République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Borj Bou Arréridj
Faculté des Mathématiques et d'Informatique
Département d'informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en informatique

Spécialité : Technologies de l'information et de la communication

THÈME

Analyse des sentiments algériennes sur les agitations et mouvements sociaux sur twitter

Présenté par :

MILOUDI Issam

Soutenu publiquement le : 27/06/2022

Devant le jury composé de:

Président : MOHDEB Djamila.

Examinateur : FILLALI Ferhat.

Encadreur : LAIFA Meriem.

2021/2022

Résumé

L'Algérie a récemment connu d'importants bouleversements sociétaux, qui ont entraîné plusieurs changements. Malgré le fait que l'Algérie ait déjà connu des troubles sociaux, les événements les plus récents sont uniques en ce sens qu'ils ont été capturés sous une forme ou une autre via les médias sociaux.Ce mouvement social (appelé Hirak) a été diffusé sur les réseaux sociaux. Diverses plateformes de médias sociaux, telles que Twitter, ont été utilisées par les internautes pour partager leurs pensées et leurs points de vue, allant de favorables à négatifs à neutres. Le but de cette étude était d'examiner les attitudes et les tweets sur la révolution algérienne du monde entier Utilisation d'algorithmes pour la classification.Une base de données de 3697 tweets ont été utilisé, répartis en 1793 commentaires positifs, 232 commentaires négatifs et 1672 commentaires neutres. Les résultats expérimentaux ont montré que le meilleur classificateur est le bidirectionnelle à long-court terme et avec une précision raisonnable égale à 68%.

Mots clés : Mouvements sociaux, traitement automatique du langage, hirak algérien, appel à l'action.

Abstract

Algeria has recently experienced major societal upheavals, which have led to several changes. Despite the fact that Algeria has experienced social unrest before, the most recent events are unique in that they were captured in one form or another via social media. This social movement (called Hirak) was broadcasted on social networks. Various social media platforms, such as Twitter, have been used by netizens to share their thoughts and views, ranging from supportive to negative to neutral. The purpose of this study was to examine attitudes and tweets about the Algerian revolution from around the world using algorithms for classification (both CNNs and RNNs). A database of 3697 tweetswas used, divided into 1793 positive comments, 232 negative comments and 1672 neutral comments. The experimental results showed that the best classifier is the bidirectional long-short term and with a reasonable accuracy equal to 68%.

Keywords: Social movements, automatic language processing, Algerian hirak, call to action.

الملخص

شهدت الجزائر مؤخرًا اضطرابات مجتمعية كبيرة أدت إلى العديد من التغييرات. على الرغم من حقيقة أن الجزائر شهدت اضطرابات اجتماعية من قبل ، إلا أن الأحداث الأخيرة فريدة من نوعها حيث تم التقاطها بشكل أو بآخر عبر وسائل التواصل الاجتماعي. تم بث هذه الحركة الاجتماعية (المسماة الحراك) على الشبكات الاجتماعية. تم استخدام العديد من منصات الوسائط الاجتماعية ، مثل التويتر ، من قبل مستخدمي الإنترنت لمشاركة أفكارهم ووجهات نظرهم ، والتي تتراوح من الداعمة إلى السلبية إلى الحيادية. الغرض من هذه الدراسة هو فحص المواقف والتغريدات حول الثورة الجزائرية من جميع أنحاء العالم باستخدام الخوارزميات في التصنيف. في قاعدة بيانات تضم 3697 تغريدة مقسمة إلى 1793 تعليقًا إيجابيًا و 232 تعليقًا سلبيًا و 1672 تعليقًا محايدًا. أظهرت النتائج التجريبية أن أفضل مصنف هو المدى الطويل قصير المدى ثنائي الاتجاه وبدقة معقولة تساوي 68٪.

الكلمات المفتاحية: حركات اجتماعية ، معالجة آلية للغة ، حراك جزائري ، دعوة للعمل.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

API L'interface de programme d'application.

CNN Réseaux de neurones convolutifs.

RNN Réseaux de neurones récurrents.

LSTM Longue mémoire à court terme.

SVM Machines à Support Vectorielle

NB Naïve Bayes

DT Arbre de décision

RL Régression logistique

Introduction générale

L'action collective fait référence à une action entreprise ensemble par un groupe de personnes, dont le but est d'améliorer leur condition et d'atteindre un objectif commun. C'est un terme qui a des formulations et des théories dans de nombreux domaines des sciences sociales, notamment la psychologie, la sociologie, l'anthropologie, les sciences politiques et l'économie [1].

Un mouvement social est une tentative vaguement structurée par un grand nombre des personnes d'atteindre un certain objectif, généralement d'importance sociale ou politique [2]. Il peut s'agir de mettre en œuvre, de résister ou d'inverser un changement social. C'est une sorte d'action collective dans laquelle des personnes, des organisations ou les deux sont impliquées.

Internet a joué un rôle important dans l'évolution des actions de mobilisation depuis plus de dix ans. En effet, les technologies de communication telles que Twitter, Facebook, Telegram et Tor ont modifié le rythme, la portée et l'efficacité de la communication par le mouvement. L'utilisation de ces technologies a accompagné les révolutions et les transitions démocratiques à tel point que des termes tels que « révolution Twitter », « révolution Facebook » et « révolution 2.0 » ont été utilisés depuis le printemps arabe, malgré le manque de preuves scientifiques de leur rôle dans ces mouvements.

Pour la première fois, lors de la révolution orange en Ukraine en 2004, les téléphones portables et Internet ont joué un rôle central dans la coordination des révolutionnaires [3]. Par exemple, l'organisation de la société civile Pora ! Tente de mobiliser et d'éclairer les jeunes en utilisant Internet pour lutter contre la censure gouvernementale et les SMS [4] pour coordonner les mobilisations et diffuser des informations liées aux élections.

Les manifestations algériennes de 2019-2021, six jours après qu'Abdelaziz Bouteflika a annoncé sa candidature à un cinquième mandat présidentiel dans un communiqué signé. Ces manifestations, sans précédent depuis la guerre civile algérienne, étaient pacifiques et ont conduit les militaires à exiger la démission immédiate de Bouteflika, qui a eu lieu le 2 avril 2019.

La montée des tensions au sein du régime algérien remonte au début du régime de Bouteflika qui a été caractérisé par le monopole de l'État sur les revenus des ressources naturelles utilisés pour financer le système clientéliste du gouvernement et assurer sa stabilité. [5] Les grandes manifestations ont eu lieu dans les plus grands centres urbains d'Algérie de février à décembre 2019. En raison de leur ampleur considérable, les manifestations ont attiré une couverture médiatique internationale et provoqué des réactions de plusieurs chefs d'État et personnalités savantes.

L'objectif principal de ce projet est d'apprendre comment mener un projet complet d'analyse de données, de la collecte de données à la conclusion et à la prise de décisions.

Le plan du rapportest structuré comme suit :

Introduction générale

On a donné une définition de l'action collective et le mouvement social, après le Hirak algérien, à la fin nous avons expliqué le rôle des médias sociaux dans les mouvements sociaux.

