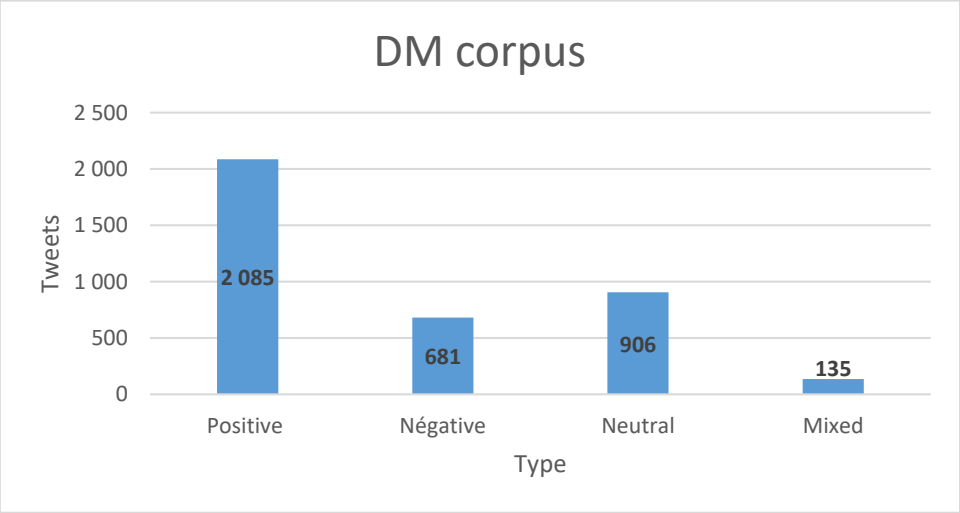

Figure 8: Distribution des sentiments exprimés dans le corpus DM.

Lors de la collection des tweets à partir de l'API Twitter, nous avons extrait les coordonnées (longitude et latitude) de chaque tweet. Nous utilisons ensuite ces coordonnées pour afficher les emplacements des tweets sur notre carte du Maroc.

Figure 9: Emplacement des tweets sur la carte du Maroc.

Cette représentation donne une idée des emplacements des Tweets marocains positifs, négatifs, neutres et mixte, ce qui peut conduire à une meilleure compréhension de notre société marocaine.

6.1 Word cloud

Le nuage de mots-clés (plus rarement nuage de mots-clefs ou nuage de tags, tag cloud, Word Cloud ou keyword cloud en anglais) est une représentation visuelle des mots-clefs (tags) les plus utilisés dans un document ou corpus. Généralement, les mots s'affichent dans des tailles et graisses de caractères d'autant plus visibles qu'ils sont utilisés ou populaires.

Ci-dessous les représentations des nuages de mots des quatre classes.

Figure 10: Positive word cloud.

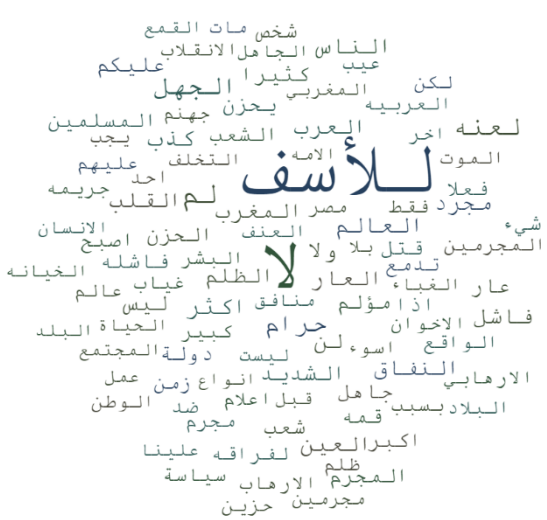

Figure 11: Negative word cloud.

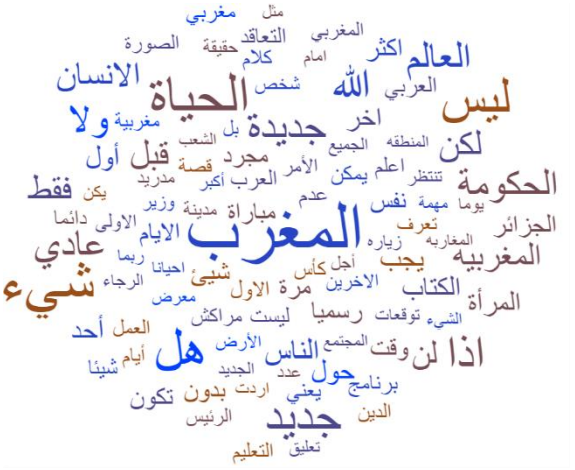

Figure 12: Neutral word cloud.

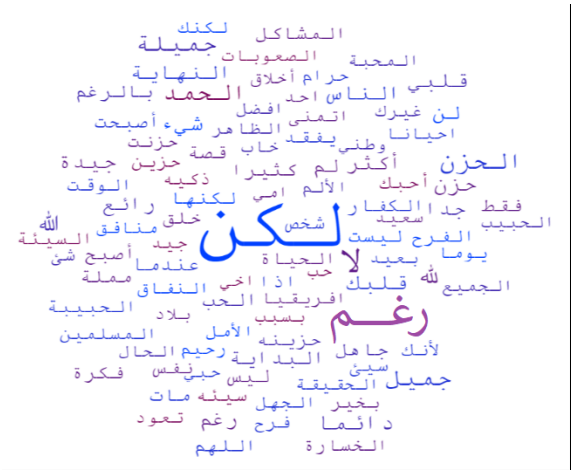

Figure 13: Mixed word cloud.

7 Extraction et présentation des descripteurs

Le processus du classement du texte par des modèles d'apprentissage automatique est essentiellement le même que celui utilisé pour le classement d'un autre type de données. La principale différence, est constituée par le processus de transformation de données pour que celles-ci puissent être passées à l'algorithme de classement comme une représentation vectorielle numérique.

Dans cette transformation, il est nécessaire de passer les données du texte pur à une représentation dans laquelle les documents de texte sont numériquement représentés dans une matrice que le classifieur peut interpréter. On s'est basé sur la description de tâches du processus du classement illustré par la figure 14, on explique ci-dessous les différentes étapes de prétraitement du texte.

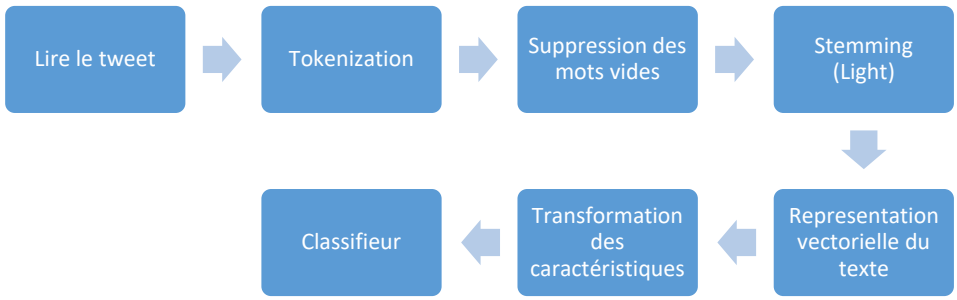

Figure 14: Processus de prétraitement et transformation du texte.

7.1 L'extraction de termes (Tokenization)

Cette tâche consiste essentiellement à diviser le texte qui a été lu en des structures de base pour l'analyse future. Ces structures peuvent être des mots (mono-grammes), des ensembles de deux ou plusieurs mots adjacents (bigrammes ou n-grammes), des phrases ou des déclarations, des symboles ou une autre structure de base offrant une information utile pour le classement. Le résultat est une liste de « tokens », correspondant aux mots, bigrammes, etc., séparés par des caractères d'espace simple.

Les modèles d-grammes présentent l'avantage d'être capable de trouver les dépendances à longue distance et d'apporter une information plus pertinente pour le rattachement syntaxique des mots entre eux. Les modèles d-grammes facilitent aussi le repérage des négations.

7.2 La suppression de mots fonctionnels (stop words)

Il existe certains mots, appelées fonctionnels, qui apparaissent trop fréquemment dans tout type de texte. Cette particularité fait en sorte que leur présence n'apporte aucune information utile pour le classement du texte. La présence de ces mots peut, au contraire, produire du bruit qui affecte les performances du système. C'est la raison pour laquelle il est préférable de supprimer ces mots pour ainsi améliorer la capacité de classement du modèle qui sera postérieurement utilisé. Ce type de mots inclut les connecteurs, les conjonctions, les causes déterminantes, ainsi que des verbes qui figurent fréquemment dans toutes les catégories de classement (par exemple les mots « ... من - مع - إلى »).

7.3 Stemming ou réduction à la tige

En morphologie linguistique, et dans la recherche d'information (Information Retrieval), la réduction à la tige est le processus de diminution de mots déviés à leur tige forme d'origine. La tige n'a pas besoin d'être identique à la racine morphologique du mot. Il est, habituellement, suffisant qu'elle permette de regrouper des mots avec une tige et sens semblable, même si cette tige n'est pas une racine valide.

Dans l'exemple suivant, la création des tokens d'un texte d'entrée et le prétraitement proposé dans notre système. Nous considérerons que la liste de mots vides est celle fournie par défaut pour l'arabe standard [8] et une liste du dialecte marocain que nous avons créé, et qui contient plus de 100 mots fonctionnels.

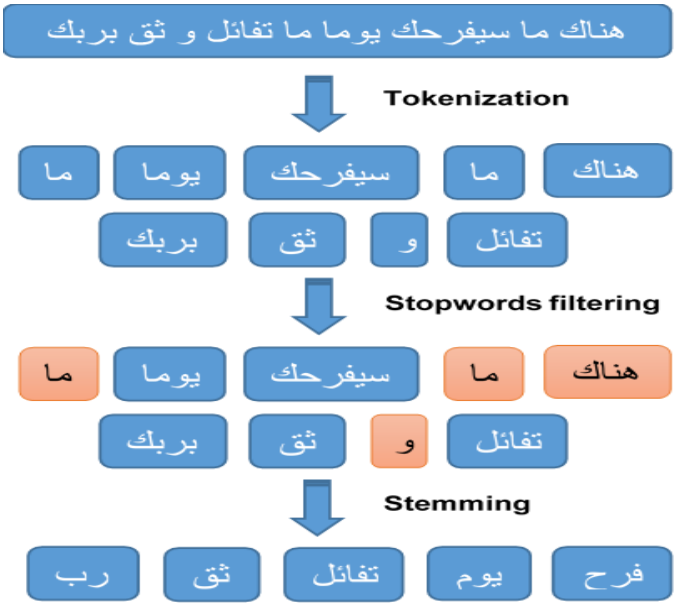

Figure 15: Exemple de prétraitement du texte.

7.4 La représentation vectorielle du texte.

Le texte original peut être vu comme une séquence de mots. Ce type de représentation est actuellement incompréhensible pour les algorithmes d'apprentissage automatique qui ont besoin de recevoir des représentations vectorielles numériques des entités à classer. La représentation vectorielle consiste à transformer chaque document en une séquence de nombres, dans laquelle chaque nombre correspond à un mot du vocabulaire de l'ensemble des documents ou corpus.

Pour transformer les documents de texte en vecteurs, on produit d'abord un vocabulaire avec tous les mots contenus dans les textes de l'ensemble d'entraînement. On produit ensuite une matrice numérique dans laquelle chaque ligne correspond à un des documents de texte et chaque colonne correspond à un mot du vocabulaire du corpus. Si le mot n'apparaît pas dans le document, on lui assigne le nombre 0. Par contre, s'il apparaît, on peut lui assigner le nombre 1, ou celui correspondant au total de fois que le mot apparaît dans le document. Cette dernière matrice s'appelle la matrice de fréquences.

La matrice numérique résultante peut être passée alors à l'algorithme de classement qui sera capable de l'interpréter et de travailler avec elle. Cette représentation est aussi appelée le sac de mots (bag of words). Le tableau 9 illustre ce processus:

[7] http://arabicstopwords.sourceforge.net/

Document / Word	فكرة	جميلة	تحياتي	الخالصة	لكم	جميعا	دمتم	سالمين
فكرة جميلة	1	1	0	0	0	1	0	0
تحياتي الخالصة لكم جميعا	0	0	1	1	1	1	0	0
تحياتي لكم فكرة جميلة دمتم سالمين	1	1	1	0	1	0	1	1

Tableau 9: Représentation du sac de mots (BOW).

7.5 La transformation des caractéristiques

D'une part, on peut penser qu'il pourrait être judicieux de faire une représentation numérique qui accorde plus d'importance aux mots dont la fréquence est haute dans la catégorie à laquelle ils appartiennent et basse dans les autres catégories, en vue de pondérer la valeur numérique de chaque mot selon l'information qu'elle apporte pour le classement. Aussi, c'est l'effet produit par la pondération TF-IDF, introduite par [33], qu'on utilisera dans la partie pratique de notre recherche. Il y a par ailleurs d'autres pondérations possibles comme le $\mathcal{X}^2$, le ${\mathcal{X}_p}^2$, le gini index, et le gain d'information, expliquées par [34], qui permettent aussi de capturer cette sorte de relations entre mots et documents.

