N° d'ordre :
N° de série :

République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique

UNIVERSITÉ ECHAHID HAMMA LAKHDAR D'EL OUED

FACULTÉ DES SCIENCES EXACTES

DEPARTEMENT D'INFORMATIQUE

Mémoire de fin d'étude

MASTER ACADEMIQUE

Domaine : Mathématiques et Informatique
Filière : Informatique
Spécialité : Systèmes Distribués et Intelligence Artificiel (SDIA)

Thème

Analyse des sentiments dans les réseaux sociaux en dialecte algérien

Préparé par :

- KARA Mostefa
- MELOUK Abdelbasset

Soutenu devant le jury composé de

M. BEGGAS Mounir	Président	Univ. d'El Oued
M. OTHMANI Samir	Examinateur	Univ. d'El Oued
Dr. **LEJDEL Brahim** MCA	Encadreur	Univ. d'El Oued

Année universitaire : 2018 – 2019.

Remerciements

Tous d’abord, nous tenons à remercier le bon Dieu de nous avoir accordé toute la détermination, la volonté et la force pour qu’on puisse réaliser ce modeste travail.

Nous remercions infiniment notre encadreur **Dr. LEJDEL Brahim** pour ses conseils, sa patience, sa disponibilité et son soutien tout au long de cette période.

Nous tenons à exprimer notre profonde gratitude et nos sincères remerciements aux :

Membres de jury d’avoir accepté de juger notre travail et de l’avoir enrichi.

Toutes les personnes qui nous ont aidés, de près ou de loin, en particulier :

Notre enseignant Dr. NAGOUDI El Moatez Bellah.

Nous adressons aussi nos remerciements à tous nos enseignants qui ont veillé sur notre formation.

Résumé

Aujourd'hui, l'analyse des textes a une grande importance surtout dans les domaines comme politiques, productions et services…etc. Actuellement, les réseaux sociaux pleins des textes dans lesquelles, les internautes s'expriment en différents sujets, l'intérêt de leurs opinions est considérable, où la compréhension du contenu véhiculé par ces textes est un élément essentiel. On peut dire que le bon gestionnaire ce qui écoute bien aux opinions des citoyens. Dans ce sens-là, l'analyse de sentiments est très importante pour satisfaire les besoins des citoyens. Dans ce travail, nous allons utiliser quatre algorithmes qui permettent d'analyser et classifier un ensemble de publications dérivées des réseaux sociaux. Les classes que nous avons définies sont : la classe positive, négative ou neutre.

A Notre savoir ce travail est parmi les premiers travaux qui utilisent et comparent plusieurs algorithmes de classification des commentaires en Facebook, en utilisant le dialecte algérien.

Mots clés : fouille d'opinions, analyse des sentiments, fouille de texte, détection émotionnelle, web social, corpus annoté, Lexique de sentiments.

Abstract

Today, text analysis has a great importance, especially in areas such as politics, productions and services...etc. Currently, the social networks full of the texts in which, the Internet-users express in different subjects, the interest of their opinions is considerable, where the comprehension of the content conveyed by these texts is an essential element. We can say that the good manager who listens well to the opinions of citizens. In this sense, the Sentiment Analysis is very important to meet the needs of citizens. In this work, we will use four algorithms that analyze and classify a set of publications derived from social networks. The classes that we have defined are: positive, negative or neutral. To our knowledge this work is among the first one that uses and compares several comment classification algorithms in Facebook, using Algerian dialect.

Keywords: opinion mining, Sentiment Analysis, text mining, emotional detection, social web, annotated corpus, Lexicon of Sentiment.

ملخص

في هذه الايام، يكتسي تحليل النصوص أهمية كبيرة، خاصة في مجالات مثل السياسة والإنتاج والخدمات. كذلك نلاحظ أن الشبكات الاجتماعية مليئة بالنصوص التي يعبر فيها متصفحي الانترنت عن أنفسهم في مواضيع مختلفة، وفي آرائهم أهمية كبيرًا، حيث يكون فهم المحتوى الذي تنقله هذه النصوص عنصرًا أساسيًا. يمكننا القول إن المسير الجيد هو الذي يستمع جيدًا لآراء المواطنين. بهذا المعنى، فإن تحليل المشاعر مهم جدًا لتلبية احتياجاتهم. في هذا العمل سنستعمل أربع خوارزميات لتحليل وتصنيف مجموعة من المنشورات المستمدة من الشبكات الاجتماعية.

الفئات التي حددناها هي: إيجابية، سلبية أو محايدة. على حد علمنا هذا العمل يعتبر من بين الاعمال الأوائل الذي يوظف ويقارن مجموعة من خوارزميات تصنيف التعليقات في الفايسبوك باللهجة الجزائرية.

الكلمات المفتاحية: التنقيب في الآراء، تحليل المشاعر، التنقيب في النصوص، الكشف عن العواطف، الشبكات الاجتماعية، المعجم المعلق، قاموس المشاعر.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Actuellement, notre vie se base sur l'information et son analyse. Cette information est plus disponible à nos jours et plus précisément sous forme numérique, avec le développement du Web 2.0. De plus en plus, les gens communiquent, partagent de contenu et expriment leur avis sur internet à propos d'un grand nombre de sujets, dans les groupes de discussion, les blogs, les forums et autres sites concernant les critiques de produits.

En fin de 2013, Facebook a ouvert ses pages à la recommandation clients, les produits choisis peuvent être évalués par leurs fans avec l'un des valeurs de 1 à 5 où 1 est une polarité très négative, et 5 est une polarité très positive **[1]**, les opinions disponibles sur l'internet ont un impact considérable sur les internautes, des sondages montrent que la plupart des utilisateurs (80%) ont déjà fait des recherches d'avis sur un produit ou un service et que ces derniers payent deux fois plus cher pour un produit où son avis est plus affirmatif qu'un autre **[2]**, les entreprises prennent en compte ces paramètres et ils savent que l'analyse d'opinion est une composante importante pour la prise de décision.

Selon une étude faite par Pew Research Center, 20% des utilisateurs de réseaux sociaux ont modifié leur avis politique sur un sujet à cause de ce qu'ils ont vu sur les médias sociaux **[3]**. Nous pouvons voir l'utilité de la détection d'opinion dans les domaines de de marketing, politique, Etudes psychologiques, santé, sécurité routière, tourisme...

L'objective de ce travail consiste à étudier les opinions publiques qui se concernent essentiellement les sentiments trouvés dans les commentaires Facebook écrits en arabe et plus précisément en dialecte algérien. Avant de détailler les différents chapitres de ce mémoire, nous présentons une nomenclature des notions qui sont nécessaires pour comprendre le travail effectué.

Notre mémoire est divisé principalement en quatre chapitres.
Dans le premier chapitre, nous focalisons sur l'état de l'art de l'analyse des sentiments, notamment les travaux inhérents aux dialectes. Le second chapitre est consacré aux particularités de la langue arabe. Dans le troisième chapitre, nous présentons en détail la modélisation de notre système. Le quatrième chapitre présente l'expérimentions et la discussion des résultats obtenus. Enfin, nous mettons une conclusion et quelques perspectives.

Nomenclature des notions

Dans cette section, nous allons définir une nomenclature des concepts utilisés dans ce mémoire pour faire l'éclaircir avant de détailler notre travail.

Sentiment : Expression de la sensibilité, d'un penchant, d'une affection, d'une passion. Synonymes : avis, opinion **[4]**.

Opinion : Jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense **[5]**.

Réseau social : L'ensemble des sites internet qui permettent de relier des personnes physiques ou personnes morales entre elles pour partager des textes, photos, vidéos... **[6]**.

Lexique : Est un ensemble de mots d'une langue (ou dialecte), généralement est un dictionnaire spécialisé dans un domaine particulier **[7]**.

Corpus : Un corpus est un ensemble fini de textes préparé pour but d'analyse, spécifiquement, collectivement fini d'énoncés considérés comme particularité du type de langue à exploiter **[8]**.

Apprentissage automatique (Machine Learning) : Est un type d'intelligence artificielle qui donne aux ordinateurs la capacité d'apprendre sans être programmés et de trouver des solutions aux problèmes qui ne sont pas résolus par des algorithmes classiques **[9]**.

Traitement automatique des langues naturelles : C'est de faire parler les ordinateurs, et plus précisément, à leur donner les capacités linguistiques d'un être humain **[10]**.

Machine à vecteurs de support (**SVM** Support Vector Machines) **:** Sont des techniques d'apprentissage statistique proposées par V. Vapnik en 1995 **[11]**. Elles permettent de résoudre des problèmes tel que la régression, la fusion, … et d'une part particulière le classement. Sa principale idée consiste à projeter l'espace d'entrée (données à deux classes différentes non séparables de façon linéaire) dans un espace à grande dimension (espace de caractéristiques) où les données deviennent linéairement séparables, dans ce nouvel espace, la technique de construction de l'hyperplan optimal est utilisée pour séparer les deux classes par une fonction de classement **[12]**.

Arbre de décision (**DT** Decision Tree) **:** Un arbre de décision est un graphe non orienté, acyclique et connexe, son ensemble de nœuds est devisé en trois catégories, nœud racine, nœuds internes (enfants) et nœuds terminaux (feuilles), Les arbres de décision (AD) sont utilisés dans l'exploration de données pour prendre une décision **[13]**.

Forêt d'Arbres Décisionnels (**RF** Random Forest) **:** Le terme Forêts Aléatoires est un ensemble d'arbres de décision. Où chaque arbre a pour but de décomposer le problème en suite de tests correspondant à partitionner l'espace de données en sous-régions homogènes (en terme de classe), aller de la racine à une feuille en effectuent les tests de nœuds où la classe d'une feuille est la majoritaire parmi les exemples d'apprentissage appartenant à cette feuille **[14]**.

Naïf Bayésien (**NB** Naive Bayes) **:** Un classificateur naïf de Bayes est un classificateur probabiliste basé sur l'application du théorème de Bayes **[15]** avec l'hypothèse naïve (forte indépendance), c'est-à-dire que les variables explicatives (Xi) sont supposées indépendantes conditionnellement à la variable cible (C), il appartient à la famille des classificateurs linéaires (son rôle est de classer dans des classes les échantillons qui ont des propriétés similaires). Ce classificateur est souvent utilisé sur les flux de données pour la classification supervisée **[16]**.

Mesures de performance [17] : Soit : PS : documents pertinents sélectionnés,

PN : documents pertinents non sélectionnés, NS : documents non pertinents sélectionnés, NN : documents non pertinents non sélectionnés.

Nous avons : **Rappel** (recall) = PS / (PS + PN)

Précision = PS/(PS + NS), **F-mesure** (F-score) = 2*précision * rappel/(précision + rappel)

Accuracy = (PS + NN) / (PS + NS + PN + NN)

N-grammes : Comme leur nom l'indique, les *n-grams* sont des suites d'un nombre donné de caractères (éventuellement de mots) **[18]**. À titre d'exemple, le bigramme le plus fréquent de la langue française est « de ».

Fonctionnalité : Une caractéristique ou une propriété qui nous permet de prédire l'appartenance d'un élément à une classe donnée **[19]**.

MSA : Modern Standard Arabic, qui est enseignée dans les écoles contemporaines **[20]**.

Chapitre 1

Etat de l’art

Chapitre 1

1 Etat de l'art

1.1 Introduction

L'analyse de sentiments, aussi parfois désignée sous le nom fouille d'opinions, est un sous-domaine de l'informatique, il est considéré comme une partie du traitement automatique du langage naturel et a pour but de classifier les sentiments exprimés dans des textes, Comment on peut utiliser les ordinateurs pour mieux comprendre le langage naturel ?