Le premier chapitre : Revue de littérature

On a donné une définition d'analyse des sentiments et de ses niveaux, après on a exprimé leur besoin et son importance, et aussi on a parlé sur ces approches et les domaines d'applications, on terminer le chapitre par les travaux connexes.

Le deuxième chapitre : Méthodologie

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

Le troisième chapitre : Résultats et Discussion

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

Conclusion générale

Nous avons clôturé petit résumé du projetet les défis et les travaux futurs possibles.

Chapitre 01 : Revue de la littérature

1. Introduction

La pratique consistant à obtenir et à évaluer les pensées, les idées et les perceptions des gens sur divers sujets, biens, sujets et services est connue sous le nom d'analyse des sentiments, ce qui rend difficile la perception des émotions et le choix de la bonne polarisation des émotions. L'analyse des sentiments utilise le traitement du langage naturel et l'exploration de texte pour identifier et extraire des informations subjectives du texte.

2. Analyse des sentiments

2.1.Définition

L'analyse des sentiments est l'utilisation du traitement du langage naturel, de l'analyse de texte, de la linguistique computationnelle et de la biométrie pour identifier, extraire, quantifier et étudier systématiquement les états affectifs et les informations subjectives. L'analyse des sentiments est largement appliquée à la voix des documents clients tels que les avis et les réponses aux enquêtes, les médias en ligne et sociaux, et les documents de santé pour des applications allant du marketing au service client en passant par la médecine clinique [6].

2.2.Les niveaux [7]

Lors de l'analyse des sentiments, la première étape consiste à définir le texte qui sera analysé dans le cas d'une certaine étude. En général, il existe trois niveaux d'analyse : le niveau du document (niveau du message ou niveau du document), le niveau de la phrase (niveau de la phrase) et le niveau de l'aspect (niveau de l'entité et de l'aspect). La Figure 1 représenteles niveaux d'analyses des sentiments.

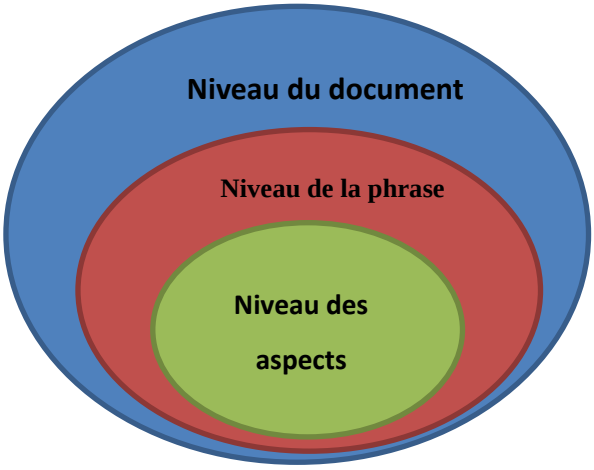

Figure 1:les niveaux d'analyses des sentiments [7]

2.2.1. Niveau du document

Déterminer la polarité d'un texte entier L'hypothèse est que le texte n'exprime qu'un point de vue sur une seule entité (par exemple, un seul produit).

2.2.2. Niveau de la phrase

Déterminer la polarité de chaque phrase incluse dans un texte L'hypothèse est que chaque phrase du texte exprime un point de vue distinct sur une entité distincte.

2.2.3. Niveau des aspects

Effectue une analyse plus fine que les autres niveaux. Il est basé sur l'idée qu'une opinion consiste d'un sentiment et une cible (d'opinion). Par exemple, la phrase «L'iPhone est très bon, mais il faut encore travailler sur la durée de vie de la batterie et les problèmes de sécurité» évalue trois aspects : iPhone (positif), la durée de vie de la batterie (négatif) et la sécurité (négative).

3. Besoin d'analyse des sentiments [8]

L'analyse des sentiments est extrêmement importante car elle permet aux entreprises de mieux comprendre les sentiments de leurs clients à l'égard de leur marque. L'analyse des sentiments est un terme qui décrit les méthodes et les tactiques que les entreprises utilisent pour analyser les données sur la façon dont leurs clients se sentent à propos d'un certain service ou produit.

L'analyse des sentiments est une procédure qui analyse automatiquement les énoncés en langage naturel, identifie les déclarations ou points de vue clés et les catégorise en fonction de leur attitude émotionnelle.

- Grâce à l'amélioration des produits, à l'identification des problèmes en temps réel et à l'unicité du marché, l'analyse des sentiments a accru la satisfaction des consommateurs vis-à-vis des besoins de l'entreprise.
- La satisfaction des consommateurs est mesurée via une analyse des sentiments, dans laquelle un client décrit son expérience avec un produit et exprime ses pensées et ses sentiments à ce sujet à l'aide de commentaires en langage naturel. Cela nous donne des informations essentielles pour savoir si le client est satisfait du produit et, si nécessaire, comment nous pouvons l'améliorer.
- Identifiez et répondez aux problèmes en temps réel : Un consommateur peut instantanément exprimer son mécontentement au monde entier via les médias sociaux.

4. Tâche d'analyse des sentiments

La figure 2 représente les différentes taches d'analyse des sentiments :