Dans cette étape, nous considérons les termes restant après l'étape de prétraitement comme descripteurs. Ces descripteurs ont un rôle important pour le classement des sentiments. Pour réaliser l'opération d'apprentissage, nous avons proposé deux représentations : Le modèle de représentation Word Embedding et le modèle de pondération TF-IDF.

7.5.1Term Frequency - Inverse Document Frequency (TF-IDF)

Le modèle de pondération TF-IDF (Term Frequency - Inverse Document Frequency) dans le modèle vectoriel, un document est représenté sous forme d'un vecteur dans un espace engendré par tous les termes d'indexation. La dimension de cet espace est le nombre de termes d'indexation de la collection de document. Les coordonnées d'un vecteur document sont les poids des termes d'index dans ce document. Un poids plus important est donné aux mots caractéristiques d'un document présenté sous forme d = (w1, w2, w3, …, wn).

Dans, un premier temps, il est nécessaire de calculer la fréquence d'un terme (TF).

Celle-ci correspond au nombre d'occurrences de ce terme dans le document considéré. Ainsi, pour le document dj et le terme ti, la fréquence du terme dans le document est donnée par l'équation suivante :

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$\boldsymbol{n_{i,j}}$: est le nombre d'occurrences du terme t_i dans d_i.
$\sum_k \boldsymbol{n_{k,j}}$: est le nombre de termes dans le document.

La fréquence inverse de document (Inverse Document Frequency) mesure l'importance du terme dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme. Elle est définie de la manière suivante :

$$IDF_i = \log(\frac{D}{d_j : t_i \epsilon d_j})$$

D représente le nombre total de documents dans le corpus.
$d_j : t_i \epsilon d_j$: est le nombre de documents dans lesquels le terme t_i apparait.

Enfin, le poids s'obtient en multipliant les deux mesures :

$$TF_i - IDF_{i,j} = TF_{i,j} * IDF_i$$

7.5.2 Word Embedding

Le Word Embedding est essentiellement une forme de représentation de mots qui relie la compréhension humaine du langage à celle d'une machine. Il désigne un ensemble de techniques de Machine Learning qui visent à représenter les mots ou les phrases d'un texte par des vecteurs de nombres réels, décrits dans un modèle vectoriel.

Ces nouvelles représentations de données textuelles ont permis d'améliorer les performances des méthodes de traitement automatique des langues, comme le Topic Modeling ou le Sentiment Analysis.

Le Word Embedding repose sur la théorie linguistique fondée par Zelling Harris et connue sous le nom de Distributional Semantics. Cette théorie considère qu'un mot est caractérisé par son contexte, c'est à dire par les mots qui l'entourent. Ainsi, des mots qui partagent des contextes similaires partagent également des significations similaires. Les algorithmes de Word Embedding sont le plus souvent employés pour décrire des mots à travers de vecteurs numériques, mais ils peuvent également être utilisés pour construire des représentations vectorielles de phrases entières, de données biologiques comme les séquences d'ADN, ou encore des réseaux représentés comme des graphes.

Une fois terminé le processus de vectorisation du texte, on peut finalement passer à l'étape suivante : le classement.

8 Le classement

Ils existent plusieurs méthodes de classement supervisée et beaucoup d'entre elles ont été testés pour le classement des sentiments. On peut citer les réseaux de neurones, la régression logistique, les arbres de décision, les machines à support de vecteurs, les réseaux de neurones convolutifs, les réseaux de neurones récurrents ainsi que des méthodes combinant différents classifieurs.

8.1 Les réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs (CNN) sont des réseaux initialement créés pour des tâches liées aux images qui peuvent apprendre à capturer des caractéristiques spécifiques indépendamment de la localité.

Pour un exemple plus concret, imaginons que nous utilisons des CNN pour distinguer les images de voitures par rapport aux images de chiens. Étant donné que CNN apprend à capturer des fonctionnalités, quel que soit leur emplacement, CNN apprendra que les voitures ont des roues et que chaque fois qu'elle voit une roue, où qu'elle se trouve sur la photo, cette fonctionnalité s'active.

8.2 Les réseaux de mémoire à long terme à court terme (LSTM)

Les réseaux de mémoire à long terme à court terme (LSTM) sont un type spécifique de réseau de neurones récurrents (RNN) qui sont capables d'apprendre les relations entre les éléments d'une séquence d'entrée.

LSTM est un type de réseau doté d'une mémoire qui "se souvient" des données précédentes de l'entrée et prend des décisions en fonction de cette connaissance. Ces réseaux conviennent plus directement aux entrées de données écrites, car chaque mot d'une phrase a une signification basée sur les mots qui l'entourent (mots précédents et suivants).

Dans notre cas, il est possible qu'un LSTM nous permette de capturer un sentiment changeant dans un tweet. Par exemple, une phrase telle que : *Au début, je l'aimais bien, mais ensuite je l'ai détesté* a des mots avec des sentiments contradictoires qui finiraient par semer la confusion sur un simple réseau Feed-Forward. Le LSTM, de son côté, pourrait apprendre que les sentiments exprimés à la fin d'une phrase signifient davantage que ceux exprimés au début.

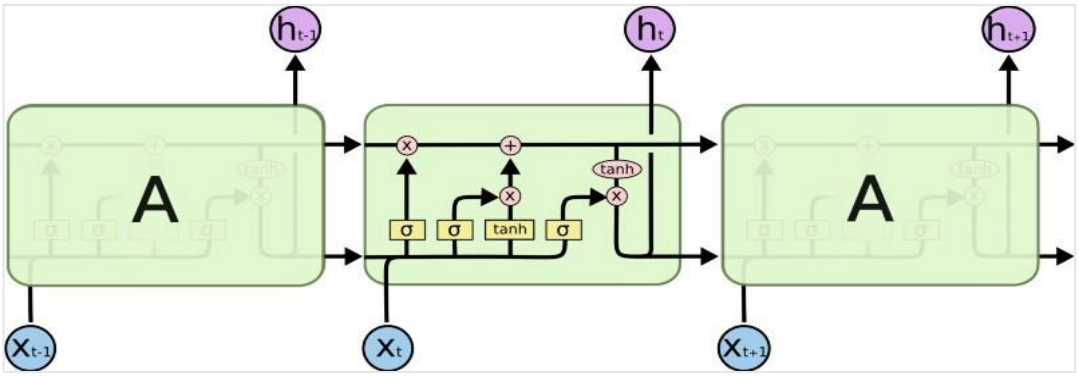

Figure 16: Les réseaux de mémoire à long terme à court terme.
Source : colah.github.io

Plutôt que de chercher à optimiser un seul classifieur en choisissant les meilleures caractéristiques pour un problème donné, les chercheurs ont trouvé plus intéressant de combiner plusieurs méthodes de classement.

La multiplication des travaux sur la combinaison de classifieurs a entrainé la mise au point de nombreux schémas traitant les données de manières différentes. Nous utilisons dans cette partie la combinaison des deux classifieurs d'apprentissage en profondeur (LSTM et CNN).

8.3 Modèle CNN-LSTM

Notre modèle CNN-LSTM consiste en une couche initiale LSTM qui reçoit les intégrations de mots pour chaque mot du tweet en tant qu'entrées. L'intuition est que ses mots de sortie vont stocker des informations non seulement sur le jeton initial, mais également sur tous les jetons précédents. En d'autres termes, la couche LSTM génère un nouveau codage pour l'entrée d'origine. La sortie de la couche LSTM est ensuite introduite dans une couche de convolution qui extraira les caractéristiques locales. Enfin, la sortie de la couche de convolution sera regroupée dans une dimension plus petite et finalement émise sous forme d'étiquette positive, négative, neutre ou mixte.

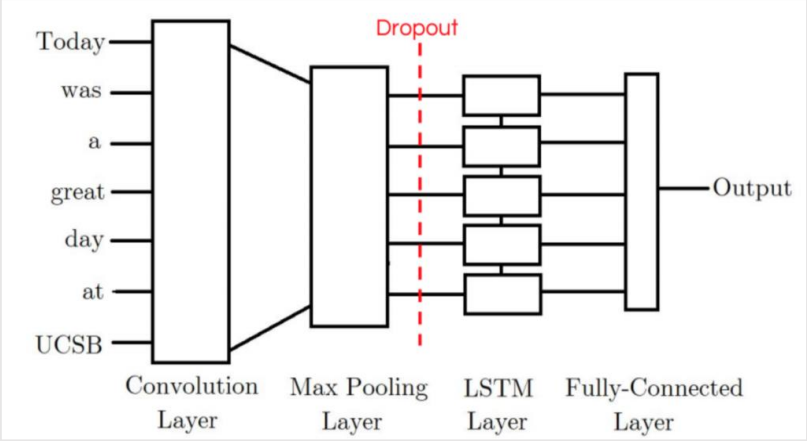

Figure 17: Le modèle CNN-LSTM.

8.4 Machine à vecteurs de support

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais Support Vector Machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires. SVM effectue le classement en recherchant l'hyper-plan qui différencie les classes que nous avons tracées dans un espace à n dimensions.

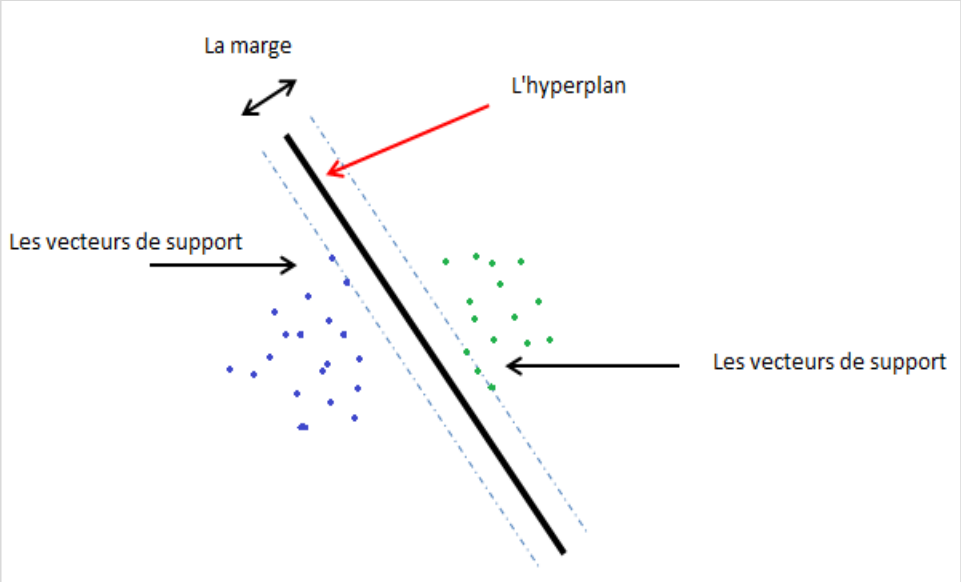

Figure 18 : Représentation d'un hyperplan et de vecteurs de support

L'un des principes de base des SVM se nomme l'hyperplan. L'hyperplan forme une sorte de ligne à l'intérieur d'un espace (Feature space). Cet hyperplan se voit également accompagné d'une marge de chaque côté. La taille de la marge est un facteur important. En effet, plus le système trouve un hyperplan avec une grande marge, plus les résultats sont optimaux.

8.5 La régression logistique

C'est une méthode de classement qui a été remise en valeur et qui est devenu une méthode populaire grâce à ses bonnes performances dans le classement automatique. Elle permet de calculer la probabilité d'appartenance à la catégorie k, $P_r(Y = k|X = x)$ comme suit :

$$P_r(Y = k|X = x) = \frac{e^{\beta x}}{1 + e^{\beta x}}$$

Où β est le vecteur de coefficients de régression qui doivent être estimés avec des exemples d'entraînement en utilisant, par exemple, la méthode de moindres carrés. La catégorie assignée sera celle dont la probabilité est la plus grande, c'est-à-dire :

$$\hat{y} = \underbrace{argmax}_{k \in \{1,\dots,k\}} P(Y = k|\, x)$$

La régression logistique a bénéficié de beaucoup de travaux de recherches et est devenue une méthode pratique dans de nombreux systèmes commerciaux à grande échelle, en particulier aux grandes sociétés d'Internet comme Google et Yahoo qui l'emploient pour apprendre de grands ensembles de données [14].

Ce modèle, en plus de pouvoir être utilisé seul, constitue en outre le bloc fondamental des réseaux neuronaux.