Dans ce contexte, plusieurs travaux sont réalisés en tous les domaines d'application connus et avec différents sous-objectifs (construction de corpus, détection d'opinion, comparaison de fonctionnalités, application des méthodes…etc.), dans son ouvrage, Banfield (1982) a mis en évidence les éléments du langage qui conduisent à la subjectivité d'un texte ou tout du moins d'une partie d'un texte **[21]**. Hatzivassiloglou et MckKeown (1997) sont les premiers à se pencher sur la classification des opinions **[22]**. Pang (2002) a fait une étude qui permettait de classifier les sentiments de commentaires de films **[23]**, il est le premier qui a expérimenté l'apprentissage automatique **[24]**.

Dans notre étude, nous allons focaliser sur sept projets plus récents.

1.2 Identification de la polarité

Un point de vue est considéré comme positive, négative ou une nuance de ces extrêmes, on spécifie par la polarité la position d'une opinion sur cet axe comprenant des différents niveaux de positivité et de négativité, donc la polarité peut être des catégories telle que positive, négative et neutre, ou bien elle peut être définie par des valeurs décrivent le degré de positivité (ou négativité), par exemple une valeur entre 1 et 5, où 1 désigne une polarité très négative, et 5 désigne une polarité très positive, elle prend aussi le nom « orientation sémantique » **[25]**.

Cette considération présuppose que le texte est homogène, il est soit totalement positif, soit totalement négatif (éventuellement neutre) ça dépend de la cible étudiée, à partir de cette considération, nous avons deux vues :

1.2.1 Opinions de polarité unique

Cette vue prend en considération l'opinion globale exprimée, d'où le texte est associé à une idée unique et il ne forme pas un mélange d'opinions différentes (hétérogènes), cette considération est d'autant plus valide que le document est court, par exemple, une critique de produit ou de service **[25]**.

1.2.2 Opinions basées sur les aspects

La désignation de la polarité d'une opinion basée sur des aspects pour une cible donnée, elle consiste à identifier la polarité de chaque un de ces aspects, néanmoins, les aspects d'une cible ne sont pas généralement connus préalablement, et ils varient d'une cible à une autre. Par exemple, les aspects d'une machine à laver sont : sa rapidité, sa consommation d'électricité, sa quantité à laver, etc..., alors que les aspects d'un ordinateur sont : son poids, sa vitesse, son espace de stockage, sa durée de vie, ...etc., il y a des travaux séparent l'opération en deux étapes, la première est l'extraction des aspects dans un premier temps, puis l'identification de l'orientation de chaque aspect, des autres travaux proposent à modéliser ces deux étapes conjointement **[25]**.

1.3 Approches de l'Analyse des Sentiments

D'une part, il doit expliquer la différence entre deux méthodes d'apprentissage ; l'apprentissage supervisé et celui-ci non supervisé. La première implique la présence de deux ensembles de données, un ensemble d'entrainement et un ensemble de test, La méthode s'appelle supervisée puisque le système est entraîné sur un sous-ensemble d'apprentissage qui contient des modèles déjà traités (dans notre étude, les modèles sont des commentaires Facebook).

La deuxième méthode (non supervisée) ne recommande qu'un seul ensemble de données, cette méthode exige que le système de façon autonome doit restructurer les informations au sein de l'ensemble de données en les regrouper en sous-ensembles, et il doit le réorganiser d'une façon que les données les plus similaires soient dans le même groupe.

D'autre part nous distinguons trois approches de fouille d'opinion, basé sur lexique, basé sur corpus et l'approche mixte **[26]**.

1.3.1 Approches basée sur lexique

Appelé aussi symbolique ou linguistique, jusqu'à maintenant, la plupart des études de l'analyse des sentiments se sont basées sur cette méthode. Elle permet d'identifier la

polarité d'un texte à l'utilisation de deux ensembles de mots, ceux qui expriment un sentiment positif et ceux qui expriment un sentiment négatif.

Le modèle compte dans le texte le nombre de mots positifs et le nombre de mots négatifs, la somme donne une évaluation globale du sentiment de texte, si le nombre de mots positifs l'emporte sur celle de mots négatifs, le texte considéré comme positif, inversement, le texte est considéré comme négatif, éventuellement neutre si les nombres sont égaux **[26]**.
Exemple d'algorithme : Support Victor Machine (SVM)…

1.3.2 Approches basée sur corpus

Appelé aussi statistique, inversement à la méthode précédente, l'analyse automatique des sentiments basé sur corpus n'a pas besoin d'un lexique de mots positifs et négatifs, elle a besoin de deux corpus annotés (éventuellement un seul, si on utilise l'apprentissage non supervisé), le premier corpus destiné à l'apprentissage (l'entraînement), à partir les annotations, le modèle automatiquement sera capable de faire une analyse équivalente et d'une manière autonome.

Le deuxième corpus destiné au test, son rôle est de vérifier la performance du modèle, si le cas est idéal, le résultat d'analyse devrait être cent pour cent avec ceux du premier corpus. Pour améliorer la performance du modèle, il est important que le corpus d'entrainement doive être représentatif pour le corpus d'évaluation **[26]**. Exemple d'algorithme : réseau de neurone...

1.3.3 Approches hybrides

Cette approche est appelée aussi classification semi-supervisée, elle combine les points forts de deux approches précédentes, il y a trois façon de faire. La première est d'exploiter les outils linguistiques pour élaborer le corpus puis classer les textes par un outil d'apprentissage supervisé. La deuxième façon est d'utiliser l'apprentissage automatique pour établir le corpus d'opinion nécessaire à l'approche basée sur lexique. La troisième façon est le conjointement des deux approches précédentes et la combinaison de leurs résultats soit par un système de vote soit par un algorithme d'apprentissage **[27]**. Exemple d'algorithme : S3VM (Semi-Supervised Support Vector Machine).

1.4 Difficultés de l'analyse de sentiments

L'extraction du sentiment consiste à déterminer la polarité d'une opinion, elle est en général peut être positive, négative ou neutre, par la suite, nous citons quelques difficultés de cette procédure :

- Ambiguïté de quelques mots, sont-ils positifs ou négatifs ! Selon le contexte (exemple : المكتوب) **[28]**.
- Difficulté à cause de la structure syntaxique ou sémantique d'un texte, soit l'exemple "le téléphone est bon, mais sa couleur n'est pas bon", il y a ici dans le texte deux polarités opposées **[28]**.
- Détermination la partie de la phrase qui a le poids le plus important, par exemple "le produit est beau, joli ..., en fin je ne l'accepte pas", dans ce cas-là l'opinion de la deuxième partie de la phrase est la plus importante **[29]**.
- Difficulté due au langage naturel, plus particulièrement les paroles (ou les mots) qui portent le soupçon (doute), par exemple : je crois que cet homme est gentil **[29]**.
- Difficulté due à l'analyse où la distribution des paquets de mots fait une influence sur la polarité du texte, soit les deux phrases suivantes (elles contiennent les mêmes paquets de mots sans exprimer les mêmes sentiments) : "Je l'ai apprécié pas seulement à cause de ... ", " Je l'ai pas apprécié seulement à cause de ... ", la première est positive alors que la deuxième est négative **[30]**.
- Le choix de l'approche à appliquer et le lexique (le dictionnaire) à utiliser **[30]**.
- L'adaptation d'un domaine à un autre, par exemple si on a un corpus multi domaine, la création d'un dictionnaire pour lui est plus difficile quand on a un corpus d'un seul domaine **[31]**.

1.5 Travaux connexes

Dans ce contexte il y a deux genres de travaux, le premier et qui nous intéresse pas c'est que de faire un outil d'analyse a comme entrée : un tweet, et comme sortie : la polarité de ce tweet, par exemple l'application twitter sentiment (Sentiment140) **[32]**, outil en ligne, gratuit, dans laquelle, les auteurs testent trois algorithmes qui montrent des résultats entre 80% et 83% de réussite **[28]**.

Le deuxième genre c'est de faire un modèle d'analyse qui nous donne l'annotation d'un corpus (ou partie de corpus) selon l'une des trois approches définies en 1.3.

1.5.1 Analyse de sentiments - cas de MSA

Le travail choisis dans le cas de MSA est de Mohammed et al. (2014), ils ont visé s'accentuer sur l'économie et plus précisément sur les revues de produits **[33]**. Ils ont commencé par la première phase (collection de données), en suite le prétraitement passant à la classification, leur corpus était collecté par eux-mêmes à partir de plusieurs ressources web comme reviewzat.com et jawal123.com...etc., sous forme d'un ensemble de documents textes, chaque document est un produit représenté par son type, ils ont sélectionné cinq types de produits qui forment ce corpus, les types sont Caméra, PC portable, Téléphone portable, Tablet, Télévision. Le corpus contient 250 documents, 2812 phrases et 15466 mots.

Deux évaluateurs ont travaillé sur l'étiquetage des opinions, le premier est un expert pour l'évaluation de produits et le deuxième est un spécialiste en langue arabe et un troisième non spécialiste a été utilisé juste pour valider les choix de deux autres annotateurs pour avoir un certain degré de fiabilité d'annotation.

Ils ont trouvé quelques difficultés, parmi eux :

a) Les expressions émotionnelle (La joie, le malheureux, la surprise…), pour cela ils ont développé un petit convertisseur (symbole vers mot) qui fonctionne comme il est montré dans le tableau 1.1.

Émoticône	Symbole	Commentaire	Commentaire après conversion de l'icône
🙂	:)	ها تف روعة :)	هاتف روعة، سعيد
😊	^_^	جهاز ممتاز فاق كل التوقعات^_^	جهاز ممتاز فاق كل التوقعات، مستمتع
😐	:(	الكاميرا سيئة للغاية :(	الكاميرا سيئة للغاية، حزين
😕	:/	دقة الكاميرا دون المتوقع/ :	دقة الكاميرا دون المتوقع، خاب أملي

Tableau 1.1 : Exemples de conversions à partir d'un symbole vers un mot.

b) L'élimination de la redondance des caractères tout en conservant la signification des mots, exemple جمييييل→جميل.

c) Les mots étrangères ال mobile ممتاز.

Dans la classification, ils ont commencé par le stemming, c'est le processus de suppression de tous les préfixes et les suffixes d'un mot pour produire le stem ou la racine, ce processus

est difficile en arabe car, par exemple le stemming des deux mots رائع merveilleux et مروع terrible donne le mot روع horreur, lors que ces deux mots ont des polarités inversées.

Ils ont fait leurs tests avec trois algorithmes de classification, qui sont le Support Vector Machines (SVM), Naïve Bayes (NB) et K-plus proche voisin (KNN).

Pour réaliser les tests de performances de leur système, ils ont utilisé deux techniques la validation croisée et le pourcentage scission.

a) **La validation croisée** : Il doit découper l'ensemble de données en K groupes tirés aléatoirement qui font les ensembles de test **[34]**.

Dans ce cas, le nombre K est égal à 10, c'est-à-dire l'ensemble d'apprentissage est coupé en 10 groupes, le modèle va effectuer l'apprentissage dix fois sur neuf parties d'entrainement et sera évalué sur la dixième, les dix évaluations sont alors combinées, Le résultat obtenu est montré sur le tableau 1.2.

	En termes de précision			En termes de rappel		
Corpus / Classificateur	KPPV	SVM	NB	KPPV	SVM	NB
Corpus à l'état brut	0,712	0,886	0,834	0,693	0,885	0,828
Corpus + light stemmer	**0,76**	0,904	0,861	**0,705**	**0,902**	0,857
Corpus + khoja stemmer	**0,76**	0,904	0,861	**0,705**	**0,902**	0,857
Corpus + normalisation	0,618	0,885	0,871	0,607	0,877	0,869
Corpus + normalisation + khoja stemmer	0,58	**0,912**	**0,876**	0,578	0,898	**0,873**
Corpus + normalisation + light stemmer	0,58	**0,912**	**0,876**	0,578	0,898	**0,873**

Tableau 1.2 : Les résultats de classification par la méthode d'évaluation : validation croisée.