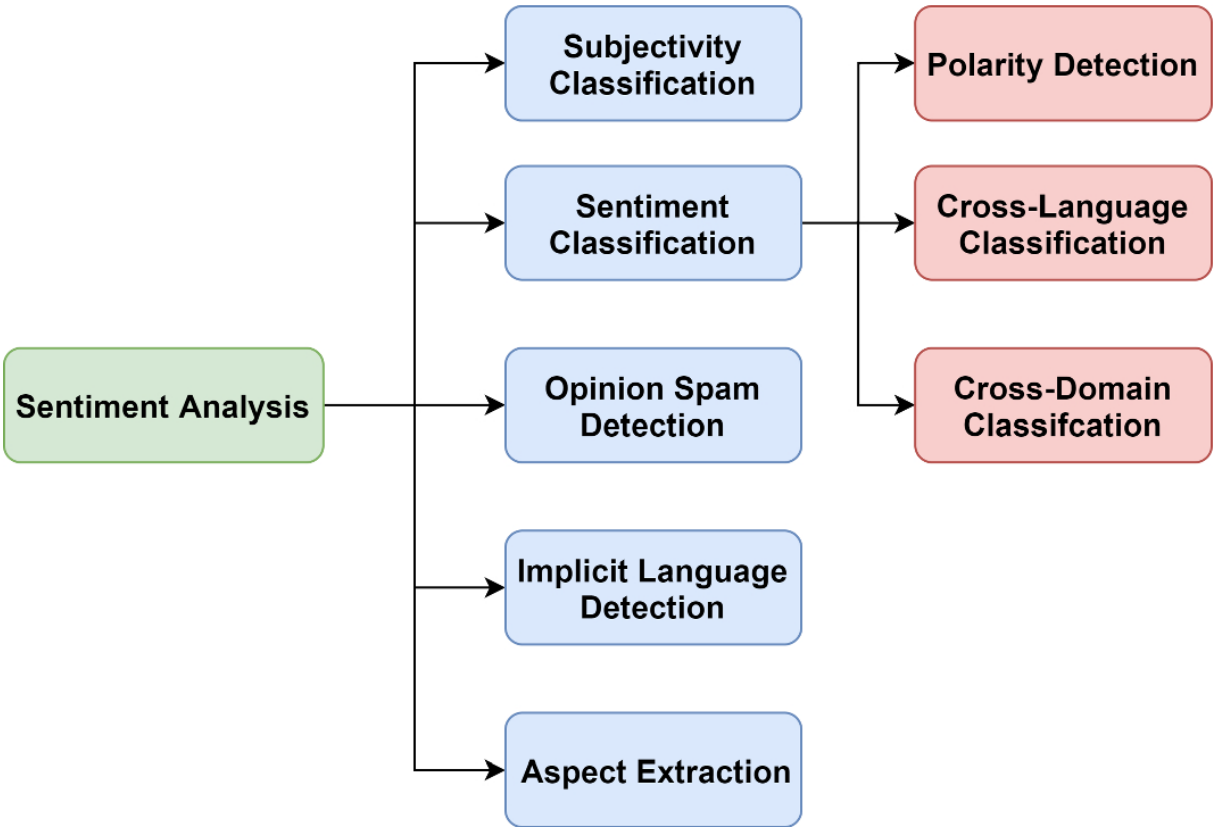

Figure 2: Différentes tâches d'analyse des sentiments [8]

4.1. Classification de la subjectivité

On suppose souvent qu'il s'agit de la première étape du processus d'analyse des sentiments. La catégorisation de la subjectivité reconnaît les indices subjectifs, les expressions émotionnelles et les idées subjectives [9]. Ces indicateurs sont utilisés pour distinguer les objets textuels objectifs et subjectifs. La catégorisation de la subjectivité vise à exclure les éléments objectifs de données défavorables du traitement ultérieur [10].

4.2. Classification des sentiments

L'une des sous-tâches de la classification des sentiments, et le terme « Analyse d'opinion » est fréquemment utilisé pour désigner l'analyse des sentiments. Il s'agit d'une petite tâche qui vous demande de comprendre comment chaque morceau de texte vous fait ressentir. Traditionnellement, la polarité est soit positive soit négative. Le contexte de l'opinion est investigué, les composantes intra-opinion et inter-opinion étant finement définies [11].

L'extraction des caractéristiques invariantes spécifiques au domaine et l'emplacement où elles sont distribuées est une approche largement utilisée. L'analyse inter-langues s'effectue de manière similaire, en entraînant le modèle sur un ensemble de données issues d'une langue source, puis en l'évaluant sur un ensemble de données issues d'une langue différente avec des données limitées.

4.3. Détection de spam d'opinion

En raison de la popularité croissante du commerce électronique et des sites d'avis, la détection des spams est devenue une difficulté majeure dans l'analyse des sentiments. La détection de spam d'opinion recherche trois aspects clés dans un avis téléphonique : le texte, les métadonnées et la connaissance réelle du produit. Quelques métadonnées utilisées pour identifier les avis de spam sont les notes par étoiles ou par points, l'adresse IP de l'utilisateur, la géographie et d'autres informations [12].

4.4. Sarcasme de détection de langage implicite

Les langages implicites sont parfois appelés ironie et comédie. Même pour les humains, détecter ce type de communication équivoque et ambiguë peut être difficile. Cependant, le langage implicite est une partie importante d'une déclaration qui peut changer radicalement son sens et sa polarité. Les moyens plus traditionnels d'identifier le langage implicite incluent

la recherche de signaux tels que les émoticônes, les émotions de rire et l'utilisation fréquente des signes de ponctuation [13].

4.5. Extraction d'aspect

Les trois phases de l'analyse des sentiments au niveau des aspects sont l'extraction des aspects, la classification de la polarité et l'agrégation. L'extraction d'aspect est l'une des procédures fondamentales de l'analyse des sentiments basée sur les aspects, car elle la distingue de l'analyse des sentiments traditionnelle [14]. Étant donné qu'un nombre considérable de façons sont utilisées plus fréquemment que d'autres, les mots clés les plus couramment utilisés sont plus susceptibles d'apparaître comme des aspects ; cette stratégie simple peut s'avérer être une approche assez forte. De plus, cette stratégie peut négliger des sujets qui ne sont généralement pas énoncés [15].

5. Approche de l'analyse des sentiments [7]

Il existe plusieurs méthodes mettre en œuvre des systèmes d'analyse des sentiments, que l'on peut classer comme suit :

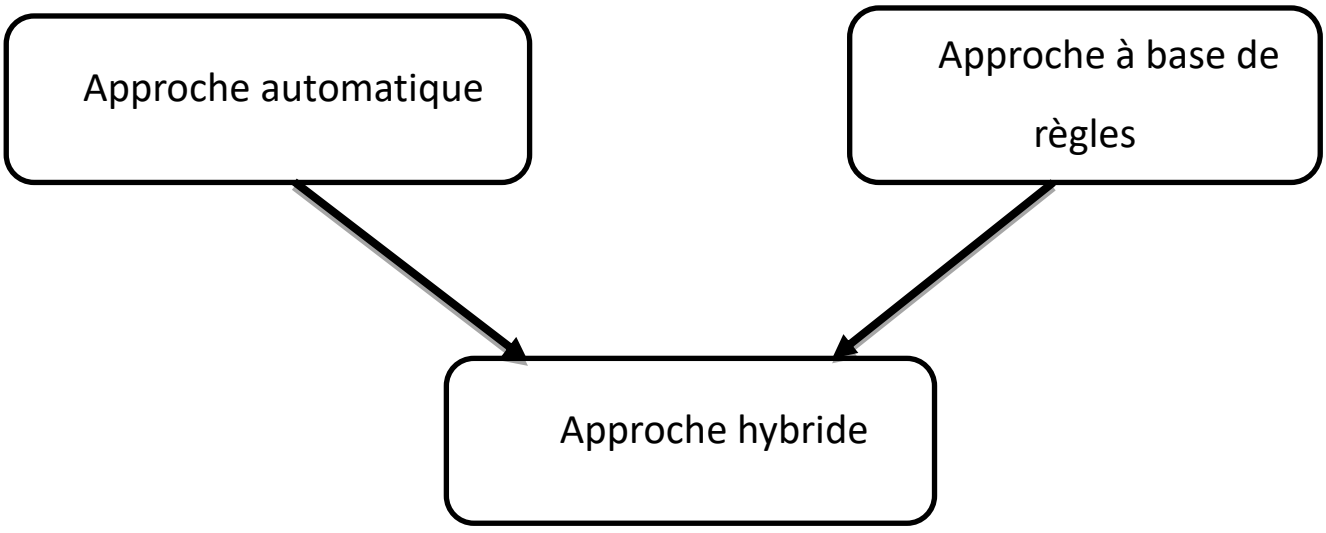

Figure 3:les approches d'analyses des sentiments [7]

5.1. Approche automatique

Les techniques d'apprentissage automatique sont utilisées dans les approches automatiques. Le travail d'analyse des sentiments est généralement décrit comme un problème de classification, dans lequel un classificateur reçoit un texte et renvoie la catégorie appropriée, telle que positive, négative ou neutre (en cas d'analyse de polarité).