9 Évaluation de l'analyse

Les classifieurs SVM, Logistic Regression, CNN, LSTM et la combinaison CNN-LSTM forment les algorithmes que nous avons choisi pour mener cette étude, en raison de leur bonne performance et leur utilisation intensive. En vue d'entraîner les classifieurs, une méthode de validation croisée à 3 plis (three-fold cross validation) a été utilisée. Cela implique que le système scinde automatiquement le corpus en trois parties plus ou moins égales.

Le système mettra alors ce procédé en œuvre trois fois de suite en veillant qu'à chaque fois il utilise un ensemble différent comme données d'évaluation et les deux autres plis comme données d'apprentissage. Après 3 rotations, chaque tweet présent dans le corpus aura été une fois utilisé comme donnée d'évaluation.

Nous avons proposé deux tâches, la première consiste à identifier le langage utilisé (AS ou DM) et la deuxième centrée sur l'analyse des sentiments, étant donné un tweet écrit en AS ou DM, cette tâche consiste à le classer selon le sentiment/émotion exprimé par son auteur (positif, négatif, neutre ou mixte).

9.1 Résultats de la première tâche

Le corpus construit est composé de deux catégorie de langage utilisé dans les tweets, l'arabe standard et le dialecte marocain.

Tweet	Classe	Mots dialectal
Ar :القادم اجمل بإذن الله الله يجيب غير الصحة و السلامة Fr : Le futur est beau, si Dieu le veut .Que Dieu nous bénisse Santé et sécurité.	DM	• يجيب
Ar :أكيد كانت جد رائعة و كأنها عندها خبرة سنين برافو عليها Fr : Elle était certainement géniale et comme s'elle avait l'expérience des années Bravo.	DM	• برافو

Tableau 10: Exemple de tweets de l'arabe dialectal marocain.

Dans le premier exemple le tweet contient que des tokens de l'arabe marocain standard, tandis que le token « يجيب » est utilisé dans le dialecte pour exprimer un autre sens que celui dans l'AS, ce qui entraine une dégradation de précision dans la phase de prédiction.

Le tableau suivant contient le résultat de l'analyse du corpus :

Model	Features		Accuracy
LSTM	Word Embedding	Avec sw	**88.86**
		Sans sw	86.95
CNN-LSTM	Word Embedding	Avec sw	**88.43**
		Sans sw	87.06
CNN	Word Embedding	Avec sw	**88.32**
		Sans sw	87.37
SVM	TF-IDF	Avec sw	**88.39**
		Sans sw	87.25
LR	TF-IDF	Avec sw	**88.17**
		Sans sw	87.02

Tableau 11: Résultats du classement du type du langage utilisé.

Le tableau 11 présente les résultats et le classement des systèmes selon la précision pour la deuxième tâche. Les meilleurs résultats en termes de précision sont 88.86% ont été obtenu avec le LSTM sans élimination des mots fonctionnels et 87.37% avec CNN en les éliminant.

9.2 Résultats de la deuxième tâche

9.2.1 Approche Machine Learning

Malgré la répartition assez inégale des sentiments (positive pour la grande majorité, et la petite taille du type mixte), le système parvient néanmoins à réaliser des résultats plutôt convaincants.

Afin d'appuyer les propos ci-dessus, les tableaux suivants contiennent le résultat de l'analyse des corpus DM et AS et l'ensemble des deux, avec les classifieurs classiques + TF-IDF, les classifieurs du Deep Learning + Word Embedding.

Modèles classiques	Features			Accuracy			Accuracy all	
				AS	DM	AS-DM	AS	DM
SVM	TF-IDF	Uni-gram	Avec SW	83.50	67.01	78.70	82.06	70.08
			Sans SW	82.30	66.75	78.77	82.78	68.07
		Bi-grams	Avec SW	84.75	67.80	80.05	83.84	70.00
			Sans SW	83.47	68.68	79.41	82.54	71.40
		Tri-grams	Avec SW	84.15	67.40	80.00	83.33	70.35
			Sans SW	83.33	67.54	79.86	83.94	69.38
Logistic Regression		Uni-gram	Avec SW	82.23	64.38	78.18	82.03	69.82
			Sans SW	82.07	65.78	78.55	82.44	68.59
		Bi-grams	Avec SW	81.27	62.36	77.52	80.93	68.77
			Sans SW	81.14	60.88	78.08	81.14	70.26
		Tri-grams	Avec SW	81.31	62.10	77.96	81.58	68.68
			Sans SW	80.73	61.22	77.54	81.00	68.72

Tableau 12: Résultats des classifieurs classiques + TF-IDF sur les tweets marocains.

Modèles Deep Learning	Features		Accuracy			Accuracy all	
			AS	DM	AS_DM	AS	DM
CNN	Word Embedding	Avec SW	91.78	84.17	89.56	91.62	85.37
		Sans SW	91.39	83.82	89.08	91.12	84.78
LSTM		Avec SW	92.09	83.36	89.60	91.80	84.69
		Sans SW	91.64	82.87	89.49	91.36	85.04
LSTM-CNN		Avec SW	91.83	81.75	89.46	91.50	85.55
		Sans SW	91.67	82.00	88.93	91.46	85.26

Tableau 13: Résultats des classifieurs DL + Word Embedding sur les tweets marocains.

Le résumé des résultats d'évaluation expérimentale de la combinaison des classifieurs classiques avec un système de pondération de type TF-IDF présenté dans le tableau 12 concluait comme suit :

(1) Le classifieur SVM utilisant les différents d-grammes (Uni-, Bi- et Trigrams) atteint un très bon niveau de précision de 84.75% dans le premier corpus (AS) utilisant des bigrammes sans éliminer les mots fonctionnels (Stop Words) et une précision de 80.05% dans le troisième corpus (AS-DM), le corpus dialectal atteint des résultats inférieurs relativement au (AS) de 71.40% de précision, en raison de la taille réduite du corpus d'une part et par la grande variété d'écriture du dialecte marocain de l'autre.

(2) Pareillement pour le classifieur Logistic Regression, qui atteint 82.23% de précision pour le corpus (AS), utilisant les unigrammes sans éliminer les mots fonctionnels, et 78.55% pour le corpus (AS-DM) en éliminant les mots fonctionnels.

En ce qui concerne le tableau 13, les résultats de la combinaison des classifieurs du Deep Learning avec le modèle de représentation Word Embedding, ont été présenté comme suit :

(1) Le classifieur CNN a obtenu des résultats de précision de 91.78% en gardant les mots fonctionnels, et 91,39% en les éliminant dans le premier corpus (AS), et des résultats convaincants dans le corpus dialectal avec une précision de 85.37%, et 89.56% dans la combinaison des deux corpus.

(2) Le classificateur LSTM sans élimination des mots fonctionnels, avec le Word Embedding a obtenu le meilleur résultat de précision 92.09 % surperformant le résultat des autres classifieurs.

(3) La combinaison des deux classifieurs CNN-LSTM sans élimination des mots fonctionnels a obtenu un bon résultat avec une précision de 91.83% dans le premier corpus (AS), et 85.55% dans le corpus (DM).

Nous avons pensé à tester les performances de notre système sur la combinaison des deux corpus, pour cela on a lancé l'apprentissage sur toute la base de données, mais dans le test nous avons fait un test séparé (un test sur l'AS, et un test sur le DM). Par rapport à l'AS nous avons obtenu des résultats dans la colonne « Accuracy all » presque pareils que ceux obtenu dans la colonne « Accuracy » (où on a lancé l'apprentissage sur chaque corpus séparément), et par rapport au DM nous avons obtenu une augmentation moyenne de 2%. C'est pour cela nous avons décidé de travailler dans les prochains travaux avec un seul système.

En générale, la fonction Bigrammes et l'inclusion des mots fonctionnels ont obtenu les meilleurs résultats de précision dans tous les classifieurs.

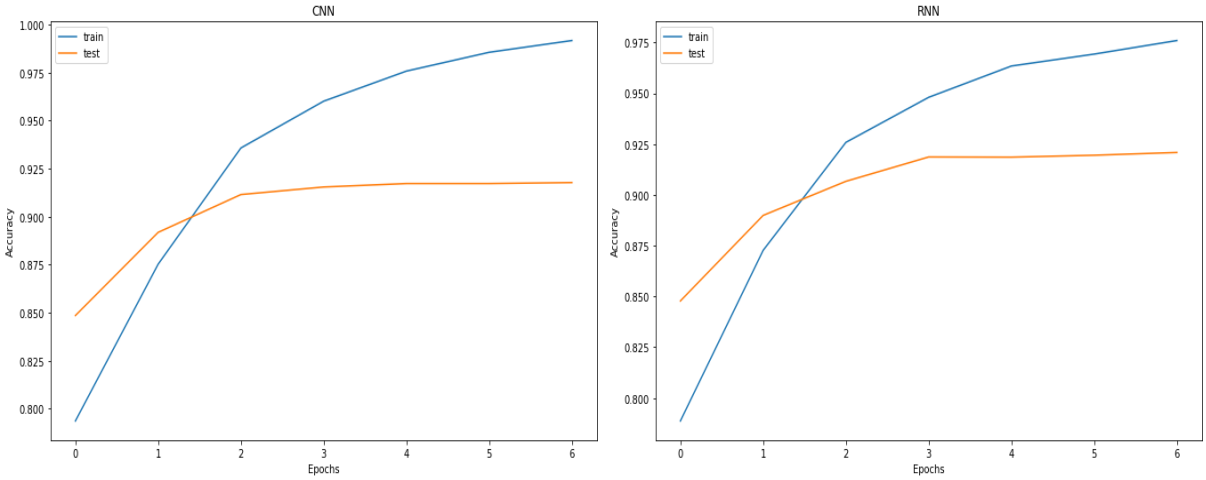

Figure 19: Résultats moyens de performance des classifieurs CNN et LSTM.

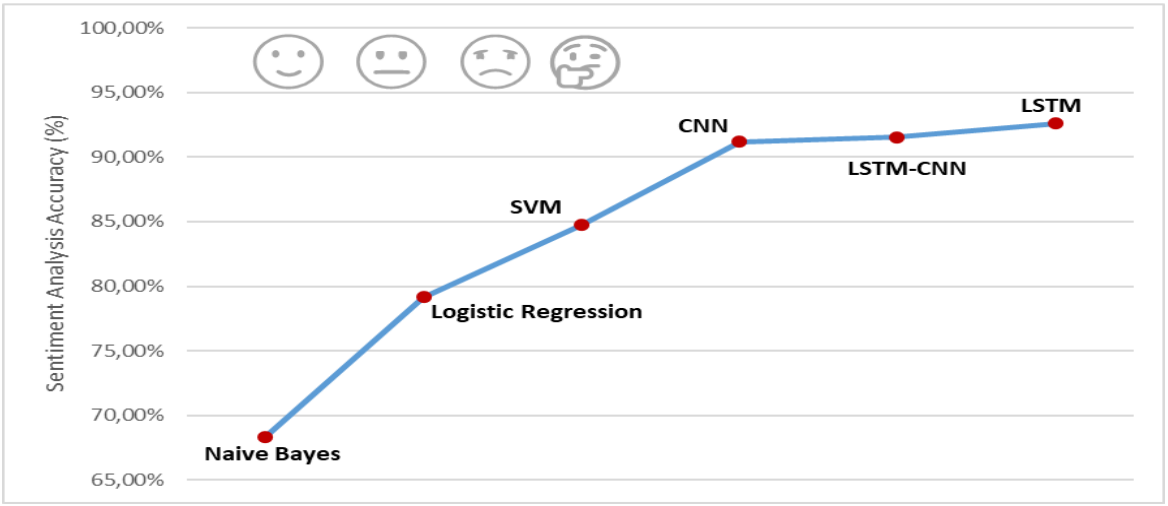

Figure 20: Comparaison les différents classifieurs sur le corpus AS.

Selon les résultats obtenus, il est clair que les algorithmes d'apprentissage profond dépassent les autres algorithmes classiques tel que les SVM, NB et Logistic Regression.

Comme mentionné avant, le type « Mixte » contient les tweets dans lesquelles l'auteur du tweet a exprimé à la fois un sentiment positif et négatif. Tandis que parfois le coté positive est le plus dominant ou l'inverse, donc le système classe le tweet avec le type dominant ce qui donne une dégradation de précision dans les résultats (dans le cas du classifieur SVM avec 3 classes la précision est de 86.60%, tandis qu'avec 4 classes elle est de 84.75% (baisse de 2%)).

Dans les deux tableaux précédents nous avons fait confiance à l'état d'art (l'utilisation des classifieurs classiques + TF-IDF et des classifieurs du Deep Learning + Word Embedding). Dans les deux tableaux ci-dessous (14 et 15) nous souhaitons tester les performances des méthodes classiques avec Word Embedding et des méthodes Deep Learning avec TF-IDF.

Modèles classiques	Features		Accuracy		
			AS	DM	AS_DM
SVM	Word Embedding	Avec SW	54.84	54.64	51.76
		Sans SW	54.99	54.38	51.94
Logistic Regression		Avec SW	52.42	55.17	52.60
		Sans SW	52.63	53.94	52.89

Tableau 14: Résultats des classifieurs classiques + Word Embedding.