Ils ont conclu qu'il y'a un classificateur qui a donné des performances moins bonnes que les deux autres, il est le classificateur k-plus proches voisins (KPPV), le SVM est le plus efficace avec les différents types de combinaison de données.

b) **Pourcentage scission** : Dans cette méthode, le corpus est divisé aléatoirement en deux ensembles de données disjoints, le premier ensemble c'est l'ensemble d'entrainement et le deuxième ensemble c'est l'ensemble d'évaluation, dans leur cas, 80% des données sont pour l'ensemble d'apprentissage et le reste (20%) c'est pour l'ensemble de test, Le résultat obtenu est montré sur le tableau 1.3.

	En termes de précision			En termes de rappel		
Corpus / Classificateur	KPPV	SVM	NB	KPPV	SVM	NB
Corpus à l'état brut	**0,803**	**0,946**	0,822	**0,776**	**0,939**	0,816
Corpus + light stemmer	0,799	**0,946**	0,881	0,735	**0,939**	0,878
Corpus + khoja stemmer	0,799	**0,946**	0,881	0,735	**0,939**	0,878
Corpus + normalisation	0,788	0,899	0,922	0,714	0,898	0,918
Corpus + normalisation + khoja stemmer	0,801	0,93	**0,946**	0,653	0,918	**0,939**
Corpus + normalisation + light stemmer	0,801	0,93	**0,946**	0,653	0,918	**0,939**

Tableau 1.3 : Les résultats de classification par apprentissage.

Cette fois le SVM n'est pas toujours le plus efficace (juste sur le corpus a l'état brute), dans d'autres cas le NB était le plus efficace, il a atteint 0,946 de précision et 0,939 de rappel.

1.5.2 Analyse de sentiment - cas de dialecte tunisien

Avec le dialecte tunisien on va jeter un coup d'œil sur le travail de Salima mdhaffer et al. (2017). Leurs principales contributions sont : Une enquête sur les ressources disponibles pour la langue arabe SA (Sentiment Analysis) MSA et dialectique. La création d'un corpus de formation disponible gratuitement pour le dialecte tunisien et L'évaluation des performances du système dialectal tunisien SA sous plusieurs configurations **[35]**.

Leur corpus appelé TSAC, il est constitué de commentaires écrits sur les pages officielles des radios et des chaînes de télévision tunisiennes, à savoir Mosaïque FM, Jawhra FM, Shemes FM, HiwarElttounsi TV et Nessma TV au cours d'une période allant de janvier 2015 à juin 2016, voir le tableau 1.4.

	Positive	Négative	Total
Commentaires	8215	8845	17060

Tableau 1.4 : Statistiques de corpus TSAC.

Pour le test, ils ont utilisé aussi deux autres corpus différents. Les corpus sont : OCA (Corpus d'opinion en arabe) **[36]**, il contient 500 critiques de films en MSA, collectées sur des forums et des sites Web, il est divisé en 250 critiques positives et 250 critiques négatives, le corpus LABR (Critique de livre arabe à grande échelle) **[37]**, il est mixte, écrit en MSA et en différent dialectes arabe, il est librement disponible et contient plus de 63k commentaires. Le résultat présenté dans le tableau 1.5.

Classificateur	Entraînement	Positive		Négative		Taux d'erreur
		Précision	Rappel	Précision	Rappel	
SVM	MSA	0.44	0.15	0.49	0.80	0.52
	D Mix	0.50	0.84	0.52	0.17	0.49
	TUN	0.77	0.77	0.77	0.76	**0.23**
	MSA D Mix	0.51	0.90	0.60	015	0.47
	TUN MSA	0.74	0.83	0.80	0.71	**0.23**
	TUN D Mix	0.68	0.76	0.73	0.64	0.30
	ALL	0.71	0.81	0.78	0.66	0.26
BNB	MSA	0.43	0.28	0.46	0.62	0.55
	D Mix	0.51	0.94	0.58	0.09	0.49
	TUN	0.56	0.70	0.60	0.46	**0.42**
	MSA D Mix	0.51	0.98	0.67	0.5	0.49
	TUN MSA	0.55	0.77	0.62	037	0.43
	TUN D Mix	0.54	0.76	0.60	0.36	0.44
	ALL	0.54	0.82	0.62	0.30	0.44
MLP	MSA	0.52	0.40	0.51	0.64	0.48
	D Mix	0.51	0.75	0.53	0.28	0.49
	TUN	0.78	0.78	0.78	0.78	**0.22**
	MSA D Mix	0.53	0.49	0.52	0.56	0.47
	TUN MSA	0.76	0.78	0.77	0.76	0.23
	TUN D Mix	0.75	0.77	0.76	0.75	0.24
	ALL	0.74	0.77	0.76	0.73	0.25

Tableau 1.5 : Résultats d'expériences d'Analyse de Sentiment tunisien en utilisant divers classificateurs avec différents ensembles de tests.

Il remarquaient que : il y a un taux d'erreur de 0,23 avec SVM, et 0,22 avec MLP (Multi-Layer Perceptron) et 0,42 avec NB, comme indiqué dans le tableau 1.5, SVM et MLP obtenaient des résultats similaires pour toutes les configurations expérimentales, cependant, des résultats plus faibles étaient obtenus avec le classificateur NB, aucune amélioration lorsque les systèmes SA sont formés avec des données de formation supplémentaires provenant de LABR (Critique de livre arabe à grande échelle) et OCA (Corpus d'opinion en arabe), globalement, des résultats plus médiocres étaient obtenus lorsque les systèmes d'analyse de sentiment sont formés sans le corpus TSAC, ceci est principalement dû aux quelques points parmi eux:

- Les ensembles de données OCA et LABR sont limités à un domaine (films et livres, respectivement), tandis que l'ensemble d'évaluation est multi-domaine.

- Les différences lexicales entre dialecte tunisien, MSA et autres dialectes.

1.5.3 Analyse de sentiment - cas de dialecte marocain

Sur le dialecte marocain, Abdeljalil Elouardighi et al. (2018) ont fait leur travail. Ils intéressaient dans lequel à l'analyse des sentiments à partir des commentaires Facebook **[38]**. Les principales contributions dans ce travail sont :

- Citer les propriétés de la langue ASM et l'ADM (Arabe Dialectal Marocain) et leurs défis pour l'AS (Analyse de Sentiment),

- Proposer un ensemble de techniques de prétraitement des commentaires Facebook écrits en ASM et en ADM pour l'AS.

La Figure 1.1 présente le processus proposé pour l'analyse des sentiments. Ce processus est composé de quatre phases :

Figure 1.1 : Etapes du processus proposé pour l'analyse des sentiments.

L'explication de ces phases :

a) Collection des commentaires Facebook :

Ils ont ciblé des journaux marocains qui publiaient en ligne des commentaires sur les élections législatives marocaines ayant eu lieu le 07 octobre 2016, la collecte a été effectuée, en utilisant "Facebook Graph API" **[39]**. Pendant 70 jours ils ont sélectionné 10254 commentaires.

b) Traitement des commentaires Facebook :

L'objectif étant d'extraire des variables (sous forme de mots ou de séquences de mots) pour les utiliser dans la classification, nettoyage et une normalisation du texte : suppression des signes, des symboles, des lettres répétées, des mots vides ou des mots qui ne fournissent aucune information sur le sujet étudié, puis l'opération de tokénisation par laquelle le texte du commentaire est divisé en unités lexicales (tokens), Le tableau 1.6 présente un exemple de prétraitement d'un commentaire.

Tache	Résultat
Texte initial	الكلام الي كي يقولو هاذ السياسي ماشي معقوووول ! هههههههه # السياسة المغربية Les discours de ce policien ne sont pas raisonnables ! hahahaha # politique marocaine
Nettoyage	الكلام الي كي يقولو هاذ السياسي ماشي معقوووول هههههههه السياسة المغربية
Normalisation	الكلام الي كي يقولو هاذ السياسي ماشي معقول هه السياسة المغربية
Tokénisation	'الكلام' , 'الي' , 'كي' , 'يقولو' , 'هاذ' , 'السياسي' , 'ماشي' , 'معقول' , 'هه' , 'السياسة' , 'المغربية'
Suppression des mots vides	'الكلام' , 'يقولو' , 'السياسي' , 'ماشي' , 'معقول' , 'هه' , 'السياسة' , 'المغربية'
Désuffixation	'كلام' , 'قول' , 'سياس' , 'ماش' , 'معقول' , 'هه' , 'سياس' , 'مغرب'

Tableau 1.6 : Exemple de prétraitement d'un commentaire.

c) Extraction et sélection de variables :

En fin, 6581 commentaires annotés 2908 négatifs et 3673 positifs, les variables d'entrée sont automatiquement extraites depuis le corpus formé à partir des commentaires prétraités, en utilisant les schémas d'extraction n-gram et de pondération TF/TF-IDF **[40]**.

d) Classification des commentaires :

Par l'application de trois algorithmes, Naïve Bayes (NB), les Forêts Aléatoires (FA) et les Machines à Vecteurs Support (SVM), 50% de Dataset pour l'apprentissage, 25% pour la validation et 25% pour le test, ils obtenaient le résultat présenté dans le Tableau 1.7 (TBC taux de bon classement).

Configurations	Classificateur	Nombre de variables sélectionnées	TBC avec les variables sélectionnées sur l'échantillon de validation	TBC avec les variables sélectionnées sur l'échantillon de test	TBC calculé en présence de toutes les variables
1-Unigram/TF	SVM	186	0.74	0.75	0.76
	NB	56	0.71	0.73	0.39
	FA	149	0.75	0.75	0.74
2 - Unigram/TF-IDF	SVM	198	0.77	**0.78**	**0.78**
	NB	55	0.72	0.72	0.42
	FA	56	0.76	0.76	0.75
3 - Bigram/TF	SVM	155	0.72	0.73	0.72
	NB	20	0.69	0.68	0.35
	FA	175	0.73	0.72	0.73
4 - Bigram/TF-IDF	SVM	199	0.72	0.72	0.72
	NB	20	0.67	0.67	0.36
	FA	198	0.72	0.73	**0.77**
5 -(Unigram+Bigram)/TF	SVM	200	0.76	0.77	0.76
	NB	100	**0.74**	**0.74**	0.39
	FA	89	0.76	0.76	0.71
6 - Unigram+Bigram)/ TF-IDF	SVM	199	0.77	0.77	**0.78**
	NB	50	0.72	0.71	0.56
	FA	148	0.76	0.75	0.73

Tableau 1.7 : Taux de bon classement pour les configurations testées avec sélection de variables.

La Figure 1.2 présente l'évolution des TBC (taux de bon classement) de la configuration [Unigram+Bigram/TF], pour les trois algorithmes appliqués, en fonction du nombre de variables insérées dans l'ordre décroissant de pertinence.

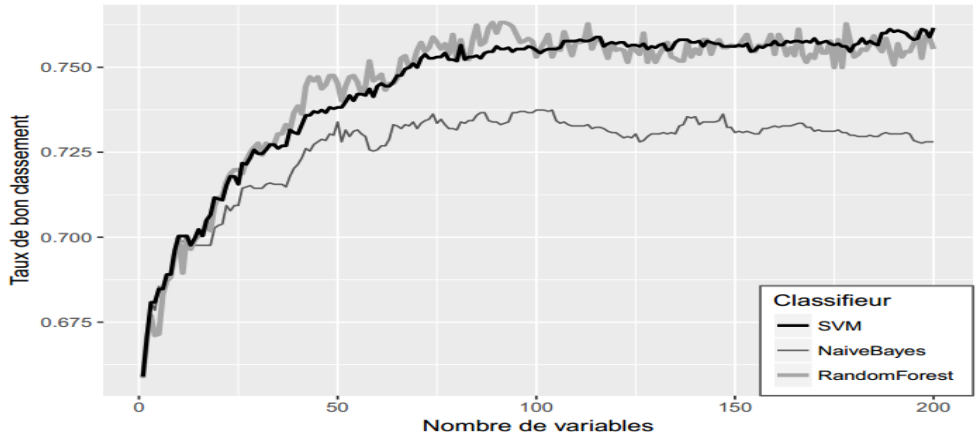

Figure 1.2 : Evolution du taux de bon classement en fonction des variables introduites pour la configuration : Unigram + Bigram / TF.