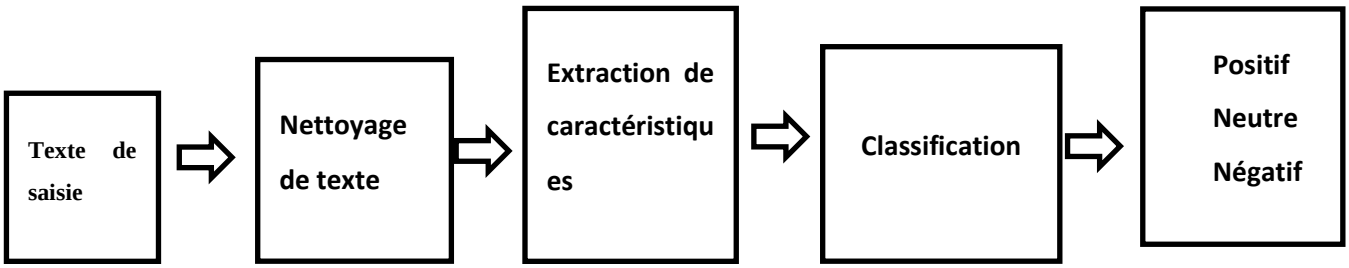

Figure 4:Approche automatique [7]

5.2. Approche à base de règles

La technique basée sur des règles (ou approche lexicale) détecte la subjectivité, la polarité ou le sujet d'une opinion en définissant un ensemble de règles dans un langage informatique (script). Cette méthode peut être utilisée avec une variété d'entrées, y compris les approches NLP traditionnelles telles que les racines, la tokenisation, le marquage POS et la segmentation. Ils utilisent le dictionnaire des sentiments avec des termes d'opinion et les associent à des faits pour déterminer la polarité dans d'autres procédures basées sur le lexique.

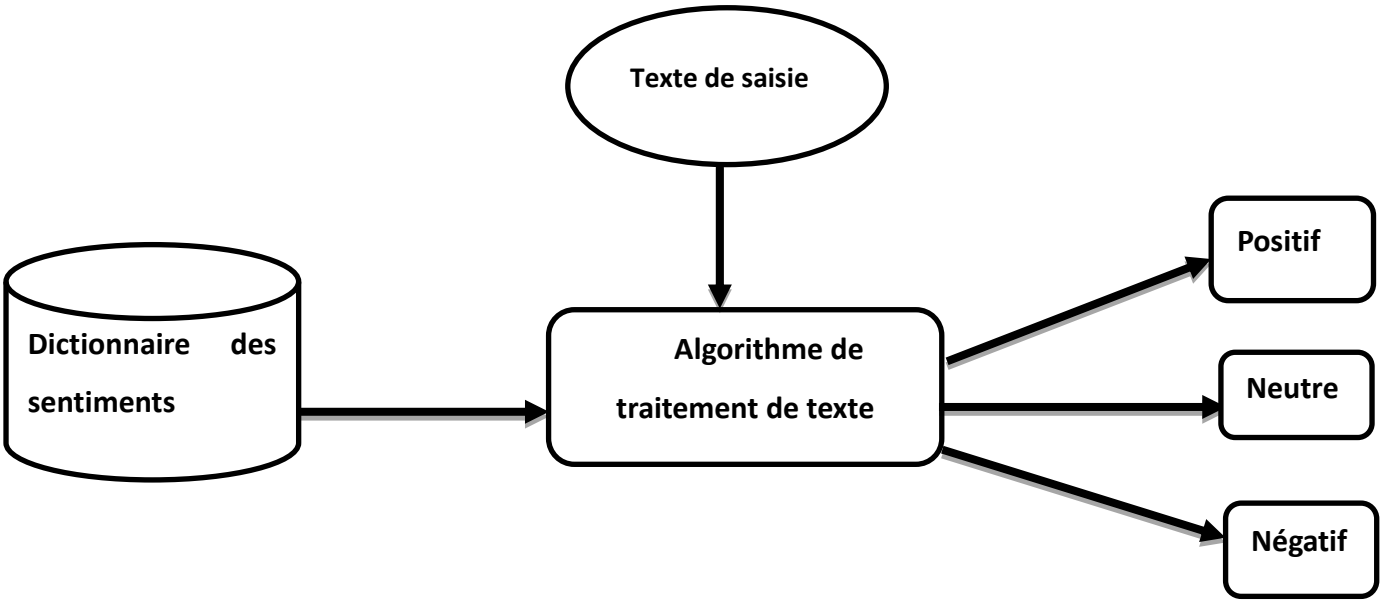

Figure 5:Approche de base de règles [7]

5.3. Approche hybride

Les solutions hybrides sont simples à comprendre : il suffit de combiner le meilleur des deux approches, basée sur des règles et automatisée. En général, les stratégies peuvent améliorer la précision en combinant les deux approches.

6. Domaines d'applications [7]

L'analyse des sentiments est importante dans une variété de domaines, et diverses applications ont évolué dans ce contexte. Quelques utilisations sont brièvement mentionnées ci-dessous :

6.1. Politique

De nos jours, les politiciens ont adopté la tendance de l'analyse des sentiments, dans laquelle ils recherchent les opinions des utilisateurs de médias sociaux sur la législation proposée avant de la promulguer. Lors d'une élection présidentielle, connaître l'avis des internautes sur un homme politique est assez important.

6.2. Éducation

L'analyse des sentiments peut être utilisée pour glaner des informations sur les méthodes d'enseignement d'un enseignant ainsi que sur le matériel de cours. Il détermine le niveau d'apprentissage d'un élève, comprend ses exigences, prévoit ses performances et apporte des modifications de style réussies. Les enseignants et les écoles peuvent utiliser les résultats de l'analyse des sentiments pour prendre des mesures correctives.

6.3. Économie

La majorité des acheteurs recherchent des recommandations sur un produit ou un service avant de l'acheter, et ils sont même prêts à payer plus pour un produit qui a une opinion favorable plus élevée qu'un autre, ce qui pourrait améliorer les ventes. Les entreprises peuvent utiliser l'analyse des sentiments pour savoir ce que leurs consommateurs pensent de leurs produits ou services. Dans le but d'améliorer leurs produits et d'augmenter les ventes et les profits.

6.4. Sanitaire et Médical

Pour améliorer les services de santé, cette application aide les professionnels de la santé à collecter et à évaluer l'humeur des patients, les épidémies, les mauvaises réactions aux médicaments et les maladies. [16] Ont souligné la difficulté d'employer l'analyse des sentiments dans les soins de santé en raison de la terminologie spécialisée et unique du domaine. [17] Ont utilisé les tweetsTwitter sur les expériences des patients comme complément à leur analyse de santé publique. En utilisant l'API Streaming de Twitter, ils ont généré plus de cinq millions de tweets liés au cancer du sein au cours d'une année. Les tweets ont été classés à l'aide d'un classificateur LR typique et d'un modèle CNN après prétraitement. Les résultats positifs du traitement, l'obtention d'un soutien communautaire et la sensibilisation du public étaient tous liés. Enfin, l'utilisation de l'analyse des sentiments pour examiner les données générées par les patients sur les réseaux sociaux peut être bénéfique.