Modèles Deep Learning	Features		Accuracy		
			AS	DM	AS_DM
CNN	TF-IDF	Avec SW	**76.02**	**72.84**	**73.45**
		Sans SW	**75.66**	**72.93**	**72.84**
LSTM		Avec SW	**75.10**	**71.52**	**72.32**
		Sans SW	**75.43**	**71.71**	**71.55**
CNN-LSTM		Avec SW	**74.24**	**71.63**	**73.09**
		Sans SW	**73.89**	**72.07**	**73.43**

Tableau 15: Résultats des classifieurs DL + TF-IDF.

Nous constatons que les résultats de classement présentés dans les deux tableaux sont moins bons. Donc suivant l'état d'art [35], et les expériences que nous avons effectuées, nous pouvons conclure que la meilleure précision, a un modèle qui utilise des fonctionnalités construites à l'aide de la combinaison des algorithmes du Deep Learning avec un modèle de représentation de type Word Embedding, et de la combinaison des algorithmes classiques avec un système de pondération de type TF-IDF.

9.2.2 Approche lexicale

Les approches lexicales utilisent des dictionnaires de mots subjectifs, considérés comme des références universelles. Dans ces dictionnaires, une polarité est associée à chacun des mots. Quel que soit le contexte dans lequel il sera inséré, le mot devrait ainsi avoir toujours la même polarité. On donne ensuite au document un score d'opinions en fonction de la présence de mots issus de ces dictionnaires dans le texte.

Compte tenu de la rareté des travaux de recherche réalisés dans le domaine de l'analyse des sentiments en général pour la langue arabe, et l'absence des travaux utilisant l'approche lexicale pour le dialecte marocain en particulier, nous avons été obligés de mettre en place un dictionnaire étiqueté. Notre dictionnaire constitué d'un lexique de plus de 30 000 mots que nous avons étiqueté manuellement.

9.2.2.1 Dictionnaire du lexique

La construction de lexiques de sentiments est une tâche très difficile qui conditionne le succès de l'approche basée sur le lexique. Les défis proviennent de la complexité de la langue arabe et du grand nombre de mots à prendre en compte. De plus, déterminer la polarité de nombreux mots peut être très difficile pour de nombreuses raisons, telles que les significations et connotations différentes (et donc la valeur de la polarité différente) de chaque mot en fonction du contexte et du contexte culturel de la personne qui publie le tweet.

Le tableau ci-dessous montre un exemple où un mot peut avoir plusieurs sens/polarités selon le contexte :

Tweet	Mot	Polarité
Ar : أنت في تقدم Fr : Vous êtes en progrès	تَقَدُّم	Positif
Ar : تقدم إلى الأمام Fr : Avancez	تَقَدَّم	Neutre

Tableau 16: Exemple de difficultés d'annotation du lexique.

Les tableaux 17 et 18 montrent des statistiques sur le dictionnaire construit :

Positif	Négatif	Neutre	**Total**
2630	**2057**	**13995**	**18683**

Tableau 17: dictionnaire du lexique extrait de la base de données AS.

Positif	Négatif	Neutre	**Total**
1291	**702**	**8902**	**10895**

Tableau 18 : dictionnaire du lexique extrait de la base de données DM.

Le dictionnaire comporte plus de 3921 mots positifs, 2759 mots négatifs et 22897 mots neutres. Le sentiment n'est pas mesuré directement avec ce dictionnaire, mais plutôt au moyen de deux ensembles de règles permettant la prise en compte des négations des termes.

Par exemple, le sentiment positif/négatif est mesuré en utilisant les deux règles suivantes :

- Mots positifs précédés d'une négation (…, لا, لن, لم) sont considérés comme des mots négatifs.
- Mots négatifs précédés d'une négation sont considérés comme des mots positifs.

La négation dans le dialecte Marocain s'exprime en ajoutant des affixes au termes, par exemple le terme « بغيت » qui signifie « je veux », pour l'inverser on lui ajoute le préfixe « ما » et le suffixe « ش », pour donner « مابغيتش » qui signifie « je ne veux pas ».

Le sentiment positif est mesuré en recherchant des mots positifs non précédés d'une négation ainsi que des termes négatifs après une négation. Cependant, nos propres expériences suggèrent que ces deux règles ont plus de valeur prédictive et pourrait même légèrement améliorer la mesure des sentiments positifs.

9.2.2.2 Prétraitement du texte

Les tâches de prétraitement du texte sont cruciales et inévitables pour tout outil de l'analyse des sentiments, surtout lorsque l'approche basée sur le lexique est adoptée. La figure 21 illustre la succession des étapes de prétraitement dans cette étape.

Figure 21: Processus de prétraitement dans l'approche lexicale.

Les étapes du prétraitement restent les mêmes que dans l'approche machine Learning. La quatrième étape de prétraitement appliquée dans cette étude est la réduction à la tige (Stemming). Deux méthodes d'analyse morphologique sont proposées pour la langue arabe : root-based Stemming (le retour à la racines) et light Stemming (élimine uniquement les affixes communs d'un mot).

Dans cette partie, nous avons appliqué le light Stemming. La principale raison de ce choix est que beaucoup de mots qui partagent la même racine ont des significations ou des sentiments complètement différents. Par exemple, les deux mots suivants «اللاعبون » et «يتلاعب » qui signifient « les joueurs » et « falsifier » ont respectivement la même racine « لعب » ce qui signifie «jouer», mais sémantiquement, en termes de sentiment, ils sont très différents. En revanche, si nous appliquons uniquement le light Stemming, les affixes seront simplement supprimés et le résultat sera « لاعب » et « تلاعب » qui signifient «joueur» et «falsifier». Ainsi, nous maintenons le sentiment correct du mot.

9.2.2.3 Classement du corpus de test des tweets

Pour déterminer la classe de chaque tweet, un score est calculé pour chaque sentiment à l'aide des mots du sentiment contenu dans le tweet, pour construire un vecteur qui va représenter le tweet.

Pour chaque sentiment (positif, négatif et neutre), son score est calculé de la manière suivante:

$$score_{sentiment} = \frac{nombre\ de\ mots\ du\ sentiment\ présents\ dans\ le\ tweet}{nbr\ total\ de\ mots\ présents\ dans\ le\ tweet}$$

Les valeurs finales des scores déterminent la polarité de l'ensemble du tweet, en le représentant sous forme d'un vecteur de trois dimensions :

$score_{positif}$	$score_{négatif}$	$score_{neutre}$

Figure 22: Vecteur tweet.

Ci-dessous les vecteurs possibles pour chacune des classes :

$score_{positif}$	0	$score_{neutre}$

Tweet positif

0	$score_{négatif}$	$score_{neutre}$

Tweet négatif

0	0	$score_{neutre}$

Tweet neutre

$score_{positif}$	$score_{négative}$	$score_{neutre}$

Tweet mixte

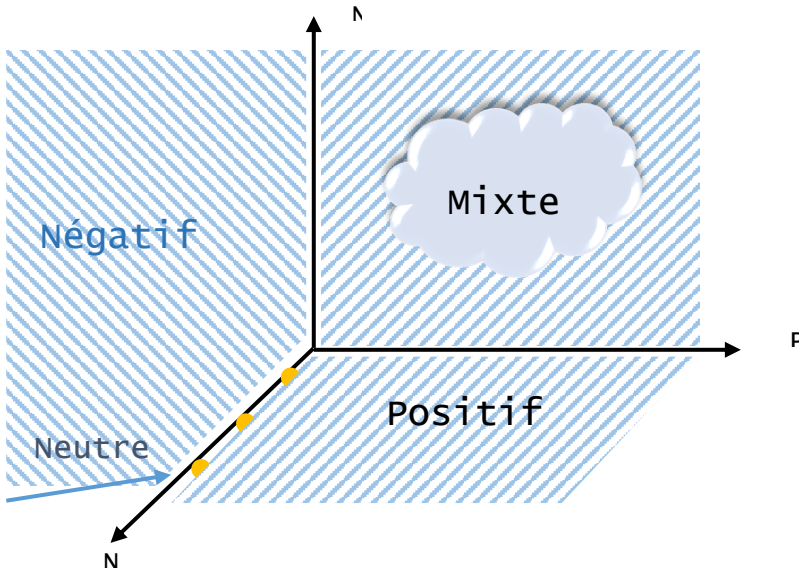

Figure 23: Représentation 3D de l'approche lexicale proposée.

Un classifieur est construit avec le vecteur créé comme donnée d'entrée, déterminant la classe au quelle le tweet appartient. Ensuite, la précision de chaque classifieur sera calculée.

Le tableau suivant contient les résultats de l'analyse des corpus (le corpus DM, le corpus AS et l'ensemble des deux).

Model	Stop words	Accuracy		
		AS	DM	AS_DM
CNN	Avec sw	**90.80**	**85.42**	**89.25**
	Sans sw	**90.85**	**85.30**	**89.14**
LSTM	Avec sw	**90.88**	**84.53**	**89.62**
	Sans sw	**90.63**	**84.02**	**88.66**
SVM	Avec sw	**82.04**	**74.14**	**78.11**
	Sans sw	**81.49**	**73.25**	**77.80**
Logistic Regression	Avec sw	**81.08**	**71.77**	**77.96**
	Sans sw	**80.63**	**71.51**	**77.54**

Tableau 19: Résultats de classement par l'approche lexicale (4 classes).

Le résumé des résultats d'évaluation expérimentale de l'approche lexicale est présenté dans le tableau 19 concluait comme suit :

(1) Le classifieur CNN a obtenu des résultats de précision de 90.85% en éliminant les mots fonctionnels, et 90,80% en les gardant dans le premier corpus (AS), et des résultats convaincants dans le corpus dialectal avec une précision de 85.42%, et 89.25% dans la combinaison des deux corpus.

(2) Le classifieur LSTM avec élimination des mots fonctionnels, a obtenu le meilleur résultat de précision 90.88 % dans le premier corpus (AS), surperformant le résultat des autres classifieurs.

(3) Le classifieur SVM atteint un niveau de précision de 82.04% dans le premier corpus en gardant les mots fonctionnels et une précision de 78.11% dans le troisième corpus (AS-DM), le corpus dialectal atteint des résultats inférieurs relativement au (AS) de 74.14% de précision, en raison de la taille réduite du corpus d'une part et par la grande variété d'écriture des marocains.

(4) Pareillement pour le classifieur Logistic Regression, qui atteint 81.08% de précision pour le corpus (AS) en gardant les mots fonctionnels, et 77.96% pour le corpus (AS-DM).

Comme mentionné avant, le type « Mixte » cause une dégradation de précision dans les résultats même dans l'approche lexicale, le tableau suivant montre les résultats de classement pour 3 classes.

Model	Stop words	Accuracy		
		AS	DM	AS_DM
SVM	Avec sw	**84.56**	**76.02**	**80.37**
	Sans sw	**83.89**	**75.56**	**80.14**
Logistic Regression	Avec sw	**83.78**	**73.56**	**80.56**
	Sans sw	**83.37**	**73.12**	**78.35**

Tableau 20: Résultats de classement par l'approche lexicale (3 classes).

Selon les résultats obtenus, il est clair que les performances de l'approche ML sont meilleurs que ceux de la deuxième approche lexicale. Ceci peut être interprété par l'apport positive de l'aspect sémantique présent dans l'approche ML et absent dans l'approche lexicale, sur la qualité du classement. Nous pensons que l'implication des autres aspects linguistiques tel que le type de mots (sujet, verbe, adjectifs. . .) peuvent améliorer le processus d'analyse des sentiments.

10 Résumé des travaux sur l'analyse des sentiments en arabe

Etude	Taille de la base de données	Arabe	Approche	Résultat
Moncef, Hanae (2019)	**13.550**	Marocain (Standard & Dialectal)	Corpus-Based	**92.09%**
El Abdouli,Hassouni, Anoun (2017)	500	Marocain	Corpus-Based	69%
Shoukry, Rafea(2012)	1.000	Egyptien (Dialectal)	Corpus-Based	78.8%
El-Beltagy, Soliman Kalamawy (2017)	13.292	Egyptien (Standard & Dialectal)	Corpus-Based	58.1%
Nabil, Aly, Atiya (2015)	10.000	Egyptien (Standard & Dialectal)	Corpus-Based	69.1%
Al-Horaibi et Khan (2016)	2.000	Saoudite	Corpus-Based	64.85%
Al-Twairesh et al. (2017)	17.573	Saoudite (Standard & Dialectal)	Corpus-Based	62.27%
Abdulla et al. (2013)	2.000	Jordanian (Standard & Dialectal)	Corpus-Based	87.2%
Heikal, Torki, and El-Makky (2018)	3.315	Arabe (Standard & Dialectal)	Corpus-Based	64.30%

Tableau 21 : Résumé des travaux sur l'analyse des sentiments des tweets arabes.