Ces résultats montrent que les meilleures performances ont été obtenues avec les combinaisons [Unigram/TF-IDF] et [Unigram + Bigram/TF-IDF] quel que soit l'algorithme utilisé. Ils prévoyaient de construire une base de commentaires plus importante et d'implémenter leur approche dans un environnement distribué et de développer une méthode d'annotation automatique des commentaires basée sur un lexique.

1.5.4 Analyse de sentiment - cas de dialecte saoudien

Le quatrième travail à présenter, c'est le travail de Nora Al-Twairesh et al. (2018). Ils ont développé une méthode hybride d'analyse de sentiments pour les tweets arabes en dialecte saoudien, leurs contributions sont : Présentation et évaluation d'un ensemble de fonctionnalités pour l'analyse de sentiments à l'aide d'une méthode de sélection en arrière, Développement et comparaison de trois modèles de classification pour l'AS (Analyse de Sentiments) des tweets saoudiens **[41]**.

Ils ont dit que selon Refaee et Rieser, (2014), les caractéristiques morphologiques aspect, sexe, humeur, personne, état et voix nuisent en réalité aux performances, entraînant une baisse de 21%. Et que La plupart des études sur Twitter SA (Sentiment Analysis) ont indiqué que l'utilisation d'un lexique des sentiments **[42]**. Ils ont rassemblé les caractéristiques utilisés dans les modèles de classification dans le tableau 1.8.

Fonctionnalité		valeurs
Sémantique	hasPositiveWordAraSenTi, hasNegativeWordAraSenTi hasPositiveWordMPQA, hasNegaitveWordMPQA hasPositiveWordLiu, hasNegaitveWordLiu, hasNegation hasIntensifier, hasDiminisher, hasModalWord, hasContrastWord	vrai, faux
	PositiveWordCount, NegativeWordCount, TweetScore	numérique
Stylistique	hasQuestionMark, hasExclamationMark, hasPositiveEmoticon hasNegativeEmoticon	vrai, faux
Tweet Spécifique	tweetLength	numérique

Tableau 1.8 : Caractéristiques utilisées dans le modèle de classification.

Dans ce travail, ils ont proposé une méthode hybride, combine une méthode à base lexique et une méthode à base corpus, la première méthode calcule un TweetScore (**Tweet score** = tweets positive value – tweets negative value) en utilisant les intensités des mots dans le lexique **[43]**, ce score est ensuite inclus en tant que caractéristique dans le modèle de classification. La deuxième méthode utilise un corpus de tweets saoudiens **[44]**. Ils

utilisaient le classificateur SVM car il a été signalé dans la majorité des études sur la SA (Sentiment Analysis) de tweets comme le classificateur le plus performant **[45]**. Ils utilisaient la bibliothèque libSVM **[46]**, à différents niveaux : classification à double voies (positif, négatif), classification à trois voies (positive, négative et neutre) et classification à quatre voies (positive, négative, neutre et mixte).

Les étiquettes utilisées sont décrites dans le tableau 1.9.

Étiquette	Signification	Étiquette	Signification
AF	all Features	AL	all-AraSentilexicon
TL	allTweetLength	WS	all-WordCount
QU	all-Question	ML	all-MPQA lexicon
EX	all-Exclamation	LL	all-Liu lexicon
EM	all-Emoticons	TS	all-TweetScore
NE	all-Negation	MW	All-modal words
IN	all-Intensifiers	CW	All- contrast words
DI	all-Diminishers		

Tableau 1.9 : Les étiquettes utilisées dans les figures.

Les tweets ont été annotés manuellement par trois annotateurs de langue maternelle arabe / saoudienne, le conflit entre les annotateurs a été résolu par un vote à la majorité.

Le tableau 1.10 présente le Dataset qui a été créés.

	Entrainement	Test	Total
Nombre de textes	15592	1981	17573

Tableau 1.10 : Division de Dataset : Entrainement et test.

Le Dataset est divisé en trois types selon les manières de classification :

a) A deux voies (positif et négatif) : voir la Figure 1.3.

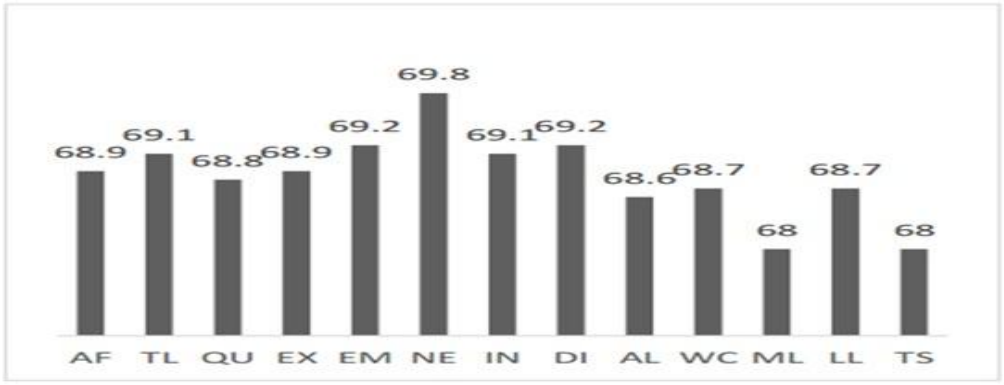

Figure 1.3 : F-score de deux voies.

b) A trois voies (positif, négatif et neutre), voir la figure 1.4.

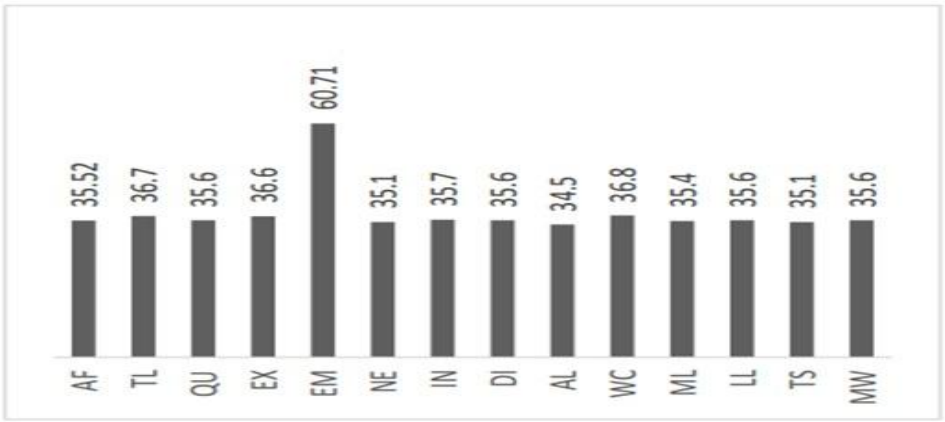

Figure 1.4 : F-score de trois voies.

c) A quatre voies (positif, négatif, neutre et mixte), voir la Figure 1.5.

Figure 1.5 : F-score de quatre voies.

Ils ont constaté que :

D'après la Figure 1.3, La suppression des fonctionnalités : tweetLength, hasPositiveEmoticon, hasNegativeEmoticon, hasNegation, hasIntensifier et hasDiminisher, a donné une augmentation du F-score à 69,5 ce qui signifie que ces fonctionnalités nuisent à la performance et doivent être supprimées, la suppression de la fonctionnalité hasNegation seule était de 69,8 et le meilleur F-score atteint a été de 69,9 en supprimant hasNegation et hasDiminisher.

D'après la Figure 1.4, la suppression des fonctionnalités hasPositiveEmoticon et hasNegativeEmoticon a donné une augmentation substantielle de la performance (+25%), aussi il y a des autres fonctionnalités où leur suppression a amélioré le F-score.

D'après la figure 1.5, la suppression des fonctionnalités tweetLength, hasQuestionMark, hasExclamationMark, hasNegation, hasIntensifier, hasDiminisher, hasContrastWord, hasModalWord, hasPositiveWordMPQA, hasNegativeWordMPQA, hasNegativeWordMPQA et hasPositiveWord a augmenté la performance.

La longueur des tweets, n'a pas eu une amélioration des performances de classification. L'impact de la présence de points d'interrogation et d'exclamation variait, l'impact des modificateurs de valence contextuelle (peu, presque, trop, assez, moins, plutôt...) en tant que caractéristiques n'était pas uniforme dans les modèles de classification et un impact clair ne peut pas être déduit.

hasPositiveWordAraSenti, hasNegative-WordAraSenti, PositiveWordCount, NegativeWordCount, ces fonctionnalités sont les seuls présents dans le meilleur ensemble de fonctionnalités des trois modèles de classification.

1.5.5 Analyse de sentiment - cas de dialecte égyptien

Amira Shoukry et Ahmed Rafea, (2012) travaillaient sur le dialecte égyptien, ils formaient un Dataset de tweets contient 1000 publications annotées 500 positives et 500 négatives, ils ont utilisé les catégories unigramme et bigrammes, ils ont obtenu le résultat décris dans le tableau 1.11 **[47]**.

	SVM (a)	NB (a)	SVM (b)	NB (b)
Unigramme	0.721	0.654	**0.725**	0.646
bigrammes	0.721	0.654	**0.725**	0.646

Tableau 1.11 : F-mesure résultat.

Ou : (a) résultats avant de supprimer les mots vides

(b) résultats après de supprimer les mots vides

Il y a ici une très légère amélioration en SVM et en NB, cela signifie que certains mots importants que ils ont supprimés et ils devraient être utilisés ou il y a d'autres mots vides qui doivent encore être supprimés. SVM a de meilleurs résultats que NB, car le SVM produit généralement des résultats plus précis que le NB. Ceci est dû au fait que NB est basé sur des probabilités, il convient donc mieux aux entrées avec une grande dimensionnalité. D'après le tableau 1.11 le bigramme n'a pas enrichi le résultat.

Dans leurs travaux futurs, ils veulent améliorer le corpus, utiliser d'autres fonctionnalités telle que la négation.

1.5.6 Analyse de sentiment - cas de dialecte jordanien

Nawaf A. Abdulla et al. (2013) ont fait l'analyse de sentiment sur le dialecte jordanien en deux approches, basée lexique et basée corpus. Ils ont collecté et annoté un dataset de 2000 tweets, où 1000 tweets de polarité positives et 1000 tweets de polarité négatives, après le prétraitement. Ils ont réalisé leur dictionnaire de 3479 mots, 1262 mots positifs et 2217 mots négatifs, ils ont emploie les mots de négation (لن - ليس ...), aussi les mots d'intensité (جدا – تماما – فعلا – بافراط) comme ça : si le mot positif (négatif) suivi ou précédé d'un mot d'intensité, il va prendre la valeur +2 (-2) au lieu de +1 (-1) **[48]**.

Les deux approches appliques sont :

a) Basée corpus :

Avec l'emploi de quatre classificateurs, SVM, NB, DT, et KNN (K-nearest neighbors ou KPPV) où k = 9, le tableau 1.12 montre leurs résultats.

Classificateurs	Originale	Root-stemmer	Light-stemmer
SVM	84.7%	85%	**87.2%**
NB	80.4%	78.75%	**81.3%**
KNN	51.3%	**52.85%**	51.45%
D-tree	50%	50%	50%

Tableau 1.12 : Taux d'exactitude (Accuracy) des résultats.

b) Basée lexique :

Dans cette méthode, ils ont devisé le travail en trois phases, la première phase a été avec un dictionnaire de 300 mots, la deuxième phase a été avec 2500 mots et la dernière phase a été en utilisant tout le dictionnaire (3479 mots), le résultat est montré dans le tableau 1.13.