7. Travaux connexes

Non seulement en termes de vie sociale, mais aussi en termes de commerce électronique, d'apprentissage en ligne et de politique, les réseaux sociaux jouent un rôle important dans notre quotidien chargé.

Ajay Band et Aziz Fellah [18] ont créé le Socio-Analyzer après avoir étudié les changements sociétaux actuels dans le mouvement #MeToo. Ils ont utilisé leur méthodologie en quatre phases pour mettre en œuvre le socio-analyseur dans cette enquête. Les données ont été détectées et triées en trois catégories par ce socio-analyseur : positives, négatives et neutres. Les résultats de cette enquête révèlent que la majorité des gens ont un point de vue neutre. [18] ont confirmé que les 765 tweets de données #MeToo généralisent les résultats aux données météorologiques basées sur ces découvertes. Lorsqu'elles sont jugées positives pour les tweets neutres, les valeurs de précision de Socio-Analyzer et TextBlob sont respectivement de 70,74 % et 72,92 %.

L'analyse des sentiments a été utilisée par Matalon, Magdaci et ses collègues [19] pour prévoir le renversement d'opinion dans les tweets de communication politique entre Israël et la Palestine. En utilisant un modèle d'apprentissage appelé l'arbre aléatoire, les chercheurs ont pu trouver 7147 appariements source-citation. Avec un ROC-AUC de 0,83, ce modèle prédit si cette source subira l'inversion d'opinion du retweeteur en fonction des aspects de traitement du langage naturel du texte source et des attributs de l'utilisateur. Selon les résultats, environ

80% des critères qui expliquent l'inversion d'attitude sont liés aux sentiments des messages originaux sur la question. D'autre part, environ 14 % des pairs citant une source sont liés à l'humeur de la source.

Rushdi-Saleh, M., Martín-Valdivia et autres a suggéré l'Opinion Corpus for Arabic (OCA) pour les critiques arabes recueillies à partir de pages Web liées au cinéma. Ce corpus a été converti en anglais pour créer l'EVOCA (English Version of Opinion Corpus for Arabic), un corpus d'opinion pour les anglophones[20]. À l'aide d'une validation croisée de 10 fois, le système suggéré a été évalué sur une variété de techniques d'apprentissage automatique, y compris les machines à vecteurs de support (SVM) et Nave Bayes (NB). Les résultats ont montré que l'EVOCA était plus pauvre que l'OCA dans cet essai, mais qu'EVOCA est toujours équivalent aux approches anglaises [20].

Md Hassan Zamir [21] a mené des recherches sur les stratégies de communication utilisées sur Twitter par le mouvement Shahbag et les résidents bangladais. L'étude a trouvé deux motivations communes pour les manifestants de Shahbag : les diffuseurs qui retweetent fréquemment et ont un haut degré de sortie, et les récepteurs dont les tweets sont fréquemment retweetés et ont un haut degré d'entrée. Où L'objectif principal de cette étude est d'apprendre comment les manifestants en réseau utilisent les médias sociaux pour échanger des informations et interagir lors de mouvements sociaux en ligne.

SenZi, le premier lexique d'analyse des sentiments pour le dialecte arabizi libanais, a été présenté parTobailil, Fernandez1 et al. [22]. Pour produire ce lexique, [22] a suivi une série de procédures, en commençant par construire, traduire, annoter et traduire diverses ressources pour arriver à une collection de départ de termes de sentiment 2K. Ils ont augmenté le nombre de mots de sentiment à 24 600 pour cette enquête. À la suite de cette recherche, un nouveau lexique arabizi a été créé, composé de 11,3 millions de mots positifs et de 13,3 millions de mots négatifs, avec un score F1 de 0,72.

Ces dernières années, les médias sociaux ont permis de s'épanouir et d'accéder via Internet, c'est le résultat du développement des technologies de l'information et de la communication, et aussi de la façon dont les gens communiquent en examinant les connexions Twitter lors du #METOO display où ils ont pu mesurer la contribution individuelle de ces groupes en fonction de la probabilité de retweet [23].

L'un des sujets les plus populaires ces dernières années est l'analyse des sentiments via l'apprentissage automatique à l'aide de données Twitter. Le problème est d'analyser les sentiments lors d'événements critiques, où ils ont analysé les sentiments sur deux ensembles de données en espagnol : tremblement de terre au Chili 2010 et indépendance de la Catalogne 2017 avec les classificateurs du réseau Bayes, les résultats ont également montré une efficacité Les résultats sont prédictifs et les réseaux résultants permettent de déterminer la relation entre les mots [24].

Dans cet article, une analyse des médias sociaux a été menée pour tirer une conclusion à la fois sur l'intensité émotionnelle et l'impact des médias sociaux sur cette veillée, où ce modèle a utilisé le pouvoir d'apprentissage d'un réseau avec une mémoire à long terme pour prédire la possibilité de retweeter, puis des expériences sur (Hirak et Brexit) ont été collectées à partir de Twitter. Là où les sentiments affectaient le contexte de l'événement, d'autre part, un groupe de sentiments a été trouvé pour améliorer la capacité prédictive. [25].

Les médias sociaux fournissent des plateformes flexibles qui jouent un rôle clé dans la dynamisation de l'action collective dans des mouvements comme le Printemps arabe et Occupy Wall Street. En permettant aux individus d'afficher largement leurs émotions, les médias sociaux amplifient les sentiments définis comme une émotion collective partagée pour alimenter les forces qui conduisent au changement dans la société. Nous constatons que la force des liens sociaux formés par les échanges de messages et de commentaires influence la participation, mais son effet diffère selon les deux mouvements[26].

8. Conclusion

De nos jours, l'analyse des sentiments a gagné encore plus de valeur avec l'avènement des réseaux sociaux, captant l'intérêt des chercheurs, des journalistes, des entreprises et des gouvernements. Bien que de nombreuses parties soient confrontées à des défis d'analyse des sentiments, ceux-ci ne sont pas très difficiles à surmonter avec les bonnes solutions et les bons partenaires de collaboration.

Chapitre 02 : Méthodologie

1. Introduction

La tâche d'analyse des sentiments des tweets écrits en arabe est le sujet de cette recherche. Pour atteindre cet objectif et obtenir les meilleures performances potentielles, nous avons utilisé la méthodologie présentée ci-dessous, qui commence par la conception globale de notre processus, puis passe à la conception spécifique.

2. Description détaillé de l'objectif

Notre objectif principal est d'examiner les données du Hirak algérien à l'aide de méthodes de classification. À l'aide de l'apprentissage automatique et de l'apprentissage en profondeur, le processus d'analyse des sentiments examine les commentaires Twitter ou « tweets » pour déterminer s'ils ont des sentiments positifs, négatifs ou neutres. La méthode d'analyse des sentiments comporte de nombreuses étapes. Nous commençons par la phase de collecte et d'annotation des données, puis passons à la phase de prétraitement et enfin à la phase de mise en œuvre.