Etude	Taille de la base de données	Arabe	Approche	Résultat
Moncef, Hanae (2019)	**13.550**	Marocain (Standard & Dialectal)	Lexicon-Based	**89.62%**
Zelmati, Mataoui(2016)	7.698	Algérien (Standard & Dialectal)	Lexicon-Based	79.13%
Ayyoub, Essa, and Alsmadi (2015)	900	Jordanian	Lexicon-Based	86.89%
Duwairi, Al-Refai (2014)	1.000	Jordanian (Standard & Dialectal)	Lexicon-Based	76.78%
El-Masri et al. (2017)	8.000	Arabe (Standard & Dialectal)	Lexicon-Based	66.50%
Abdul- Mageed et al. (2015)	3.015	Arabe (Standard & Dialectal)	Lexicon-Based	64.37%

Tableau 22 : Résumé des travaux sur l'analyse des sentiments des tweets arabes.

D'après ce tableau on remarque que nos résultats dépassent les résultats des autres travaux similaires.

11 Conclusion

Dans ce chapitre, nous avons présenté les méthodes que nous avons utilisées pour extraire les données à partir de Twitter, et les expériences effectuées sur les données collectées. Nous avons vu aussi les différentes étapes pour faire la représentation vectorielle des documents de texte (Tokenization, Stemming, stop words removal, Vectorisation et Transformation), et la discussion des résultats d'analyse sur les différents corpus par les deux approches. Nous avons trouvé que les algorithmes d'apprentissage profond dépassent les autres algorithmes classiques et donnent des résultats très intéressants.

Chapitre 4
AraSentiPedia

1 Introduction

Dans ce chapitre, nous décrivons le développement de l'application web ***AraSentiPedia*** qui facilite l'utilisation de notre système de classement proposé, de sorte que les développeurs et les chercheurs peuvent utiliser notre service pour former et tester leurs propres ensembles de données avec différents algorithmes afin de comparer leur exactitude et de tester l'un d'entre eux s'il en a besoin dans le processus de prédiction.

2 Outils de développement

2.1 Environnement de développement Pycharm

Pycharm est un IDE, Integrated Développent Environment (EDI environnement de développement intégré en français), spécialisé pour les langages de programmation Python et Django. Il offre de riches et nombreuses fonctionnalités en matière d'édition, de débogage, de développement et de tests.

2.2 MongoDB

MongoDB (de l'anglais humongous qui peut être traduit par « énorme ») est un système de gestion de base de données NoSQL orientée documents, répartissable sur un nombre quelconque d'ordinateurs et ne nécessitant pas de schéma prédéfini des données.

2.3 Flask

Flask est un Framework open-source de développement web en Python. Son but principal est d'être léger, afin de garder la souplesse de la programmation Python, associé à un système de Template.

3 Interfaces Graphiques

Dans cette partie, nous allons présenter quelques interfaces de l'application, répondant aux recommandations ergonomiques de guidage, de clarté et de souplesse. Nous avons choisi le développeur ou chercheur comme utilisateur principal de l'application, vu qu'il présente à travers ces interactions la partie majeure des principales fonctionnalités de l'application.

3.1 Logo de l'application

Figure 24: Logo de l'application AraSentiPedia

3.2 Authentification

Pour l'authentification on a opté de créer deux pages d'authentification, une pour un mode développeur et l'autre pour le mode client. Les deux pages d'authentification sont gérées par Flask Security qui permet aux utilisateurs de l'application de s'identifier par leurs logins et leurs mots de passe. Dans la page d'inscription des développeurs nous avons ajouté des champs pour le compte *Twitter développer*, car ce dernier reste confidentiel pour l 'utilisateur, et la collecte des données est strictement conditionné, au préalable, par twitter, pour chaque utilisateur. La figure ci-dessous représente l'interface d'inscription pour le mode développeur.

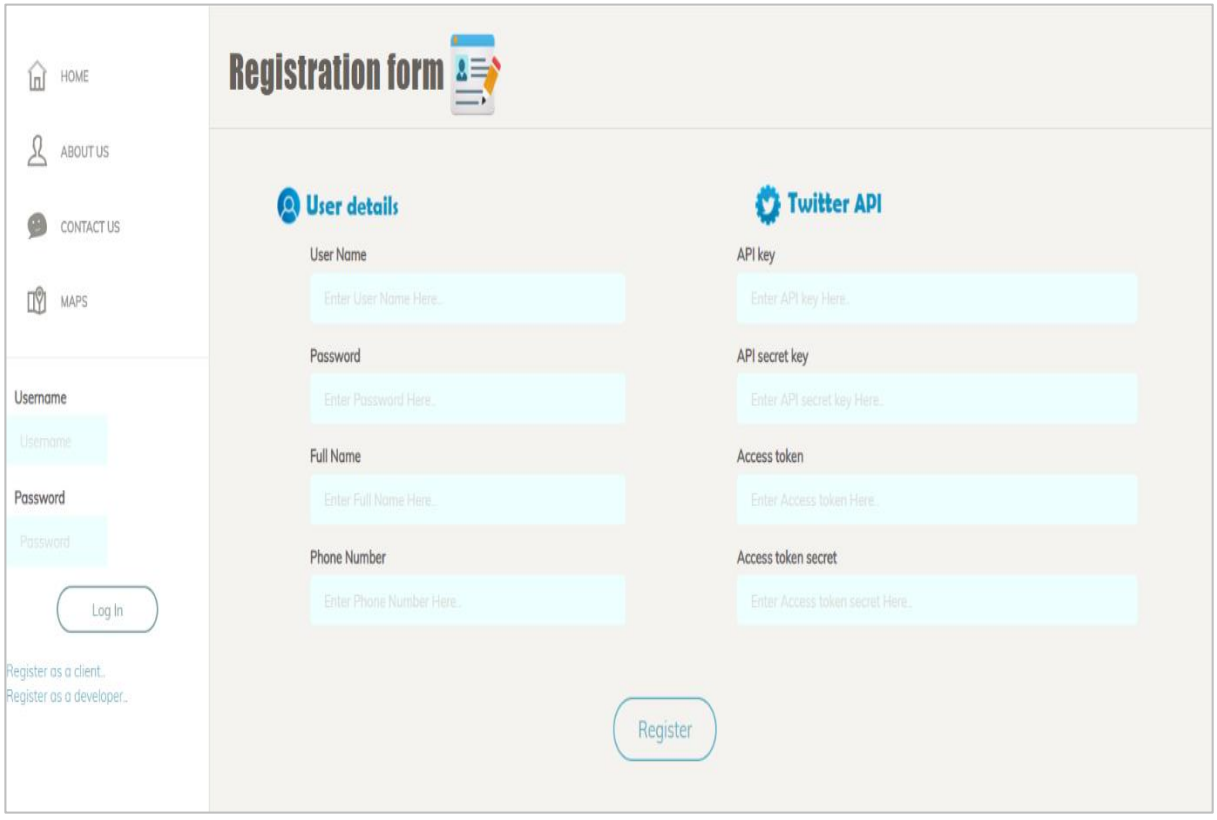

Figure 25: Interface d'inscription.

Si les coordonnés de l'utilisateur sont erronées, le système affiche un message d'erreur et l'invite à ressaisir ses coordonnés. Sinon l'utilisateur sera redirigé vers sa page d'accueil, dans laquelle il trouve un menu de tous les modules de l'application.

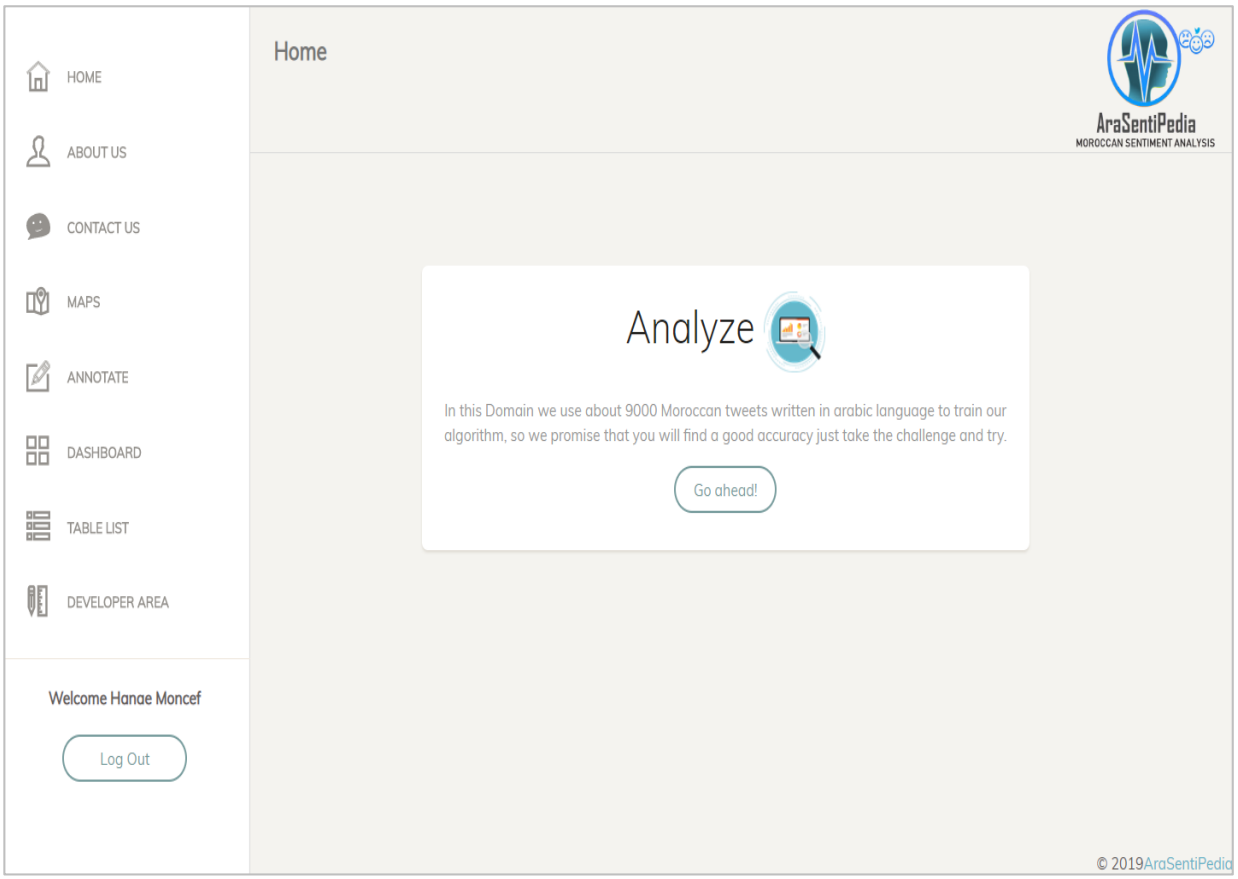

Figure 26: Interface « Home »

3.3 Interface « Home »

Dans la première version, cette application permet :

- L'analyse des sentiments des données Tweeter à partir d'un fichier Excel ou csv.
- L'extraction et l'analyse de nouvelles données Twitter en temps réel.
- L'annotation manuelle des données.
- L'utilisation plusieurs classifieurs (apprentissage, tests, enregistrement des résultats).
- Visualisation des statistiques sur la carte du Maroc.