Mesures d'évaluation	**Phase I**	**Phase II**	**Phase III**
Précision	10.5	52.2	58.6
Recall	8.9	48.9	64.9
Accuracy	16.5	48.8	**59.6**

Tableau 1.13 : Résultat de capacité de lexique.

Dans la première approche (Basée corpus) le classificateur SVM a montré le meilleur résultat. Dans la deuxième approche (Basée lexique), il est évident que la taille du lexique a une grande influence sur le résultat d'analyse.

1.5.7 Analyse de sentiment - cas de dialecte algérien

Le dernier travail, celui de M'hamed Mataoui et al. (2016), ils ont travaillé sur le dialecte algérien (ALGD) **[49]**. Leur approche est basé lexique, pour faire son modèle, ils ont créé trois lexique, lexique de mots clés, lexique de mots de négation, lexique de mots d'intensité et ils ont utilisé deux autres ressources, une liste d'émoticônes avec les polarités qui leur ont été attribuées et un dictionnaire d'expressions courantes de l'ALGD. Le lexique de mots clés contient 3093 mots (713 mots positifs et 2380 mots négatifs).

Ils ont collecté et annoté leur propre Dataset qui contient 7698 commentaires Facebook. Le tableau 1.14 présente la distribution des données collectées selon leurs thèmes.

Thèmes	Nombre de commentaires
économie	1705
politique	2422
société	1263
littérature et arts	1215
divers	1093

Tableau 1.14 : Distribution des données collectées selon leurs thèmes.

Le tableau 1.15 montre les résultats obtenus par les deux configurations liées au module de calcul de similarité de phrases courantes **[50]**.

	Sans utiliser "le module de calcul de similarité de phrases courantes"	En utilisant "le module de calcul de similarité de phrases courantes"
Accuracy	76.68 %	**79.13 %**

Tableau 1.15 : Résultats obtenus par les deux configurations liées au "module de calcul de similarité de phrases courantes".

D'après le tableau 1.15, nous avons remarqué que la meilleure configuration de leurs expériences est liée à l'utilisation de module de calcul de similarité de phrases courantes.

1.6 Discussion

On remarque que chaqu'un de ces travaux est particulièrement caractérisé par une ou plusieurs caractéristiques, dans le premier travail celui sur les MSA, les tweets étaient en économique, les chercheurs de ce travail ont utilisé deux techniques, la validation croisée et le pourcentage scission. Le deuxième travail (Dialecte Tunisien), les chercheurs ont fait une enquête sur les ressources disponibles. Le troisième travail ainsi que le travail sur le dialecte égyptien sont caractérisés par l'utilisation d'unigrame et bigrame, en plus dans ce dernier, les chercheurs ont utilisé un plus petit corpus (1000 tweets). La caractéristique à citer dans le quatrième travail celui des chercheurs ont utilisé le dialecte saoudien est qu'ils ont utilisé une méthode hybride. Ils ont aussi utilisé trois voies et quatre voies mais les trois premiers travaux ont employé seulement deux voies (positif et négatif). La chose la plus important dans ce travail est que la comparaison se fait en utilisant 19 fonctionnalités. Il faut noter ici que le travail qui utilise le dialecte jordanien enregistre le plus haut résultat qui est de 87.2% de réussite. Il faut noter aussi que les trois travaux qui utilisent les dialectes suivants : algérien, saoudien et jordanien sont caractérisés par l'utilisation des fonctionnalités supplémentaire (la négation et l'intensité).

Pour les classificateurs, le SVM est utilisé presque dans tous les travaux. Nous avons remarqué que dans tous les travaux (à part les travaux qui utilisent le dialecte tunisien et l'algérien), le classificateur SVM a réalisé les meilleurs résultats.

1.7 Conclusion

Nous avons étudié dans ce chapitre, les différentes définitions de la polarité d’une opinion, ainsi que ses approches qui utilisent la polarité unique et qui utilisent la polarité basée sur les aspects, puis nous avons défini les deux méthodes d’apprentissage automatique, l’apprentissage supervisé et celui-ci non supervisé. Nous avons présenté les approches d’analyse de sentiments qui sont l’approche basée sur lexique, l’approche basée sur corpus et l’approche hybride. Ensuite, nous avons cité quelques difficultés qui peuvent être rencontrées durant la détection de sentiments tels que l’ambiguïté de la polarité de certains mots et les difficultés due au langage naturel.

Dans le chapitre suivant, nous allons étudier quelques aspects de particularité de lange arabe et de dialecte.

Chapitre 2

La langue arabe et le dialecte

Chapitre 2

2 La langue arabe et le dialecte

2.1 Introduction

Les langues sémitiques (en référence au nom de "**Sem**" fils de "**Noé**") forment un ensemble de langues parlées depuis le plus ancien temps au Moyen-Orient, au Proche-Orient ainsi qu'en Afrique du Nord, La langue sémitique est une des branches de la famille des langues afro-asiatiques, où on ne sait pas de manière certaine l'origine et l'expansion géographique de ces langues, soit de l'Asie vers l'Afrique ou le contraire.

L'origine du mot "*Arabe*" reste inconnu, malgré il y a beaucoup de recherches, le radical "*arab*" désigne le désert et c'est un mot araméen "arâbâh", le mot arabe peut dériver aussi de la racine sémitique Abhar "se déplacer", mais l'étymologie arabe considère que le mot "*arabe*" est dérivé du verbe "exprimer" **[51]**.

2.2 Voyellation

Dans la langue arabe nous avons deux types de signes de notation des voyelles, le premier type sont les voyelles brèves, qui sont notées au moyen de signes diacritiques, le deuxième type sont les lettres d'allongement d'une voyelle.

Les voyelles brèves sont : fatha, damma et kassra, se rajoutent sur les lettres, pour donner le vrai sens (la prononciation précise) de chaque mot et pour éviter les ambiguïtés (les conflits) lorsque le contexte n'est pas suffi, il y a aussi des autres diacritiques moins utilisés comme sukun, shadda et tanwn.

Les voyelles longues (lettres d'allongement) sont : "Alif", "waw" et "ya", avec les sons respectivement "aa", "ou", "ii" **[52]**.

2.3 La richesse de la langue arabe

La langue arabe s'écrit au moyen de 28 lettres, le terme "*abjad*" désigne le système d'écriture, elle dispose le plus grand nombre de mots avec plus de 12 millions de mots où l'Anglais 600 000 mots, le Français 150 000, mots, le Russe 130 000 mots...

Elle est une langue très riche, Il y aurait 80 termes différents pour identifier le miel, 200 pour le serpent, 500 pour le lion, 1000 pour le chameau et l'épée et jusqu'à 4400 pour définir l'idée de malheur, les grammairiens arabes prétendent que toutes les racines ont été originalement des verbes, où le nombre de ces racines en réalité est de 6000 **[53]**.

2.4 Eléments de structure de la langue arabe

Comme toutes les langues sémitiques, l'arabe se caractérise par l'utilisation de certains schémas morphologiques (modèles de formation des mots), ces modèles permettent de fournir des mots à partir de racines abstraites, ces racines se composent d'habitude de trois consonnes qui forment les unités de base pour obtenir de nombreux mots dérivés de cette racine.

On trouve par exemple la racine KTB se retrouve dans le verbe dérivé KaTaB "à écrire" qui peut être conjugué en ajoutant de préfixes et de suffixes convenables.

Le schéma KaTaB peut être représentée par le schéma $C_1aC_2aC_3$ où C_1= K, C_2 = T, et C_3 = B. Le doublement de la consonne médiane de la racine donne "$C_1aC_2C_2aC_3$" qui se traduit par le schéma *KaTTaB* qui a le sens causal "a fait écrire". Par allongement de la première voyelle on obtient *KaaTaB*, par l'ajout du préfixe "ta" nous aurons taKaaTaB (notion de réciprocité).

Chacun de ces schémas peut être modifié pour montrer le passé, le présent…, par exemple KuTiB " été écrit", yaKTuB "écrit" et yuKTaB "est écrit. Ainsi que le pluriel, par exemple Kutub "des livres", Kuttab "des écrivains", maKTaBaat "des bibliothèques", en plus, quelques-uns peuvent être mis au féminin par des terminaisons dédiées, par exemple KaaTiba "une femme écrivain", muKaaTiBa "correspondantes" et KaaTiBaat "des femmes écrivains".

Ces exemples présentent un modèle riche et varié de schémas dérivés d'un seul mot qui permet d'engendrer des centaines de mots ont référencé à la même racine **[54]**.

2.5 Arabe algérien

L'Arabe Algérien ou bien le dialecte algérien, est un groupe d'Arabe Nord-Africain, dialectes mélangés avec différentes langues parlées en Algérie, le frottement de plusieurs langues à travers l'histoire de la région a produit un langage complexe et riche comprenant

des mots, des expressions et des structures linguistiques, ces langues tel que le Berbère, le Français, l'Italien, l'Espagnol et le Turc ainsi que des autres langues romanes méditerranéennes. Le dialecte algérien est fortement influencé particulièrement par le français où on peut trouver la commutation de codes et emprunter **[55]**.

Nous pouvons catégoriser les paroles en ce dialecte comme la suite :

MSA encodé en lettres arabiques, par exemple : « عيدكم مبارك كل عام وانتم بخير »

MSA encodé en lettres romanisées, par exemple : « ana said hada elyaoum »

Dialecte encodé en lettres arabiques, par exemple : « كيراك شريكي »

Dialecte encodé en lettres romanisées, par exemple : « maysoumouch bla shour »

Langue étrangère encodée en lettres romanisées, par exemple : « je vais à l'univ »

Langue étrangère encodée en lettres arabiques, par exemple : « برافووو »

Donc chaque texte est l'une de ces catégories ou un mélange d'eux, on peut aussi trouver une autre écriture particulière celui l'utilisation d'un numéro au lieu d'une lettre par exemple : « b1-بيا-bien », « 3adi-عادي-normal », « 5amsa-خمسه-cinq », « 6ariq-طريق-route », « sa7b-صاحب- ami », « 9raya-قراية-étude »...

2.6 Les difficultés de l'arabe et du dialecte

Nous pouvons citer les difficultés suivantes :

- Un seul dialecte (par exemple le dialecte algérien) peut contenir plusieurs sous dialectes.

- La grande distance entre MSA et quelques dialectes.

- Une racine peut prendre plusieurs formes en fonction du contexte.

- La répétition d'une lettre plusieurs fois pour intensifier le sens ou le sentiment (bzeeeef).

- L'arabe a divers signes diacritiques, la présence ou l'absence de tels signes, peut changer totalement le sens des mots.

- Les mots de négation utilisés pour nier les verbes au passé ou au présent, qui changent le sens exactement au contraire **[56]**.

2.7 Conclusion

Dans ce chapitre, Nous avons scruté l'histoire de notre langue et quelque caractéristiques, sa richesse de vocabulaire et plus important, sa multitude d'idiomes et surtout de façon particulière comme nous avons en Algérie, où il est presque chaque wilaya à son propre accent, c'est une des principales raisons du manque des travaux qui utilisent le dialecte algérien.

Dans le chapitre suivant, nous allons présenter notre modélisation, où nous allons étiqueter un ensemble de commentaires à partir des réseaux sociaux et regrouper un ensemble de mots en dialecte algérien.

Chapitre 3

Modélisation

Chapitre 3

3 Modélisation

3.1 Introduction

Dans ce chapitre, nous commençons de créer notre modèle d'analyse d'opinion. En débutant par notre contribution. Ensuite, nous décrivons la source de données sur laquelle le modèle est appliqué. Finalement, nous avons fait l'annotation de corpus et la construction de lexique, qu'ils consomment beaucoup de temps et nécessitent un grand effort.