3. La conception détaillée du système

3.1.La collection des données

La première étape de la procédure d'analyse des sentiments consiste à rassembler les tweets en entrant un mot-clé et en récupérant tous les tweets qui incluent ce terme. Les tweets peuvent provenir de diverses sources. Une méthode consiste à utiliser un robot d'exploration Twitter, qui utilise le service Web Twitter pour rassembler une collection de tweets pertinents.

Afin de catégoriser les tweetsa été collecté par d'anciens étudiants, ils seront conservés dans une base de données. Les données ont déjà été collectées et enregistrées dans un fichier CSV nommé "01_11_2019 - tous_les_donnée.csv" dans notre exemple. Il y avait 5450 tweets dans cette collecte de données.

	Unnamed: 0	user	Date	Text	retweets
0	0	tjrs_positif	2019-10-31 23:41:56+00:00	Pas de célébration officielle cette année, le ...	0
1	1	MerHadjer23	2019-10-31 23:28:12+00:00	...وعقدنا العزم أن تحيا الجزائر فاشهدوا... فاشهدو	1
2	2	rabia_hakim	2019-10-31 23:28:12+00:00	...تسقط_انتخابات_العصابات #تسقط_انتخابات_العصابات#	0
3	3	tjrs_positif	2019-10-31 23:23:22+00:00	Nous serons là tous ensemble demain pour réaff...	0
4	4	Gostoland	2019-10-31 23:14:48+00:00	...تسقط_انتخابات_العصابات #حراك_1_نوفمبر إنّ م#	0
...	...	...	...	...	...
5445	15	MerzougTouati	2019-10-20 21:41:14+00:00	... نداء النائب المستقيل من البرلمان خالد تازاغرت	0
5446	16	1vYAEVVlEjjS3mz	2019-10-12 19:02:15+00:00	...سيكون هناك تجمع غدا في ساحة اول نوفمبر على الس	14
5447	19	Dzhour16	2019-10-10 17:55:15+00:00	...إضراب عام وطني مرتقب يومي 30 و 31 أكتوبر متبوع	0
5448	20	Dzhour16	2019-10-10 17:54:28+00:00	...إضراب عام وطني مرتقب يومي 30 و 31 أكتوبر متبوع	0
5449	23	Dzhour16	2019-10-10 17:53:09+00:00	...إضراب عام وطني مرتقب يومي 30 و 31 أكتوبر متبوع	0

5450 rows × 5 columns

Figure 6: Collection des données.

3.2.La division de la base de données

Nous avons séparé la base de données en deux parties en fonction de la langue. Une base de données contient la langue arabe, tandis que la deuxième base de données contient la langue française. Pour ce faire, nous avons conçu la fonction "CheckLanguage". Ce dernier a été utilisé dans la colonne "Texte" pour déterminer si ce tweet était rédigé en arabe ou en français. Cette phase génère deux fichiers CSV nommés "Arabic.csv" et "French.csv".

	Unnamed: 0	user	Date	Text	retweets	Type
0	1	MerHadjer23	2019-10-31 23:28:12+00:00	...وعقدنا العزم أن تحيا الجزائر فاشهدوا... فاشهدو	1.0	A
1	2	rabia_hakim	2019-10-31 23:28:12+00:00	...تسقط_انتخابات_العصابات #تسقط_انتخابات_العصابات#	0.0	A
2	4	Gostoland	2019-10-31 23:14:48+00:00	...تسقط_انتخابات_العصابات #حراك_1_نوفمبر إنّ م#	0.0	A
3	8	tjrs_positif	2019-10-31 23:06:23+00:00	VIGILANCE #تسقط_انتخابات_العصابات# حراك_1_نوفمبر	0.0	A
4	9	AissaMilano	2019-10-31 22:55:29+00:00	...نوفمبر_الإستقلال #حراك_1_نوفمبر #تسقط_انتخابا#	0.0	A
...	...	...	...	...	...	...
3883	4676	SMakri	2019-10-25 14:20:23+00:00	... الله أكبر أول نوفمبر الله أكبر راه جاي نوفمبر	1.0	A
3884	4677	SMakri	2019-10-25 14:07:02+00:00	...الله أكبر أول نوفمبر من شعارات مسيرة اليوم #ال	1.0	A
3885	4678	MerzougTouati	2019-10-20 21:41:14+00:00	... نداء النائب المستقيل من البرلمان خالد تازاغرت	0.0	A
3886	4679	1vYAEVVlEjjS3mz	2019-10-12 19:02:15+00:00	...سيكون هناك تجمع غدا في ساحة اول نوفمبر على الس	14.0	A
3887	4680	Dzhour16	2019-10-10 17:55:15+00:00	...إضراب عام وطني مرتقب يومي 30 و 31 أكتوبر متبوع	0.0	A

3888 rows × 6 columns

Figure 7: Base qui contient des tweets en Arabe.

	Unnamed: 0	Unnamed: 0.1	Unnamed: 0.1.1	user	Date	Text	retweets	Type
0	0	0	0	tjrs_positif	2019-10-31 23:41:56+00:00	Pas de célébration officielle cette année, le ...	0	F
1	3	3	3	tjrs_positif	2019-10-31 23:23:22+00:00	Nous serons là tous ensemble demain pour réaff...	0	F
2	5	5	5	tjrs_positif	2019-10-31 23:13:40+00:00	Nuit blanche à Alger, et l'emblème Amazigh flo...	0	F
3	6	6	6	hesalert	2019-10-31 23:10:04+00:00	Demain je sortirai. Je me battrai, pour elle, ...	2	F
4	7	7	7	tjrs_positif	2019-10-31 23:09:10+00:00	Les sons du mortier (مهراز) font vibrer la Ca...	0	F
...	...	...	...	...	...	...	...	...
788	4551	5296	2163	AliBenflis2019	2019-10-02 02:48:19+00:00	تحيا الجزائر يحيا الشعب يحيا علي بن فليس https...	3	F
789	4614	5372	56	arkouby	2019-10-30 14:43:41+00:00	نداءات للتعبئة قبيل جمعة نوفمبر الكبرى https:/...	0	F
790	4626	5385	71	athbouyahia	2019-10-29 15:53:52+00:00	...الأغنية الجديدة التي أحدثت ضـ جة اليوم و سيغني	0	F
791	4646	5409	109	AlgeriaTimes	2019-10-26 14:06:33+00:00	... زيتوت : الشعب http://AlgeriaTimes.net Algeria	0	F
792	4671	5437	7	Abdelrrahmane	2019-10-30 22:15:55+00:00	موقفي من مسيرة اول نوفمبر https://www.facebook...	0	F

793 rows × 8 columns

Figure 8: Base qui contient des tweets en Français.