3.4 Interface d'analyse des sentiments

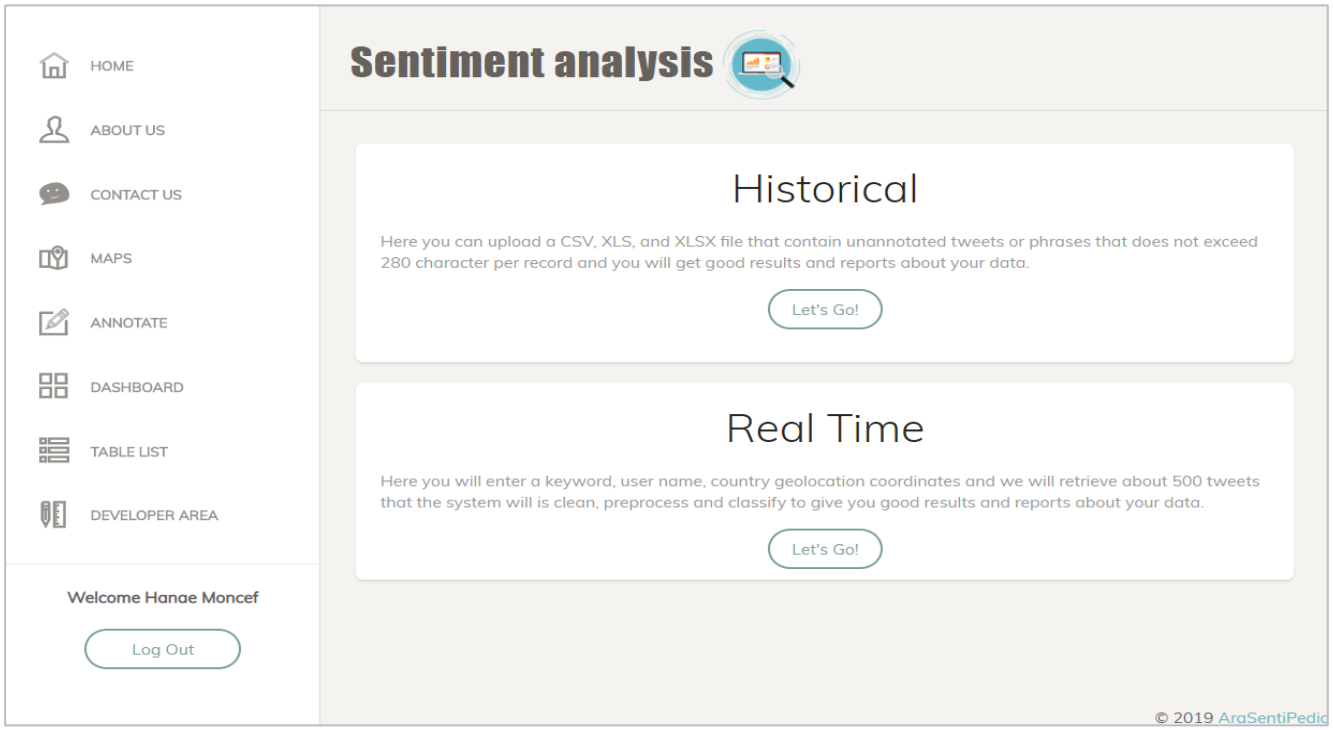

Figure 27: Interface d'analyse des sentiments.

Cette interface permet l'analyse des données historiques (à partir d'un fichier xlsx ou csv) ou en temps réel (l'extraction de nouvelles données à partir de Twitter).
Pour l'analyse en temps réel, l'application offre la possibilité de collecte de données par les trois méthodes (Search API, Streaming API, User timeline API).

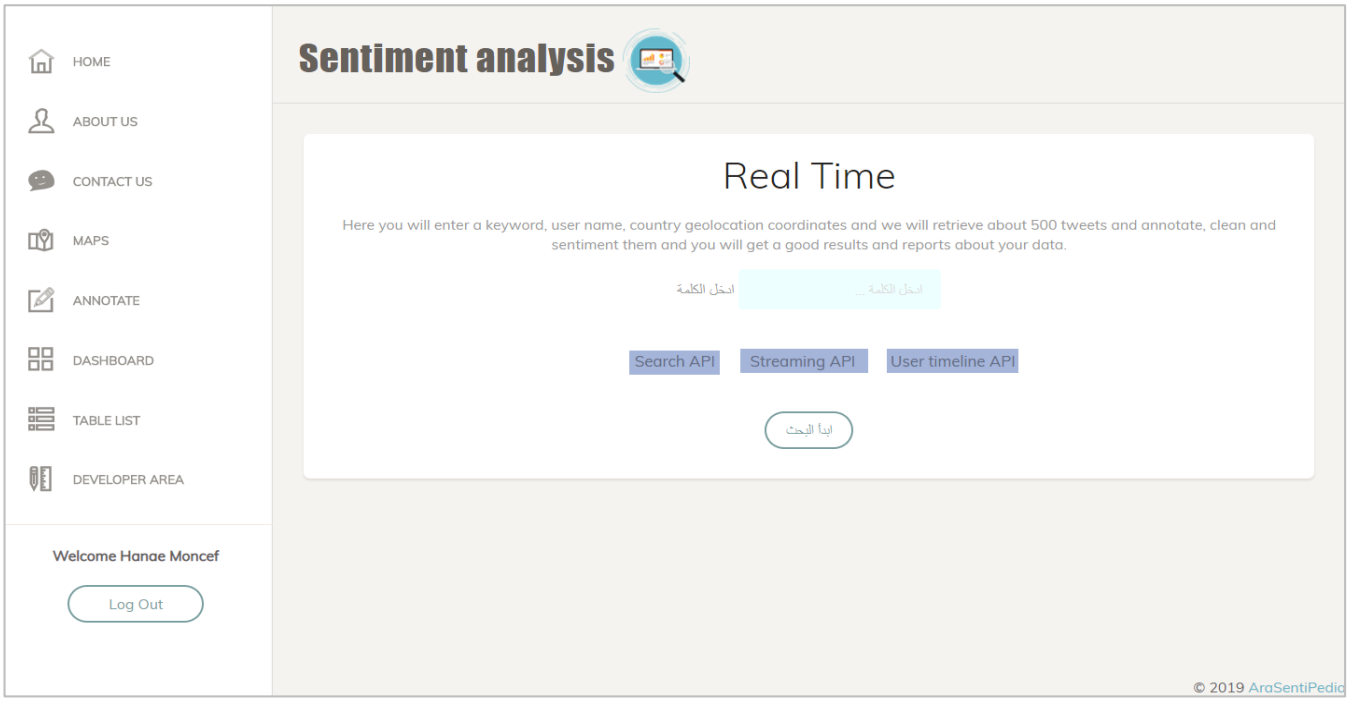

Figure 28: Interface de collecte de données.

3.5 Interface « Maps »

L'interface « Maps » présente les emplacements des tweets ainsi que les statistiques sur chaque ville sur la carte du Maroc.

Figure 29: Interface de la carte du Maroc.

3.6 Interface d'annotation

Figure 30: Interface d'annotation.

3.7 Interface du développement

Cette interface permet de choisir les fichiers d'apprentissage (fichier Train et fichier de Test) et la sélection de la liste des classifieurs à utiliser.

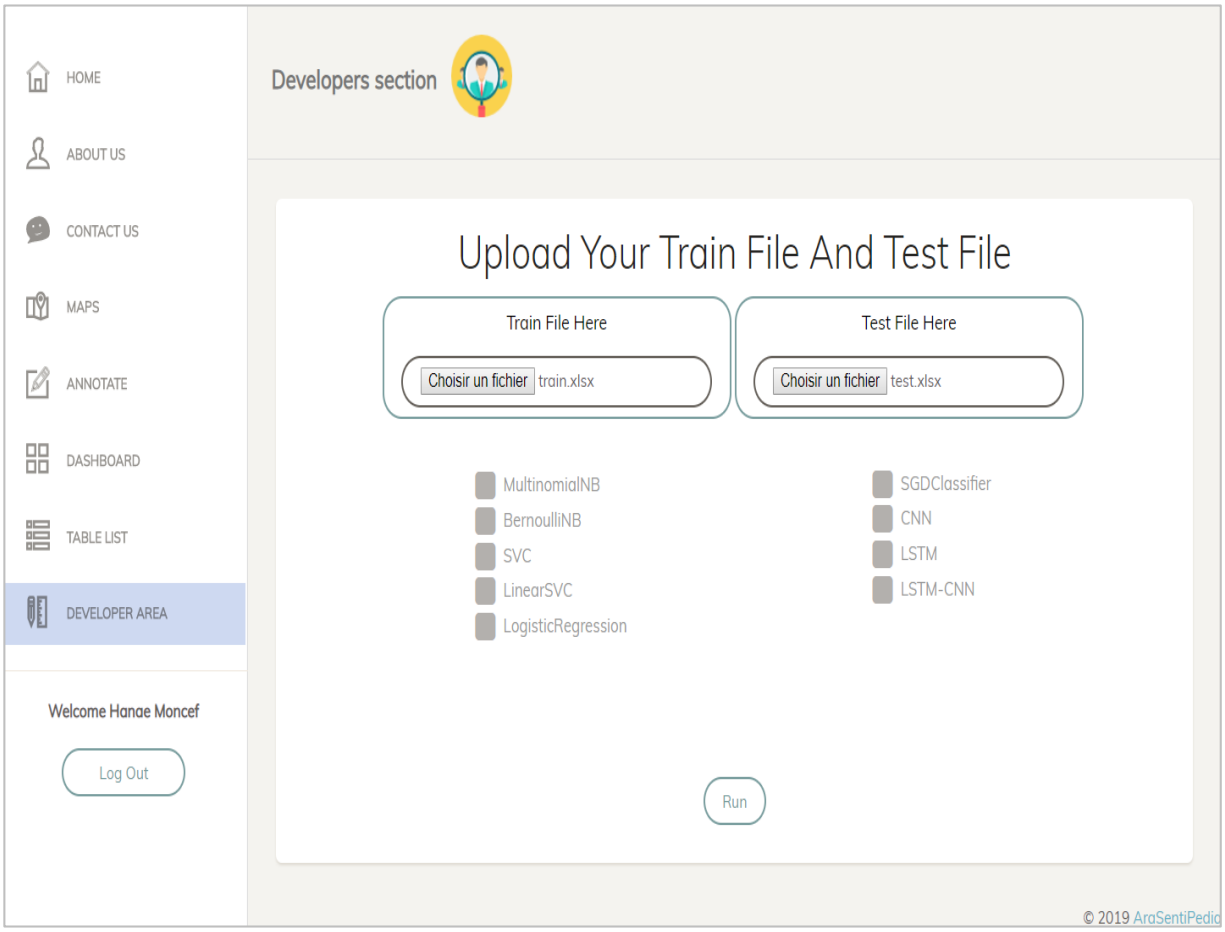

Figure 31 : L'interface du développement.

En appuyant sur le bouton **Run** le système lance l'apprentissage des classifieurs sélectionnés. Après la persistance des modèles, la fenêtre des résultats montre le rapport d'apprentissage contenant :

- Le temps qu'a pris l'entraînement du classifieur.
- Les paramètres du classifieur optimal.
- Le rapport du classement contenant la matrice de confusion pour le classement des données de test en utilisant les mesures des performances (Accuracy, Recall et F-score, Support (taille du fichier du test)).

3.8 Interface des résultats d'apprentissage

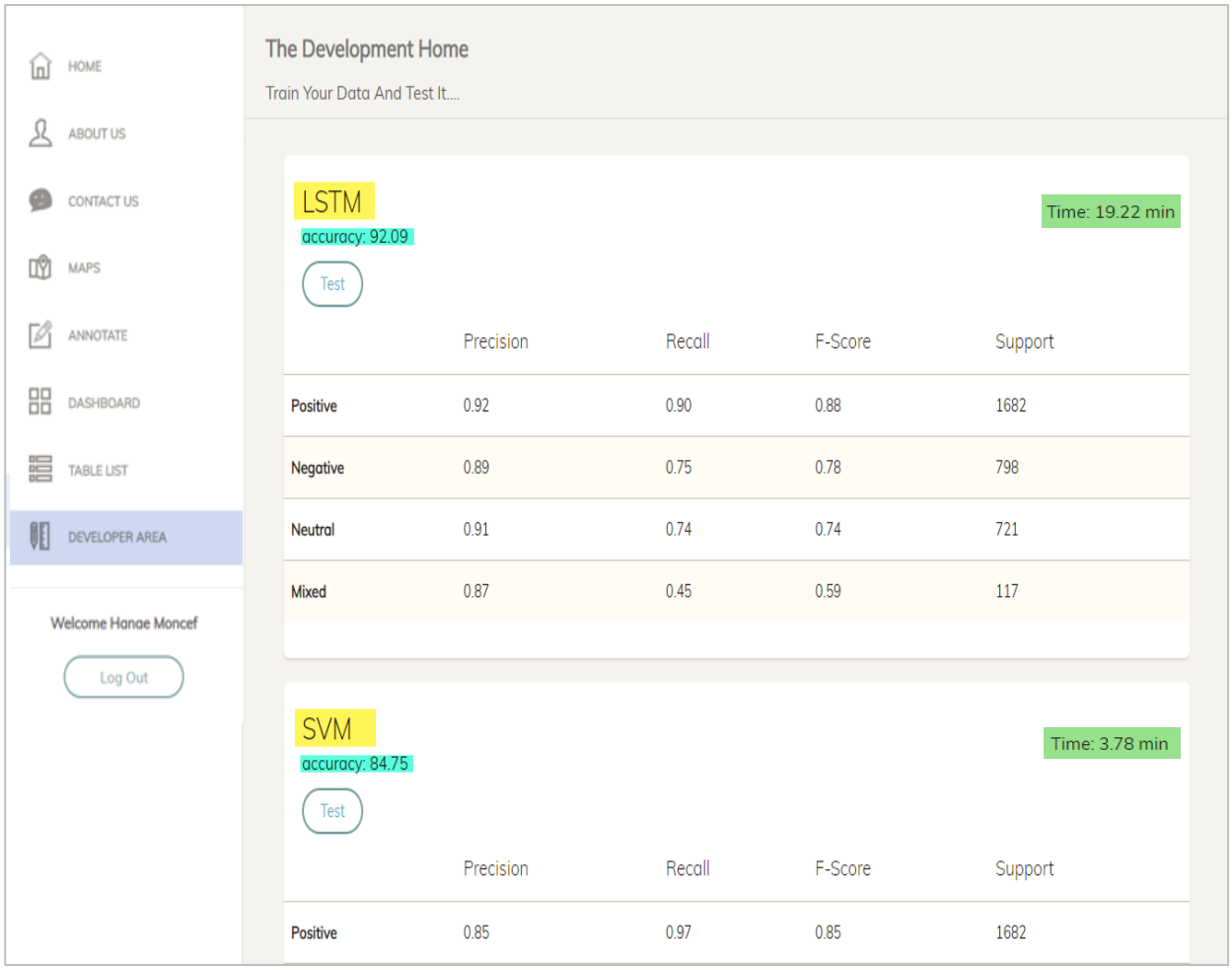

Figure 32: Interface des résultats d'apprentissage.