3.2 Contribution

Nos contributions principales consistent à :

a) Travailler sur le dialecte Algérien, avec l'exploitation de quatre classificateurs, Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF) et Naïve Bayes (NB), qui est considéré à notre savoir le premier travail qui teste l'utilisation des quatre classificateurs.
b) Annotation un Dataset qui contient 2891 commentaires.
c) Faire un dictionnaire de 1328 mots annotés en dialecte Algérien.

3.3 Source de données

Pour gagner du temps, nous avons exploité le Dataset utilisé dans le travail de Mansour et Trad (2018). Ils ont construit leur propre corpus en dialecte algérien, par le développement d'un outil avec le langage de programmation Python, qui permet d'interroger l'API de Facebook pour récupérer des postes et des commentaires, ici ils ont rencontré plusieurs difficultés telles que la limitation des APIs fournis par Facebook par rapport à twitter, où ce dernier est peu utilisé à l'Algérie **[52]**.

Leur Dataset est divisé en trois parties, normale, offensante et obscène. Nous avons travaillé sur la première partie, par ce qu'elle ne contient pas des paroles impur (sale) et elle est un mélange de textes avec des sentiments positifs, négatifs et neutres, en plus, les deux autres parties (offensante et obscène) sont considérées comme des textes de polarité négative, ça va déséquilibrer la distribution de polarité dans notre Dataset.

Le tableau 3.1 présente le nombre des commentaires et leurs sources.

Source	Normale	Offensante	Obscène	Total de Commentaires
Facebook	2892	1497	611	5000

Tableau 3.1 : Données collectées.

Comme toute opération d'assemblement de données, cette étape requit à un prétraitement, qui consiste à filtre les téléchargements afin de garder seulement les textes arabes et éliminer les mots vides pour obtenir un corpus propre et prêt à utiliser.

3.4 Annotation

L'annotation ou l'étiquetage des opinions est une tâche humaine qui est un peu difficile car elle prend beaucoup du temps pour suivre les commentaires un par un, et elle a besoin des fois une grande discussion pour atteindre une décision finale, est ce que cette opinion est positive ou négative !

Alors, entre trois et quatre semaines et avec deux annotateurs (un troisième pour que le conflit entre les annotateurs va être résolu par un vote à la majorité) nous avons étiqueté toutes les 2891 entrées par l'utilisation de trois polarités positive, négative et neutre (qui ont les valeurs respectivement 1, -1, 0). Le tableau 3.2 divise le Dataset selon les trois valeurs de polarités.

Polarité	Positive	négative	neutre	Total
Nombre de commentaires	975	525	1391	2891

Tableau 3.2 : Nombre de commentaires par polarité.

Dans le tableau 3.3, nous avons décrit quelques exemples avec leurs polarités.

Commentaire	Polarité
هذي جياحة نريسكي بعمري على حاجة تستاهل ماشي علاجال بوزوج رجلين يجري مور البالون	négative
شفت المواطن مش واعي خلاص داير المزبلة حذاه يكل منا او يرمي منا	négative
في بسكرة خويا لعزيز طريق سيدي عقبة ؟	neutre
هادو روشيات تاع رخام ماشي ثلج	neutre
ربي يكثر من امثالك	positive
درس في حسن ضيفة كل إحترام وتقدير ناس المدية	positive

Tableau 3.3 : Exemple de notation de quelques commentaires.

3.5 Création de dictionnaire

Est un ensemble des mots qu'un classificateur peut les utiliser pour évaluer la polarité d'un texte. Dans la littérature, nous n'avons pas trouvé un lexique spécifique des mots en dialecte algérien. Donc, nous aurons obligé d'établir notre propre dictionnaire. Quatre semaines de travail pris pour ramasser 1328 mots à partir de sites web. Aussi, nous avons demandé l'aide des amis du nord de l'Algérie, du sud, de l'est, de l'ouest afin de couvrir le maximum possible des régions de notre pays, ces mots sont étiquetés comme l'illustre le tableau 3.4.

Polarité	Positive	Négative	Total
Nombre de mots	565	763	1328

Tableau 3.4 : Statistiques de notre dictionnaire.

La Figure 3.1 et la Figure 3.2 présentent un exemple d'une partie de dictionnaire positif et négatif respectivement.

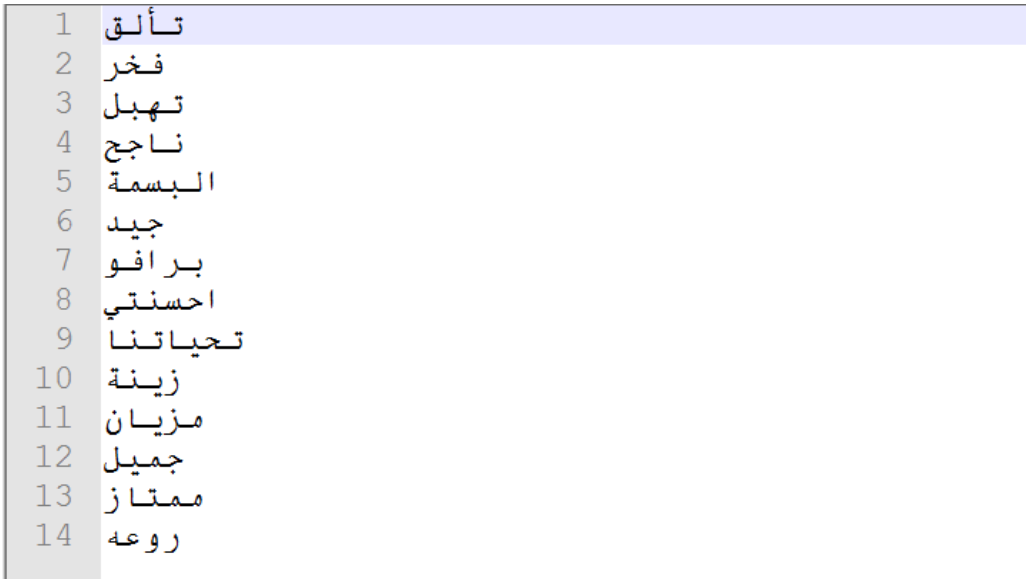

Figure 3.1 : Exemple d'une partie de dictionnaire (positif).

Figure 3.2 : Exemple d'une partie de dictionnaire (négatif).

3.6 Conclusion

Dans ce chapitre, nous avons fait beaucoup d'effort dans deux volets important dans ce travail. Le premier est l'annotation d'un corpus de 2891 publications écrits en dialecte algérien par positive, négative et neutre. Le deuxième, celui de la création d'un dictionnaire de 1328 mots aussi en dialecte algérien.

Dans le chapitre suivant, nous allons présenter notre expérimentation et discuter les résultats obtenus.

Chapitre 4

Implémentation

Chapitre 4

4 Implémentation

4.1 Introduction

Dans ce chapitre, nous allons implémenter les algorithmes d'analyse de sentiments. Nous allons principalement implémenter quatre algorithmes. Ces algorithmes sont : Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF) et Naïve Bayes (NB). Nous allons utiliser des publications qui ont été normalisées et un corpus prêt à utiliser. Enfin, nous allons discuter les résultats de la classification.

Nous allons utiliser l'algorithme général de l'analyse suivant :

Algorithme

Importer les bibliothèques

Lire le Dataset

Lire le dictionnaire

Début

Extraire les fonctionnalités

Préparation

Appeler aux classificateurs

Afficher les résultats

Fin

4.2 Ressources utilisées

Dans notre expérimentation, nous avons utilisé deux PCs, le premier était de marque HP Pavillon, le deuxième était de marque Dell, avec des processeurs multi-coré I3, des horloges de fréquence de 2.40 GHZ et des RAM de 4 GO.

Pour la programmation de l'application, nous avons utilisé l'environnement Python. Python est un langage de programmation portable, dynamique, extensible, gratuit, syntaxe très simple, code plus court que C ou Java, multi thread, orienté objet, évolutif … **[57]**.

Egalement, nous avons utilisé les packages suivants :

- CSV : Une bibliothèque grâce à lui, nous pouvons manipuler les fichiers de format csv.

- Gensim : Est une bibliothèque gratuite Python pour l'extraction automatique des sujets sémantiques des documents.

- Scikit-learn : Est une bibliothèque d'apprentissage automatique en Python, c'est le moteur qui alimente de nombreuses applications de l'intelligence artificielle et de la fouille des données.

4.3 Exemples de codes sources

Dans cette section, nous allons présenter quelques exemples de codes sources.

La Figure 4.1 présente un morceau de code qui permet d'appeler les bibliothèques nécessaires pour compiler notre application

```
2
3 import pandas as pd
4 import re
5 from data import SW
6 from sklearn.svm import SVC
7 from sklearn.model_selection import train_test_split
8 from  sklearn.tree import DecisionTreeClassifier
9 from sklearn.ensemble import RandomForestClassifier
10
11
```

Figure 4.1 : Appel des Bibliothèques.

La Figure 4.2 présente les instructions nécessaires pour lire le Dataset et le dictionnaire.

```
13
14 corpus = pd.read_csv("Dataset.csv")
15 Neg = pd.read_csv("negg.csv")
16 Pos = pd.read_csv("poss.csv")
17 df = pd.read_csv("sem2.csv", sep="\t")
18
```

Figure 4.2 : Lire le Dataset et le dictionnaire.

La Figure 4.3 présente le code source qui permet de compter les mots positifs.

```
90
91 CP = []
92 count = 0
93 for t in corpus.text:
94     for word in t.split():
95         for word1 in Pos.text:
96             if word == word1:
97                 count +=1
98     CP.append(count)
99     count = 0
100
101 corpus["pos_count"] = CP
102
```

Figure 4.3 : Compter le nombre de mots positifs.

La Figure 4.4 présente le code source qui permet de compter les mots négatifs.

```
75
76 CN = []
77 count = 0
78 for t in corpus.text:
79     for word in t.split():
80         for word1 in Neg.text:
81             if word == word1:
82                 count +=1
83     CN.append(count)
84     count = 0
85
86 corpus["neg_count"] = CN
87
```

Figure 4.4 : Compter le nombre de mots négatifs

La Figure 4.5 présente les instructions qui permettent de tester l'existence d'un mot positif.

```
63
64 OP = []
65 for t in corpus.text:
66     if any(w in t.split() for w in Pos.text):
67         OP.append(1)
68     else :
69         OP.append(0)
70
71 corpus["one_pos"] = OP
72
```

Figure 4.5 : Tester l'existence d'un mot positif.

La Figure 4.6 présente les instructions qui permettent de tester l'existence d'un mot négatif.

```
51
52 ON = []
53 for t in corpus.text:
54     if any(w in t.split() for w in Neg.text):
55         ON.append(1)
56     else :
57         ON.append(0)
58
59 corpus["one_neg"] = ON
60
```

Figure 4.6 : Tester l'existence d'un mot négatif.

La Figure 4.7 présente les instructions qui permettent d'introduire la fonctionnalité de la langueur.

```
105
106 word_count = []
107 for i in corpus.text:
108     word_count.append(len(i.split()))
109
110 corpus["word_count"] = word_count
111
```

Figure 4.7 : Introduire la fonctionnalité de la langueur.

La Figure 4.8 présente les instructions qui permettent d'introduire la fonctionnalité du pourcentage de sentiment.

```
21
22 df.columns=["val", "text"]
23
24 data = zip(df.val,df.text)
25 pos = []
26 neg = []
27 for i,j in data:
28     if i > 0:
29         pos.append((i,j))
30     else:
31         neg.append((i,j))
32
33 def f(text, ls):
34     target = []
35     text = text.split()
36     for i in ls:
37         if re.sub("#", "",i[1]) in text:
38             target.append(i)
39
40     return sum([i[0] for i in target ])
41
42 corpus["sentiment"] = [f(t, pos)+f(t, neg) for t in corpus["text"]]
43
```

Figure 4.8 : Introduire la fonctionnalité du pourcentage de sentiment.