3.3.Annotation

Le but de l'étape d'annotation est d'attribuer une étiquette de polarité (positive, négative ou neutre) à chaque message qui exprime son sentiment. Nous recommandons d'utiliser la méthode manuelle pour annoter les différents commentaires obtenus à partir de la base de données. Selon la variété d'AlgD(dialecte algérien) en raison de divers domaines et dialectes.

	user	Date	Text	retweets	Type	Annotation
0	MerHadjer23	2019-10-31 23:28:12+00:00	...وعقدنا العزم أن تحيا الجزائر فاشهدوا... فاشهدو	1.0	A	positive
1	rabia_hakim	2019-10-31 23:28:12+00:00	...تسقط_انتخابات_العصابات #تسقط_انتخابات_العصابا#	0.0	A	positive
2	Gostoland	2019-10-31 23:14:48+00:00	...تسقط_انتخابات_العصابات #حراك_1_نوفمبر إنّ م#	0.0	A	positive
3	tjrs_positif	2019-10-31 23:06:23+00:00	حراك_1_نوفمبر #تسقط_انتخابات_العصابات#VIGILANCE	0.0	A	neutre
4	AissaMilano	2019-10-31 22:55:29+00:00	...نوفمبر_الإستقلال #حراك_1_نوفمبر #تسقط_انتخابا#	0.0	A	neutre
...	...	...	...	...	...	...
3692	SMakri	2019-10-25 14:20:23+00:00	... الله أكبر أول نوفمبر الله أكبر راه جاي نوفمبر	1.0	A	positive
3693	SMakri	2019-10-25 14:07:02+00:00	...الله أكبر أول نوفمبر من شعارات مسيرة اليوم #ال	1.0	A	positive
3694	MerzougTouati	2019-10-20 21:41:14+00:00	... نداء النائب المستقيل من البرلمان خالد تازاغرت	0.0	A	positive
3695	1vYAEVVIEjjS3mz	2019-10-12 19:02:15+00:00	...سيكون هذاك تجمع غدا في ساحة اول نوفمبر على الس	14.0	A	positive
3696	Dzhour16	2019-10-10 17:55:15+00:00	...إضراب عام وطني مرتقب يومي 30 و 31 أكتوبر متبوع	0.0	A	positive

3697 rows × 6 columns

Figure 9:Annotations des tweets.

4. Application du processus d'analyse des sentiments

4.1.Le prétraitement

Il s'agit de la phase la plus importante de notre processus d'analyse des sentiments. Pour produire un modèle de classification puissant avec un meilleur score, un ensemble de données propres doit être fourni au modèle d'apprentissage automatique avant de commencer la

catégorisation des messages en positif, négatif ou neutre. Parce qu'ils sont rédigés par des personnes de différents niveaux intellectuels, la majorité des commentaires recueillis sur les réseaux sociaux sont improvisés et déroutants, avec des écarts sémantiques et grammaticaux. De plus, la diversité de cette base de données se traduit par une diversité de termes et de sens pour un même mot selon les localités. De plus, nous avons découvert d'autres problèmes dans les commentaires, tels que des fautes d'orthographe, la présence de liens, hashtags, gifs, autocollants, caractères spéciaux, etc., qu'il faut supprimer immédiatement et ces messages dupliqués afin d'avoir une image minimale et unifiée. Contenu, corpus valide et propre prêt à être exploité. Pour enrichir notre corpus, nous souhaitons conserver autant de variations de vocabulaire informatif que possible. Dans ce qui suit nous détaillerons les étapes de prétraitement :

- Supprimer les lignes vides.
- Supprimer les signes diacritiques.
- Remplacer les ponctuations par un espace.
- Supprimer les mots en double consécutifs.
- Supprimer les lettres non arabes.
- Supprimer les émoticônes.
- Suppression d'adresse (URL).
- supprimer les hashtags.
- Suppression de tiret 8 (_).
- Supprimer les répétitions de lettres plus d'une.
- Normalisation des caractères arabes.
- supprimer les mots vides (stop Words).
- supprimer le caractère non arabe.

	user	Date	retweets	Type	Annotation	Text
0	MerHadjer23	2019-10-31 23:28:12+00:00	1	A	positive	وعقدنا العزم تحيا الجزاءر فاشهدوا الاستقلال حر...
1	rabia_hakim	2019-10-31 23:28:12+00:00	0	A	positive	تسقط انتخابات العصابات
2	Gostoland	2019-10-31 23:14:48+00:00	0	A	positive	تسقط انتخابات العصابات حراك موعدهم الصيح اليس ...
3	tjrs_positif	2019-10-31 23:06:23+00:00	0	A	neutre	حراك تسقط انتخابات العصابات
4	AissaMilano	2019-10-31 22:55:29+00:00	0	A	neutre	الاستقلال حراك تسقط انتخابات العصابات
...	...	...	...	...	...	...
3692	SMakri	2019-10-25 14:20:23+00:00	1	A	positive	اله اكبر راه جاي شعارات مسيره اليوم الجمعه الجمعه
3693	SMakri	2019-10-25 14:07:02+00:00	1	A	positive	اله اكبر شعارات مسيره اليوم الجمعه الجمعه
3694	MerzougTouati	2019-10-20 21:41:14+00:00	0	A	positive	نداء الناءب المستقيل البرلمان خالد تازاغرت الش...
3695	1vYAEVVlEjjS3mz	2019-10-12 19:02:15+00:00	14	A	positive	سيكون تجمع ساحه اول الساعه العاشره صباحا ستليه...
3696	Dzhour16	2019-10-10 17:55:15+00:00	0	A	positive	اضراب عام وطني مرتقب يومي متبوع باكبر مسيره مل...

3697 rows × 6 columns

Figure 10: Résultat après le prétraitement

Le nuage de mots est la représentation graphique des mots les plus fréquemment répétés représentant la taille du mot [27]. Pour créer notre nuage de mot, nous avons utilisé trois bibliothèques sont :

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

Feature extraction

L'extraction de caractéristiques est une tâche concernant la transformation de données brutes en entrées appropriées (c'est-à-dire des caractéristiques) qui peuvent être consommées par un algorithme d'apprentissage automatique particulier. De manière expresse, les caractéristiques extraites doivent représenter le contenu textuel principal dans un format qui correspondra le mieux aux besoins de l'algorithme de classificateur sélectionné. Pour faire l'extraction des caractéristiques. Nous avons adopté deux méthodes sont [28] :

La méthode Sac des Mots

Sac des mots est une méthode utilisée dans le traitement du langage naturel et la recherche d'informations. Dans cette méthode chaque Test soit (phrase/ commentaire / tweet ou bien un document) est représenté sous la forme d'un vecteur numérique où chaque dimension est un mot spécifique du corpus et la valeur peut être une fréquence dans le document, une occurrence (notée 1 ou 0) ou même des valeurs pondérées [29].

La méthode TF-IDF

La méthode TF-IDF (Term Frequency-Inverse Document Frequency) est un processus appliqué dans l'extraction de texte et la recherche d'informations. Ce processus TF-IDF est centré sur un mot de la collection ou du corpus de documents. Il représente la durée pendant laquelle un mot apparaît dans un document, le numéro du document avec ce mot particulier et le rapport entre les documents avec ce terme par tous les documents. Cette méthode est utilisée dans la suppression des mots d'arrêt, des mots haute fréquence et basse fréquence. Il est également utilisé dans la synthèse et la classification de texte [30].