Après la phase d'apprentissage le système affiche le rapport d'apprentissage avec une option de Test du model créé avec de nouvelles données.

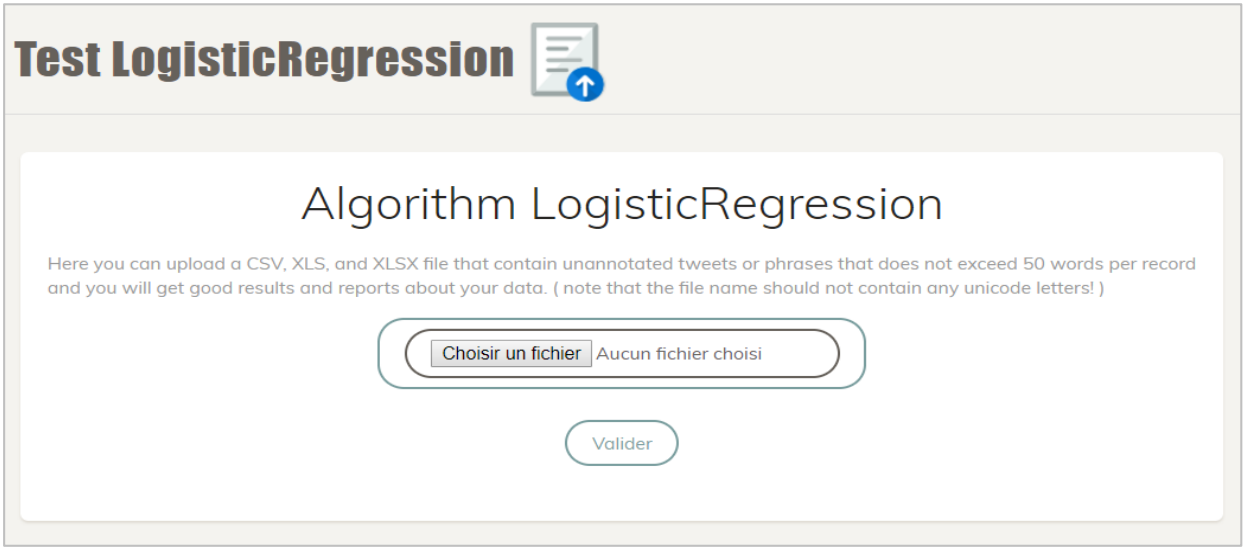

Figure 33: Interface de test.

3.9 Interface des résultats du test

Les résultats du test sont affichés dans le menu **Dashboard** offrant une vision globale sur les données du test :

- Statistiques sur les classes sous forme du 'Pie chart' et 'Bar chart '.
- Statistiques sur les mots les plus fréquents dans les données du test.
- Le Word Cloud.

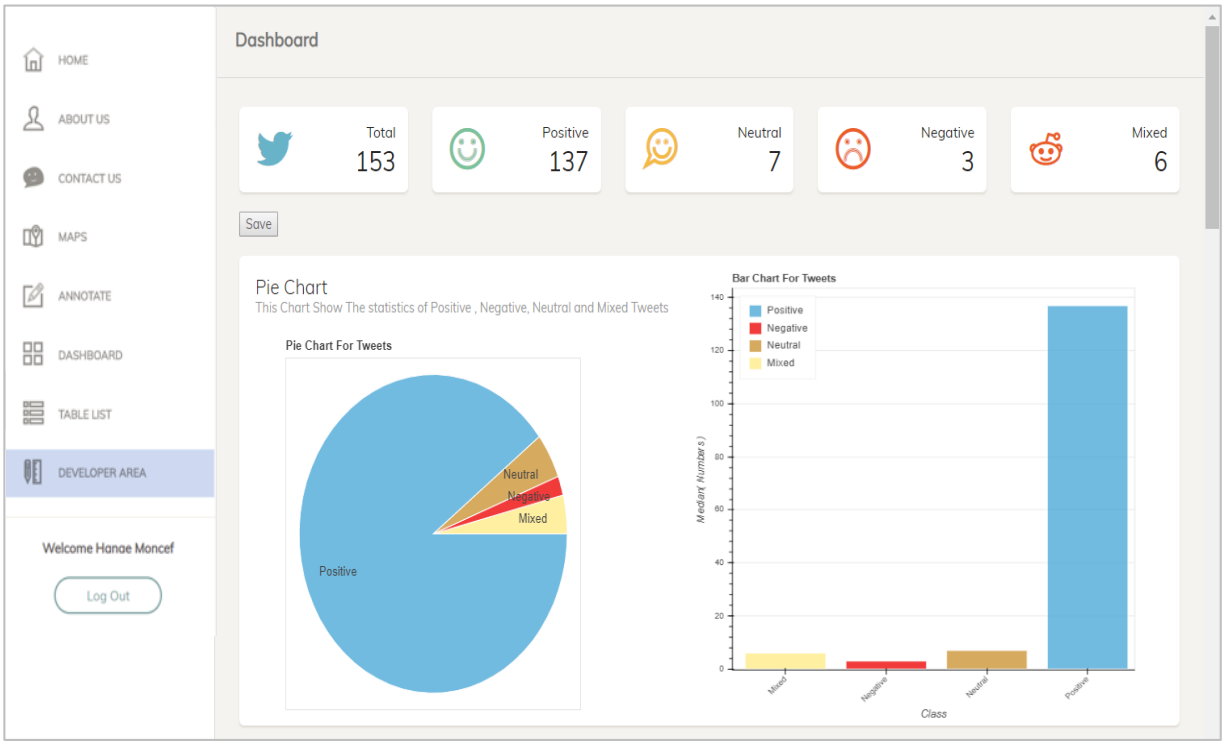

Figure 34: Interface des résultats du test.

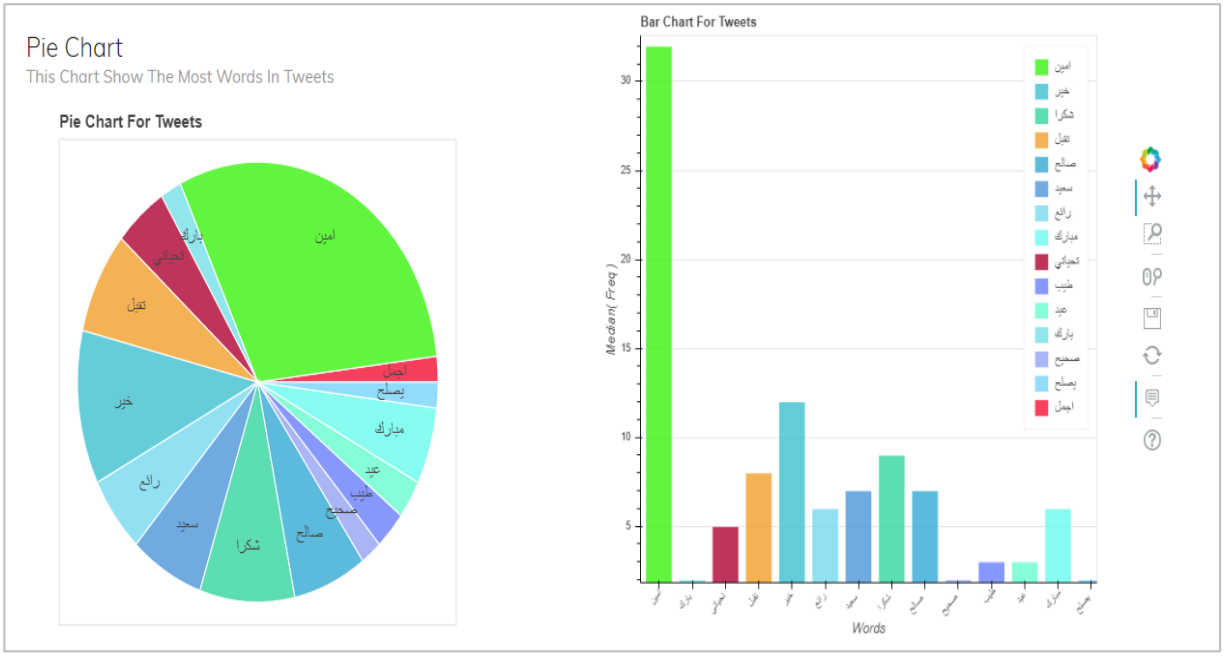

Figure 35: Interface des résultats du test.

Figure 36 : Interface des résultats du test

Nous avons proposé aussi d'ajouter une option d'enregistrement et d'affichage des résultats sous forme d'une série temporelle (time series) pour faciliter la comparaison des résultats tout au long de la phase des apprentissages/modifications du système.

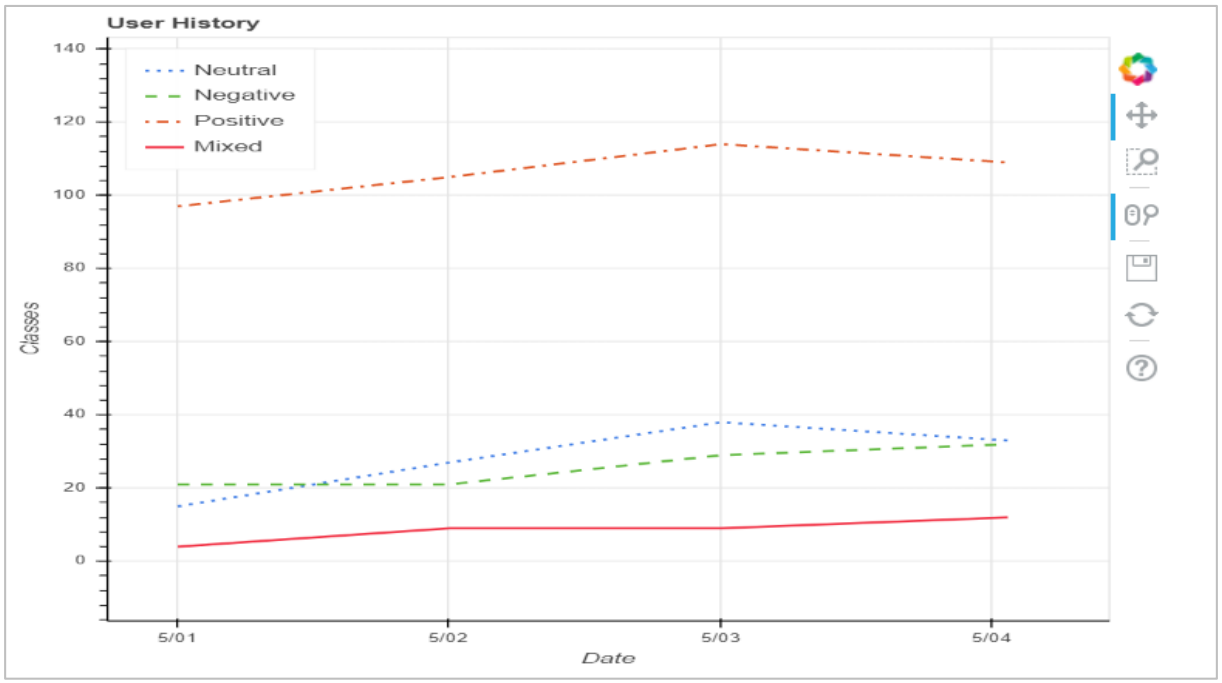

Figure 37: L'interface « Time series ».

4 Conclusion

Dans ce chapitre, nous avons présenté les principaux outils utilisés pour réaliser l'application AraSentiPedia, ainsi quelques interfaces principales de notre Système.

Conclusions et perspectives

Ce mémoire de fin d'études aborde l'analyse des sentiments dans les tweets marocains en arabe standard et dialectal.

Afin de recueillir les données nécessaires, Twitter offre particulièrement une API simple d'accès, qui nous a permis de collecter nos données d'une façon simple et rapide. En effet, nous avons fait appel à l'interface de programmation de Twitter pour récolter des tweets avec trois méthodes différentes, à savoir : Streaming API, Search API, et User_timeline API.