La Figure 4.9 présente les instructions qui permettent de préparer le corpus à analyser.

```
145
146 Z = corpus.get(corpus.columns[1:])
147
148 Y = corpus.get(Z.columns[0])
149 X = corpus.get(Z.columns[1:])
150 x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=4)
151
```

Figure 4.9 : Préparation à l'analyse.

Les Figures 4.10, 4.11, 4.12 et 4.13 présentent les instructions qui permettent d'appeler les Classificateurs SVM, DT, RF et NB, respectivement.

```
154
155 svm = SVC()
156 svm.fit(x_train, y_train)
157 acc1 = svm.score(x_test, y_test)
158 pr1 = svm.predict(x_test)
159
```

Figure 4.10 : Appel du classificateur SVM.

```
162
163 dt = DecisionTreeClassifier()
164 dt.fit(x_train, y_train)
165 acc2 = dt.score(x_test, y_test)
166 pr2 = dt.predict(x_test)
167
```

Figure 4.11 : Appel du classificateur DT.

```
170
171 rf = RandomForestClassifier(n_estimators=40)
172 rf.fit(x_train, y_train)
173 acc3 = rf.score(x_test, y_test)
174 pr3 = rf.predict(x_test)
175
```

Figure 4.12 : Appel du classificateur RF.

```
x_train, x_test, y_train, y_test=train_test_split(X, y, test_size=0.2, random_state=4)
```

```
cv = cv()
```

```
x_train_cv = cv.fit_transform(x_train)
x_test_cv =  cv.fit_transform(x_test)

x_train_cv = cv.transform(x_train)
x_test_cv =  cv.transform(x_test)
```

```
model = MultinomialNB()
model.fit(x_train_cv, y_train)
pr = model.predict(x_test_cv)
model.score(x_test_cv, y_test)
```

Figure 4.13 : Appel du classificateur NB.

4.4 Fonctionnalités

Comme nous l'avons vu précédemment dans le travail de Nora Al-Twairesh et al (2018) où ils ont constaté que les fonctionnalités qui ont le plus importante influence dans l'analyse des opinions sont ces quatre, l'existence de mots positifs (et/ou négatifs) et le nombre de mots positifs (et/ou négatifs) dans le commentaire **[41]**.

En basant sur leurs études, nous allons faire notre expérimentation où nous avons ajouté sur ces quatre fonctionnalités deux autres, qui sont la langueur du texte et le niveau du sentiment.

Le niveau du sentiment désigne son approfondie, pour appliquer cette fonctionnalité nous avons utilisé le document (SemEval2016 Arabic Twitter Sentiment Lexicon) **[58]**. La figure 4.14 présente une partie de ce document.

1	0.963	حب
2	0.925	فرح
3	0.925	نجاح
4	0.912	سرور
5	0.912	مبروك
6	0.900	صادق
7	0.900	متفائل
8	0.900	سعادة
9	0.875	نعيم
10	0.875	مبدع
11	0.875	سلام
12	0.838	رائع

1	0.000	وكيل
2	0.000	عند
3	0.000	احصي
4	0.000	منزمان
5	0.000	ليل
6	0.000	دخل
7	0.000	مكافحة

1	-0.087	غبن
2	-0.250	تانيب
3	-0.275	طفش
4	-0.287	غصة
5	-0.287	دمع
6	-0.287	زكام
7	-0.300	افقد
8	-0.313	هروب
9	-0.325	صعب
10	-0.375	عيب
11		

Figure 4.14 : Exemple de niveau de sentiments.

En ce qui concerne la Négation, il y a des travaux la considèrent comme une fonctionnalité indépendante (par exemple **[41]**), mais l'utilisation de la Négation comme ça va diminuer la précision d'analyse **[41]**.

Par contre, dans notre travail, nous avons intégré la négation dans les fonctionnalités : nombre de mots positif et nombre de mots négatifs, où le modèle vérifie si cette négation est suivie par un mot positif (négatif) ou non !, si oui il ajoute 'un' au nombre de mots négatifs (positifs), si la négation n'est suivie ni par un mot positif ni par un mot négatif, ça ne signifie rien.

Le tableau 4.1 présente les fonctionnalités utilisées dans notre implémentation.

Fonctionnalité	Abréviation	Signification
Has Positive Word	HPW	0 ou 1
Has Negative Word	HNW	0 ou 1
Positive Word Count	PWC	>= 0
Negative Word Count	NWC	>= 0
CommentLength	CL	Numérique > 0
SentimentLevel	SL	-1 <= V <= 1

Tableau 4.1 : Les fonctionnalités utilisées.

4.5 Expérimentations et résultats

Tant que nous utilisons la méthode d'apprentissage supervisé et l'approche hybride, nous avons divisé le corpus en deux parties, 80% pour l'entrainement et 20% pour le test. Nous avons fait plusieurs tests, les résultats Accuracy sont présentés dans le tableau 4.2.

Test	fonctionnalité / Classificateur	SVM	DT	RF	NB
1	Toutes les fonctionnalités	**85.14**	83.07	**85.31**	**84.28**
2	HPW, HNW, PWC, NWC	84.11	**84.45**	84.28	82.38
3	PWC, NWC, CL, SL	84.62	83.24	85.31	83.76
4	PWC, NWC	84.11	84.28	84.45	67.87
5	HPW, HNW, CL, SL	84.11	84.11	84.11	48.35
6	HPW, HNW	84.11	84.11	84.11	72.30
7	CL, SL	47.49	47.66	48.18	48.18

Tableau 4.2 : Résultats de classification.

4.6 Discussion

Le meilleur résultat dans tous les tests est 85.31%, il a obtenu par le RF avec l'utilisation de toutes les fonctionnalités, c'est la même chose pour SVM et NB, leurs meilleurs résultats étaient avec le premier test (85.14%, 84.28% respectivement), par contre DT a atteint sa maximum mesure (84.45%) dans le test (2).

D'après les tests (2), (4) et (6) nous avons remarqué que les deux couples (PWC, NWC) et (HPW, HNW) ont eu presque le même poids d'influence, c'est logique parce que le nombre de commentaires mixtes (qui contiennent des mots positifs et au même temps des mots négatifs) est usuellement petit, c'est pour ça quand nous avons exploité le nombre de mots d'une polarité, la mesure était un peu plus grande (de 84.11% à 84.45%) à part NB qu'il a eu une variation considérable.

Selon les tests (3) et (4), nous avons constaté que l'ajout des fonctionnalités (CL, SL) aux fonctionnalités (PWC, NWC) a amélioré les résultats de classificateurs sauf que DT qui a diminué les résultats.

Dans les tests (5) et (6), nous avons constaté que l'ajout des fonctionnalités (CL, SL) aux fonctionnalités (HPW, HNW) n'a fait rien sauf que la mesure de NB a diminué.

D'après le dernier test, nous avons trouvé que les fonctionnalités (CL, SL) ne peuvent pas être seules, puisque si le cas, ils vont donner le plus mouvais résultat (moins de 49%).

Ces résultats montrent que le SVM est généralement considéré comme un meilleur classificateur.

Si on fait la comparaison de notre travail avec les autres travaux qui utilisent des autres dialectes à savoir, le tunisien, le marocain, l'égyptien, le saoudien, et le jordanien. Nous pouvons conclure que nos classificateurs ont réalisé des bons résultats. Le tableau 4.3 récapitule la comparaison qui a été faite.

Le dialecte du travail	**Classificateur**	**Résultats**	
		Accuracy	**F-mesure**
Algérien (notre travail)	**RF**	**85.31 %**	**84.93 %**
Jordanien [48]	SVM	87.2 %	/
Algérien [49]	/	79.13 %	/
Tunisien [35]	MLP	/	78 %
Marocain [38]	SVM	/	78 %
Egyptien [47]	SVM	/	72.5 %
Saoudien [41]	SVM	/	69 %

Tableau 4.3 : Comparaison des résultats des différents travaux.

Nos résultats auraient certainement être meilleurs si nous aurons utilisé un corpus qui contient plus que 2891 publications.

4.7 Exemples de sorties

Malgré nous avons abouti une bonne mesure, notre application a effectué quelques erreurs. Ces erreurs sont listées dans le tableau 4.4.

Exemple	Publication	Notre annotation	Résultat système
1	عيد سعيد و مبارك و كل عام و أنت بألف خير إن شاء الله	1	1
2	سفيان فيغولي يتطوع لإحدى المائدات الإفطارية للمسلمين في فرنسا ، برافو سفيان	1	1
3	هذا فعل لا اخلاقي اين تعليم سيدنا محمد(ص)عندما كان جاره يهودي وعندما مرت جنازة يهودي بالله عليكم ماذا فعل؟؟؟!!!!!!!	-1	-1
4	مالقيتو ما ديرو جايبنا المصري تع شكوبي هادا خماج	-1	-1
5	انا حاب كاس رايب و ربع كسرة خير من لفريت	0	0
6	إذا عرفتو في اقل من 5 ثواني دير جام	0	0
7	الجميع يشهد أن # الدوون هو الافضل ولكن عبودي الكتلوني لديه رأي آخر	0	1
8	مهابل وتلاقاو	0	-1
9	يعمري هديك اتبسيمة	1	0
10	هذا الانسان ماهوش رجل وطني	-1	0

Tableau 4.4 : Exemples de résultats de l'analyse.

Nous pouvons expliquer les causes de ces erreurs dans ce que suit.

1) Dans le premier exemple (7), l'erreur se produit par ce que le modèle compte dans le texte un mot positif 'الافضل' et il le classe comme positif alors que l'existence d'un mot de polarité pas toujours fait le texte la prendre.

2) Dans le deuxième exemple (8), nous avons annoté cette publication par neutre, l'erreur survenait quand le système a trouvé le mot négatif 'مهابل' où ce dernier ne rend pas toujours le texte avoir un sentiment négatif.

3) Dans le troisième exemple (9), cette erreur parmi les erreurs qui sont difficiles à régler, on a le mot 'اتبسيمة' est un mot positif, mais le système ne l'a pas détecté car il est écrit de façon anormal (inattendue), où la façon attendue est 'التبسيمة', puisque dans les réseaux sociaux chacun écrit le mot comme il veut, donc il est dur pour le modèle à analyser.

4) L'erreur du quatrième exemple (10) dû au contexte, où le contexte dit que cette phrase est négative, alors qu'elle n'a pas aucun mot négatif, une des solutions de ce type d'erreurs est d'utiliser le bigrame, quand on considère que 'رجل وطني' est un mot (bigrame) a une polarité positive.

4.8 Conclusion

Dans ce chapitre, nous avons fait l'analyse de sentiments sur un corpus qui contient 2891 textes en dialecte algérien étiquetées comme le suivant : 975 textes positifs, 525 textes négatifs et 1391 neutres. Nous avons exploité quatre classificateurs d'apprentissage automatique qui sont machine à vecteurs de support (SVM), arbre de décision (DT), forêt d'arbres décisionnels (RF), naïve bayésienne (NB) où l'évaluation de ces classificateurs se fait par 20% du corpus. Nous avons utilisé six fonctionnalités qui sont HasPositifWord, HasNegatifWord, PositiveWordCount, NegativeWordCount, CommentLength et SentimentLevel. Nous avons fait sept tests différents, le premier s'est fait en utilisant toutes les fonctionnalités et les autres tests se sont faits par la substitution entre ces fonctionnalités. Nous avons comparé les résultats de tests pour les quatre classificateurs. Nous avons trouvé que le bon Accuracy (85.31%) est atteint par le classificateur Random Forest (RF). En fin, nous avons cité quelques exemples d'erreurs d'analyse dans notre modèle et nous avons expliqué comment le modèle les a faits.

Conclusion générale et perspectives

Synthèse

L'objective de ce mémoire est la détection des polarités des publications dans les réseaux sociaux selon trois voies, une publication positive, une publication négative et une publication neutre. Le but de notre travail est la réalisation d'une application sous Python qui utilise une source de données (Dataset, format .csv) contient des textes annotés par des valeurs 1,-1 et 0, basé sur un lexique de mots pour classifier ces textes. Nous avons commencé par la définition de quelques concepts utilisés dans ce mémoire. Ensuite, nous avons met l'accent sur sept travaux connexes, puis nous avons étudié les particularités linguistiques du dialecte Algérien. La phase la plus difficile de ce mémoire est la création d'un lexique d'idiomes algérien qui va surement aider les chercheurs par la suite, dans ses travaux futurs. Dans ce domaine, nous avons utilisé six fonctionnalités (features). Les quatre classificateurs que nous avons appliqués sont SVM, DT, RF et NB. Les résultats obtenus sont très encourageant, où nous avons atteint un meilleur Accuracy celui de 85.31% dans le cas de l'utilisation du classificateur RF.

Perspectives

Pour les travaux futurs, nous pouvons citer :

- L'enrichissement de notre dictionnaire par plus des mots de dialecte algérien en couvrant plus largement d'autres zones car l'Algérie est très vaste et il contient des dizaines de dialectes,

- L'enrichissement de Dataset par d'autres commentaires en dialecte algérien afin d'obtenir des résultats bien précis.

- L'application d'autres classificateurs et l'utilisation d'autres fonctionnalités.

- L'analyse par l'utilisation de la classe Mixte autre que les classes positive, négative et neutre.

- L'utilisation des autres configurations tel que bigramme, trigramme et mixte.

Références

[1] Avis client sur Facebook : danger ou opportunité pour les marques, http://www.brand-advocacy.fr/avis-facebook-marques/, Visité le 15/02/2019.

[2] Sébastien Gillot, Fouille d'opinions, Colloque du Master Recherche en Informatique, 2010, 1-35.

[3] Flora Even, l'influence de Facebook sur les idées politiques, https://www.rtbf.be/culture/dossier/chroniques-culture/detail_l-influence-de-facebook-sur-les-idees-politiques-flora-eveno?id=9458372, Visité le 16/02/2019.

[4] Sentiment définition et synonyme, https://www.mediadico.com/dictionnaire/definition/sentiment/, Visité le 20/02/2019.

[5] Opinion définition, https://www.larousse.fr/dictionnaires/francais/opinion/56197, Visité le 22/02/2019.

[6] Mathieu Troillet, Avantages et inconvénients des réseaux sociaux, en particulier «Facebook », pour la promotion dans les secteurs socioprofessionnels, école supérieure - Sion Suisse, 2015.

[7] Lexique, https://www.notrefamille.com/dictionnaire/definition/lexique, 05-03-2019.

[8] Qu'est-ce qu'un corpus, http://theses.univ-lyon2.fr/documents/getpart.php?id=lyon2.2002.carras_c&part=53655, Visité le 25/02/2019.

[9] Bastien L, Machine Learning – Définition, fonctionnement et secteurs d'application, http://www.artificiel.net/machine-learning-definition, Visité le 05-03-2019.

[10] Claire Gardent, Traitement des Langues Naturelles, LORIA Nancy France, 2011.

[11] Vapnik,https://scholar.google.com/citations?user=vtegaJgAAAAJ&hl=en, Visité le 05-03-2019.

[12] Jamal Kharroubi, Etude de techniques de classement "Machines à vecteurs supports" pour la vérification automatique du locuteur, école nationale supérieure des télécommunications Paris, 2002.

[13] Abe Burrows, Arbres de décision, http://cedric.cnam.fr/vertigo/Cours/ml2/coursArbresDecision.html, Visité le 01/03/2019.

[14] Fabien Moutarde, Arbre de Décision et Forêts aléatoires, Centre de Robotique CAOR MINES ParisTech ENSMP- PSL research University, 2017.

[15] Laboratoire de Biostatistique et Informatique Médicale - Faculté de Médecine - Université de Strasbourg, théorème de Bayes https://dun.unistra.fr/ipm/unit/bayesien/co/1_2_3.html, Visité le 07-03-2019.

[16] Christophe Salperwyck, Vincent Lemaire et Carine HuePowerspace, Classifieur naïf de Bayes pondéré pour flux de données, Powerspace, 2014, 275-286.

[17] Renuka Joshi, Interpretation of Performance Measures, https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/, Visité le 08-03-2019.

[18] Jean-Baptiste Camps, N-grams et identificationdes auteurs, https://graal.hypotheses.org/417, Visité le 20/03/2019.

[19] An Introduction to Classification : Feature Selection, http://www.improvedoutcomes.com/docs/WebSiteDocs/Classification_and_Prediction/SLAM/An_Introduction_to_Classification.htm, Visité le 09-03-2019.

[20] Arabe standard moderne, https://fr.unionpedia.org/Luette, Visité le 10-03-2019.

[21] Banfield, Unspeakable sentences: narration and représentation in the langage of fiction », American literary expert and linguist, 1982.

[22] Hatzivassiloglou et Mckeown, Predicting the semantic orientation of adjectives, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997,174-181.

[23] Bo Pang et Lillian Lee et Shivakumar Vaithyanathan, Sentiment Classification using Machine Learning Techniques, EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10, 2002, 79-86.

[24] Carol Hermann, Entre Web 2.0 et 3.0: opinion mining, écoles de la HES-SO Genève 2010.

[25] Thibaut Thonet, Modèles thématiques pour la découverte non supervisée de points de vue sur le Web, l'Université Toulouse 3 Paul Sabatier, 2017.

[26] Cynthia Van Hee, L'analyse des sentiments appliquée sur des tweets politiques : une étude de corpus, Faculté associée de linguistique appliquée Université Bruxelles Belgique, 2013.

[27] Damien Poirier et Françoise Fessant et Cécile Bothorel et Emilie Guimier de Neef et Marc Boullé, Approches Statistique et Linguistique Pour la Classification de Textes d'Opinion Portant sur les Films, Revue des Nouvelles Technologies de l'Information RNTI-E-17, 2010, 147-169.

[28] Soumia Elyakoute Herma et Khadidja Saifia, Analyse des sentiments cas Twitter, Mémoire de master, Université de Ghardaia, 2016.

[29] Randa BENKHELIFA, Saliha GAGUI, Fouille de données d'opinion des usagers de sites E-commerce, Mémoire de master, Université Ouargla, 2013.

[30] BEGHDAD Abdelkrim et OUSERIR Amina, Une approche Deep Learning pour l'analyse des Sentiments Sur Twitter, Mémoire de master, de Université Khemis Milana Ain Defla, 2018.

[31] Morgane marchand, Domaines et fouille d'opinion Une étude des marqueurs multi-polaires au niveau du texte, Thèse de doctorat, Université Paris-Sud, 2015.

[32] Alec Go et Richa Bhayani et Lei Huang, http://twittersentiment.appspot.com/, Visité le 15-03-2019.

[33] Mohamed Ali Sghaier et Houssem Abdellaoui et Rami Ayadi et Mounir Zrigui, Analyse de sentiments et extraction des opinions pour les sites e-commerce : application sur la langue arabe, CITALA, 2014, 57-61.

[34] J. Chiquet, Validation croisée pour le choix de paramère de méhodes, Module MPR - option modélisation, 2009.

[35] Salima Mdhaffar et Fethi Bougares et Yannick Esteve et Lamia Hadrich-Belguith, Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments, Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), Valencia, Spain, 2017, 55-61.

[36] contact *Bayoudhi et al.* to obtain a copy of the OCA sentence level segmented corpus.

[37] Mohamed Aly et Amir Atiya, LABR: Large Scale Arabic Book Reviews Dataset, Meetings of the Association of Computational Linguistics (ACL), Sofia, Bulgaria, August 2013.

[38] Abdeljalil Elouardighi et Mohcine Maghfour et Hafdalla Hammia et Fatima-Zahra Aazi, Analyse des sentiments à partir des commentaires Facebook publiés en Arabe standard ou dialectal marocain par une approche d'apprentissage, 18ème édition de la conférence Internationale sur l'Extraction et la Gestion des Connaissances, Paris, France,2018, 329-334.

[39] Justin Littman, Collecte de données Facebook avec l'API Graph, https://gwu-libraries.github.io/sfm-ui/posts/2018-01-02-facebook, Visité le 06-04-2019.

[40] Shahzad Qaiser, Ramsha Ali, Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents, International Journal of Computer Applications, 2018, pages 25-29.

[41] Nora Al-Twairesh et Hend Al-Khalifa et AbdulMalikAlsalman et Yousef Al-Ohali, Sentiment Analysis of Arabic Tweets: Feature Engineering and A Hybrid Approach, arXiv, 2018, 1-9.

[42] Eshrag Refaee et Verena Rieser, Subjectivity and Sentiment Analysis of Arabic Twitter Feeds with Limited Resources, In Workshop on Free/Open Source Arabic Corpora and Corpora Processing Tools Workshop Programme, 2014, 2268- 2273.

[43] Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Al-Salman, AraSenTi: Large-Scale TwitterSpecific Arabic Sentiment Lexicons, In Proceedings of the 54th Annual Meeting on Association for Computational Linguistics Berlin Germany Association for Computational Linguistics, 2016.

[44] Nora Al-Twairesh et Hend Al-Khalifa et AbdulMalik AlSalman et Yousef Al-Ohali, AraSenTiTweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets, In Procedia Computer Science, 2017, 63-72.

[45] Muhammad Abdul-Mageed, Mona Diab et Sandra Kübler, SAMAR: Subjectivity and sentiment analysis for Arabic social media, Computer Speech & Language, 28(1), 2014, 20-37.

[46] Chih-Chung Chang et Chih-Jen Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27, 2011.

[47] Amira Shoukry et Ahmed Rafea, Sentence-Level Arabic Sentiment Analysis, SoMNet, 2012, 2-5.

[48] Nawaf A. Abdulla et Nizar A. Ahmed et Mohammed A. Shehab et Mahmoud Al-Ayyoub, Arabic Sentiment Analysis: Lexicon-based and Corpus-based, AEECT, 2013,1-6.

[49] M'hamed Mataoui et Omar Zelmati et Madiha Boumechache, A Proposed Lexicon-Based Sentiment Analysis Approach for the Vernacular Algerian Arabic, Research in Computing Science, 2016, 55-68.

[50] Adrien Sieg, Text Similarities: Estimate the degree of similarity between two texts, https://medium.com/@adriensieg/text-similarities-da019229c894, Visité le 25-04-2019.

[51] Chouikha, La langue arabe son histoire son originalité et son influence, http://www.agoravox.fr/actualites/religions/article/la-langue-arabe-son-histoire-son-77459, Visité le 02/04/2019.

[52] MANSOUR Abd El Fattah et TRAD Aissa, La détection automatique du discours abusif offensant et obscène dans le dialecte algérien, Mémoire de Master, Université El-Oued, 2018.

[53] La langue arabe, https://www.espacefrancais.com/la-langue-arabe/#La-richesse-de-la-langue-arabe, Visité le 18/04/2019.

[54] Samir Abu-Absi, The Arabic Language, https://historyofislam.com/contents/the-modern-age/the-arabic-language/, Visité le 28/04/2019.

[55] Wafia Adouane et Simon Dobnik, Identification of Languages in Algerian Arabic Multilingual Documents, Proceedings of the Third Arabic Natural Language Processing Workshop, 2017, 1-8.

[56] Rehab M. Duwairi, Raed Marji, Narmeen Sha'ban et Sally Rushaidat, Sentiment Analysis in Arabic Tweets, 5th International Conference on Information and Communication Systems (ICICS), 2014, 1-6.

[57] Presentation du language Python, http://www.linux-center.org/articles/9812/python.html, Visité le 15/05/2019.

[58] Mohammad Salameh et Saif M. Mohammad et Svetlana Kiritchenko, Arabic Sentiment Analysis and Cross-lingual Sentiment Resources , https://saifmohammad.com/WebPages/ArabicSA.html Visité le 30-04-2019.