Pour mener à bien cette étude, nous avons collecté plus de 36 000 tweets. Après un prétraitement qui consiste à enlever les tweets contenant un langage non approprié, nous avons gardé 13 550 tweets que nous avons étiquetés manuellement. Chaque tweet se voit attribuer une étiquette Positive, Négative, Neutre ou Mixte et une classe AS ou DM.

Dans ce rapport, nous avons présenté : les principaux travaux qui s'intéressaient à l'analyse des sentiments, les différences entre les blogs et les Microblogs, la plateforme Twitter et les principaux travaux qui analysent les données de Twitter. Nous avons présenté, également, les différentes techniques que nous avons utilisées dans nos expérimentations ainsi qu'une analyse et discussion des résultats obtenus. Finalement, nous avons donné une description assez détaillée de l'application AraSentiPedia qui implémente toutes les techniques d'analyse de sentiments des tweets que nous avons étudiées.

Afin de développer un système automatique d'analyse des sentiments, nous avons implémenter deux approches. Une approche Machine Learning, dont le système reçoit des tweets déjà étiquetées et les classer. La deuxième approche, nommée approche lexicale, nécessite la construction d'un dictionnaire de mots. Le dictionnaire que nous avons mis en place contient 30 000 mots que nous avons étiquetés manuellement.

Au cœur de notre système, nous avons développé plusieurs algorithmes, qu'on peut répartir en deux classes distinctes : les algorithmes Deep Learning à savoir CNN, LSTM, CNN-LSTM et les algorithmes classiques tels que SVM, Logistic Regression, et Native Bayes.

En vue de vérifier l'efficacité des différents algorithmes, nous avons effectué plusieurs scénarios en utilisant plusieurs paramètres tels que les N-grammes, Stopwords removal, Term Frequency–Inverse Document Frequency TF-IDF, et Word Embedding.

L'étude que nous avons réalisée peut être considérée comme une des premières études de l'analyse des sentiments des tweets marocains avec une base de données assez importante de 13 550 tweets. Les deux études précédentes utilisaient moins de 1000 tweets et se basaient sur les émojis pour analyser les sentiments.

Les résultats que nous avons obtenus sont très intéressants et très prometteurs et montrent que nos systèmes sont très performants et très compétitifs dans leur domaine.

Ce stage de fin d'études nous a permis de travailler en équipe, en rédigeant nos deux premiers articles dans le domaine de la recherche, qui ont été soumis comme articles au journal King Saoud University.

Comme perspectives, notre système peut être amélioré en prenant compte l'instabilité du dialecte marocain qui entraîner le changement de sens de certains mots.

Par conséquent, des améliorations d'analyse du dialecte marocain sont nécessaires. Les prochaines étapes prévues comprennent :

1. Augmentation de la taille du jeu de données, en particulier du jeu de données DM.
2. atténuer le déséquilibre présent dans des ensembles de données.
3. Ajouter plus de paramètres, plus de fonctionnalités et de classifieurs.

Références

[1] A. El Abdouli, L. Hassouni, and H. Anoun, "Mining tweets of Moroccan users using the framework Hadoop, NLP, K-means and basemap," in *2017 Intelligent Systems and Computer Vision, ISCV 2017*, 2017, pp. 1–7.

[2] Bouillon Pierrette, "Traitement automatique des langues naturelles," bruxelle 1998.

[3] Kumar Ela, "Natural Language Processing," *India,I.K.International Publishing House* Pvt. Ltd 2011.

[4] Jean Véronis, "Natural Language Processing", *http://sites.univ- provence.fr/veronis,* 2001.

[5] Meena Rambocas and Joo Gama, "The Role of Sentiment Analysis,"*FEP Economics and Managment,* 2013.

[6] Bing Liu, " Opinions, Sentiment, and Emotion in Text,"*Cambridge University,* 2015.

[7] Dominique Boullier et Audrey Lohard, "Opinion mining et Sentiment analysis: Méthodes et outils,"2012.

[8] Faiza Belbachir, " Expérimentation de fonctions pour la détection d'opinions dans les blogs,"*Université de Paul Sabatier, Institut de Recherche en Informatique de Toulouse* 2010.

[9] Alexander Pak and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *Université de Paris-Sud, Laboratoire LIMSI- CNRS,France* 2010.

[10] Pang and Lee, " Opinion Mining and Sentiment Analysis,"*Now Publishers Inc,* 2008.

[11] Arti Buche, Dr. M. B. Chandak and Akshay Zadgaonkar, "Opinion mining and analysis:a survey," *International Journal on Natural Language Computing,* India 2013.

[12] G.Vinodhini and RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey,"*International Journal of Advanced Research in Computer Science and Software Engineering***,** India 2012.

[13] Vivek Kumar Singh and Debanjan Mahata, " A clustering and opinion mining approach to socio-political analysis of the blogosphere," *Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on* 2010

[14] Matthew Eric Glassman, Jacob R. Straus and Colleen J. Shogan, "Social Net- working and Constituent Communications: Members Use of Twitter and Facebook Dur- ing a Two-Month Period in the 112th Congress,"*Congressional Research Service,* 2009.

[15] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passon- neau, "Sentiment analysis of Twitter data," *LSM 11 Proceedings of the Workshop on Languages in Social Media,*2011.

[16] Laurent Dijoux, "Boostez votre business avec Twitter,"*Almabic,* 2009.

[17] Fred Colantonio, "Communication professionnelle en ligne: comprendre et exploiter les médias et réseaux sociaux,"*Edipro,* 2011.

[18] Tim O'Reilly and Sarah Milstein," The Twitter Book," *Angham B3 2PB, UK***,** 2

[19] Soumia Elyakoute HERMA et Khadidja SAIFIA, "Analyse des sentiments cas Twitter", Université de Ghardaia, 2015.

[20] Crannell, W. C., Clark, E., Jones, C., James, T. A., & Moore, J., 2016. A pattern-matched twitter analysis of US cancer-patient sentiments. Journal of Surgical Research, 206(2), 536-542.

[21] Cheong, Marc, and Vincent CS Lee, 2011. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. Information Systems Frontiers 13.1: 45-59.

[22] Neppalli, V.K., Caragea, C., Squicciarini, A., Tapia, A. & Stehle, S., 2017. Sentiment analysis during Hurricane Sandy in emergency response. International Journal of Disaster Risk Reduction, 21(2017), 213-222.

[23] Nazan, Serkan, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telematics and Informatics*,2018.

[24] Al-Twairesh, Al-Khalifa et Al-Salman, "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets," *Procedia Computer Science*,2017.

[25] A. Shoukry and A. Rafea, "Preprocessing Egyptian Dialect Tweets for Sentiment Mining," *Fourth Work. Comput. …*, pp. 47–56, 2012.

[26] A. El Abdouli, L. Hassouni, and H. Anoun, "Sentiment Analysis of Moroccan Tweets using Naive Bayes Algorithm," vol. 15, no. 12, 2017.

[27] Alomari, ElSherif and Shaalan, "Arabic Tweets Sentimental Analysis Using Machine Learning," *J. Chem. Theory Comput*, 2017.

[28] Wang, Can, Kazemzadeh and Bar, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle," *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 2017.

[29] J. Weng and B.-S. Lee, "Event Detection in Twitter," *5th Int. AAAI Conf. Weblogs Soc. Media*, pp. 401–408, 2011.

[30] Sakaki, Okazaki, and Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," *Proceedings of the 19th international conference on World wide web*, 851–860. New York, NY, USA: ACM, 2010.

[31] A. P. Kirilenko and S. O. Stepchenkova, "Public microblogging on climate change: One year of Twitter worldwide," *Glob. Environ. Chang.*, vol. 26, pp. 171–182, 2014.

[32] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, "PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis,"

[33] Hiroshi Ogura, Hiromi Amano, and Masato Kondo. "Distinctive characteristics of a metric using deviations from poisson for feature selection". Expert Systems with Applications, 37(3) :2273-2281, 2010.

[34] Gerard Salton, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing". Communications of the ACM, 18(11) :613-620, 1975.

[35] Olena,Yurii et Hanna,"Analysis of Propaganda Elements Detecting Algorithms in Text Data"Hu,Ivan on Advances in Computer Science for Engineering and Education II [En ligne]. Disponible : https://books.google.co.ma/books?id=3mqPDwAAQBAJ

[36] Houssem Eddine Dridi, "Analyse des données de microblogs", Université de Montréal,2012.

[37] Alwakid, Ghadah, Taha Osman, and Thomas Hughes-Roberts, "Challenges in Sentiment Analysis for Arabic Social Networks", *Procedia Computer Science* 117: 89–100. [En ligne]. Disponible: https://sciencedirect.com/science/article/pii/S1877050917321543

[38] Ennaji, Makhoukh, Es-saiydy, Moubtassime,Slaoui, "A grammar of Moroccan arabic", Publications of the Faculty of Letters Dhar El Mehraz, Fès 2004.

[39] Duwairi R., Al-Refai M. "Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization", *4th Int. Conf. on Innovations in Information Technology* IIT'07. N. (2007).

[40] Boudad, Naaima, Rdouan Faizi, Rachid Oulad Haj Thami, and Raddouane Chiheb, "Sentiment Analysis in Arabic: A Review of the Literature", *Ain Shams Engineering Journal* 9(4): 2479–90. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S2090447917300862

[41] Kheireddine Abainia, Siham Ouamour & Halim Sayoud, "A novel robust Arabic light stemmer", *Journal of Experimental & Theoretical Artificial Intelligence* [En ligne]. Disponible : http://dx.doi.org/10.1080/0952813X.2016.1212100

[42] Hiroshi Ogura, Hiromi Amano, and Masato Kondo, " Distinctive characteristics of a rnetric using deviations from poisson for feature selection" *Expert Systems with Applications,* 37(3) :2273-2281, 2010.

[43] Colah, "Understanding LSTM Networks", [En ligne]. Disponible : http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[44] Bijoyan, "Hands On Natural Language Processing (NLP) using Python", [En ligne]. Disponible : https://udemy.com/hands-on-natural-language-processing-using-python/

[45] "Flask framework", [En ligne]. Disponible : http://flask.pocoo.org/docs/1.0/

[46] "Embeddings: A Matrix of Meaning", [En ligne]. Disponible : https://medium.com/@Petuum/embeddings-a-matrix-of-meaning-4de877c9aa27

[47] "Support Vector Machines", [En ligne]. Disponible: https://scikit-learn.org/stable/modules/svm

[48] "Logistic Regression",[En ligne].Disponible: https://scikit-learn.org/stable/modules/generated/ LogisticRegression

[49] "Keras: The Python Deep Learning library", [En ligne]. Disponible : https://keras.io/

[50] "nltk.stem package",[En ligne].Disponible : https://www.nltk.org/api/nltk.stem.html

[51] "Twitter Developer Platform", [En ligne]. Disponible : https://developer.twitter.com/

[52] "The Basics of Sentiment Analysis", [En ligne]. Disponible : https://monkeylearn.com/sentiment-analysis/

[53] "Implementing a CNN for Text Classification in TensorFlow", [En ligne]. Disponible: http://wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/

[54] M. Mataoui, "A Proposed Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic.," *Res. Comput. Sci.*, vol. 110, pp. 55–70, 2016.

[55] Nawaf A. Abdulla, Nizar A. Ahmed, Mohammed A. Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sen- timent analysis: Lexicon-based and corpus-based. In Applied Electrical Engineering and Computing Tech- nologies (AEECT), IEEE.

[56] Muhammad Abdul-Mageed. 2015. Subjectivity and sentiment analysis of Arabic as a morophologically- rich language. Ph.D. thesis, INDIANA UNIVER- SITY.

[57] El-Masri, Mazen, Nabeela Altrabsheh, Hanady Mansour, and Allan RaASy. 2017. "A Web-Based Tool for Arabic Sentiment Analysis." *Procedia Computer Science* 117: 38–45.

[58] Al-Horaibi, Lamia, and Muhammad Badruddin Khan. 2016. "Sentiment Analysis of Arabic Tweets Using Text Mining Techniques." In eds. Xudong Jiang, Guojian Chen, Genci Capi, and Chiharu Ishll. , 100111F.

[59] Nabil, Mahmoud, Mohamed Aly, and Amir Atiya. 2015. "ASTD: Arabic Sentiment Tweets Dataset." (September): 2515–19.

[60] Ayyoub, Mahmoud Al, Safa Bani Essa, and Izzat Alsmadi. 2015. "Lexicon-Based Sentiment Analysis of Arabic Tweets." International Journal of Social Network Mining 2(2): 101.

[61] El-Beltagy, Samhaa R., Mona El Kalamawy, and Abu Bakr Soliman. 2017. "NileTMRG at SemEval-2017 Task 4: Arabic Sentiment Analysis." (Kim 2014): 790–95.

[62] Heikal, Maha, Marwan Torki, and Nagwa El-Makky. 2018. "Sentiment Analysis of Arabic Tweets Using Deep Learning." *Procedia Computer Science* 142: 114–22.