4.2.La classification

Dans cette phase, nous avons basées sur la classification supervisée, à partir du corpus annoté et traité. Pour accomplir cette tâche nous avons implémenté les algorithmes de classification supervisée les plus populaires comme les systèmes neurones CNN et RNN et les classificateurs classiques SVM, NB, DT, RL.

Naïve Bayes

Naïve Bayes est une technique d'apprentissage automatique utilisée pour classer le texte dans des catégories prédéfinies en fonction de caractéristiques similaires. Le classificateur Naïve Bayes a été appliqué pour améliorer le traitement et la manipulation de textes ou d'informations provenant de différentes sources. Cet algorithme représente une méthode probabiliste. En d'autres termes, le classificateur Naïve Bayes suppose que l'absence de caractéristique de classe n'est pas liée à l'absence d'autres caractéristiques. Ce classificateur est couramment utilisé pour classer les documents en raison d'une bonne performance de classification, calcule la probabilité des documents liés à les classer dans différentes classes, puis les attribue à la classe spécifique avec la probabilité la plus élevée [31].

Machines à Support Vectorielle

Machines à Support Vectorielle est l'un des modèles d'apprentissage supervisé qui ont été appliqués pour classification du texte. Il classe les différents objets et documents dans un espace de dimension finie. La machines à support vectorielle est également utilisé pour analyser des données, des textes et des documents afin de calculer la similitude entre eux. Ce model montre différents aspects utiles en tant que modèle important utilisé en informatique. Premièrement, cette méthode se défend sur une algèbre linéaire, où elle ne contient aucune équation algébrique complexe. Parmi l'un des avantages de machines à support vectorielle est l'efficacité des poids attribués aux concepts ou aux termes. Ce modèle montre également une facilité particulière par rapport à d'autres méthodes [31].

Arbre de décision

Un arbre de décision est une classification supervisée approche et construit à partir de noeuds qui représentent des cercles et les branches sont représentées par les segments qui relient les noeuds. Un arbre de décision commence à la racine, se déplace vers le bas et est généralement dessiné de gauche à droite. Le noeud à partir duquel l'arbre commence est appelé noeud

racine. Le noeud où se termine la chaîne est appelé noeud « feuille ». Deux branches ou plus peuvent être étendues à partir de chaque noeud interne, c'est-à-dire un noeud qui n'est pas un noeud feuille. Un noeud représente une certaine caractéristique tandis que les branches représentent une plage de valeurs. Ces plages de valeurs agissent comme des points de partition pour l'ensemble de valeurs de la caractéristique donnée. Le regroupement des données dans l'arbre de décision est basé sur les valeurs des attributs des données. Cet arbre de décision est réalisé à partir des données pré-classifiées [32].

Régression logistique

La régression logistique est l'un des classificateurs les plus célèbres dans les mondes de la statistique, de la science des données et de l'apprentissage automatique. Pour les données de faible dimension, la régression logistique est une approche standard pour la classification binaire. Cela est particulièrement vrai dans les domaines scientifiques tels que la médecine, la psychologie et les sciences sociales où l'accent est mis non seulement sur la prédiction mais aussi sur l'explication. Il existe également la version multinomiale de la régression logistique qui peut être utilisée pour modéliser des réponses non binaires (multi-catégories) [33].

4.2.1. Réseaux de neurones convolutifs (CNN)

Les réseaux de neurones convolutifs sont un type spécialisé de réseaux de neurones artificiels qui utilisent une opération mathématique appelée convolution à la place de la multiplication matricielle générale dans au moins une de leurs couches. Ils sont spécifiquement conçus pour traiter les données de pixels et sont utilisés dans la reconnaissance et le traitement d'images [34].La figure 11 représente notre modèle de CNN :

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 300, 100)          2000000

 dropout (Dropout)           (None, 300, 100)          0

 conv1d (Conv1D)             (None, 296, 128)          64128

 max_pooling1d (MaxPooling1D  (None, 59, 128)          0
 )

 dropout_1 (Dropout)         (None, 59, 128)           0

 batch_normalization (BatchN  (None, 59, 128)          512
 ormalization)

 conv1d_1 (Conv1D)           (None, 55, 128)           82048

 max_pooling1d_1 (MaxPooling  (None, 11, 128)          0
 1D)

 dropout_2 (Dropout)         (None, 11, 128)           0

 batch_normalization_1 (Batc  (None, 11, 128)          512
 hNormalization)

 flatten (Flatten)           (None, 1408)              0

 dense (Dense)               (None, 128)               180352

 dense_1 (Dense)             (None, 3)                 387
```

Figure 11: modèle de CNN

4.2.2. Réseaux de neurones récurrents (RNN)

Les RNN sont un type de réseau neuronal artificiel dans lequel les connexions des nœuds forment un graphe orienté dans un ordre séquentiel. Il s'agit essentiellement d'une chaîne de pièces de réseau neuronal réunies. Chacun envoie un message au suivant.Étant donné que le texte est généralement séquentiel, cette conception permet à RNN de démontrer un comportement temporel et de collecter des données séquentielles, ce qui en fait une méthode plus "naturelle" pour traiter les données textuelles [35].

4.2.3. LSTM

La mémoire longue à court terme (LSTM) est un type de réseau de neurones artificiels utilisé dans l'apprentissage en profondeur et l'intelligence artificielle. LSTM comporte des connexions de rétroaction, contrairement aux réseaux de neurones à anticipation normaux. Ce

type de réseau neuronal récurrent peut gérer non seulement des points de données uniques (comme des photos), mais également des séquences de données complètes (telles que la parole ou la vidéo). LSTM peut être utilisé pour des applications telles que l'identification d'écriture manuscrite liée non segmentée, la reconnaissance vocale, la traduction automatique, le contrôle de robots, les jeux vidéo et les soins de santé, par exemple. Le LSTM est le réseau de neurones le plus mentionné du XXe siècle [36]. La figure 12 représente notre modèle de LSTM :

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 embedding_1 (Embedding)      (None, 300, 100)          2000000

 lstm (LSTM)                  (None, 100)               80400

 dense_2 (Dense)              (None, 3)                 303
```

Figure 12: modèle LSTM

4.2.4. Bidirectionnel LSTM

La mémoire bidirectionnelle à long-court terme (LSTM bidirectionnelle) est la méthode permettant à tout réseau de neurones de stocker des informations de séquence dans les directions arrière (futur vers passé) et avant (passé vers futur).

La figure 12 représente notre modèle Bidirectionnel LSTM :

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 embedding_2 (Embedding)      (None, 300, 100)          2000000

 bidirectional (Bidirectiona  (None, 300, 20)           8880
 l)

 bidirectional_1 (Bidirectio  (None, 20)                2480
 nal)

 dense_3 (Dense)              (None, 10)                210

 dense_4 (Dense)              (None, 3)                 33
```

La figure 12 représente la différence entre les deux architectures :

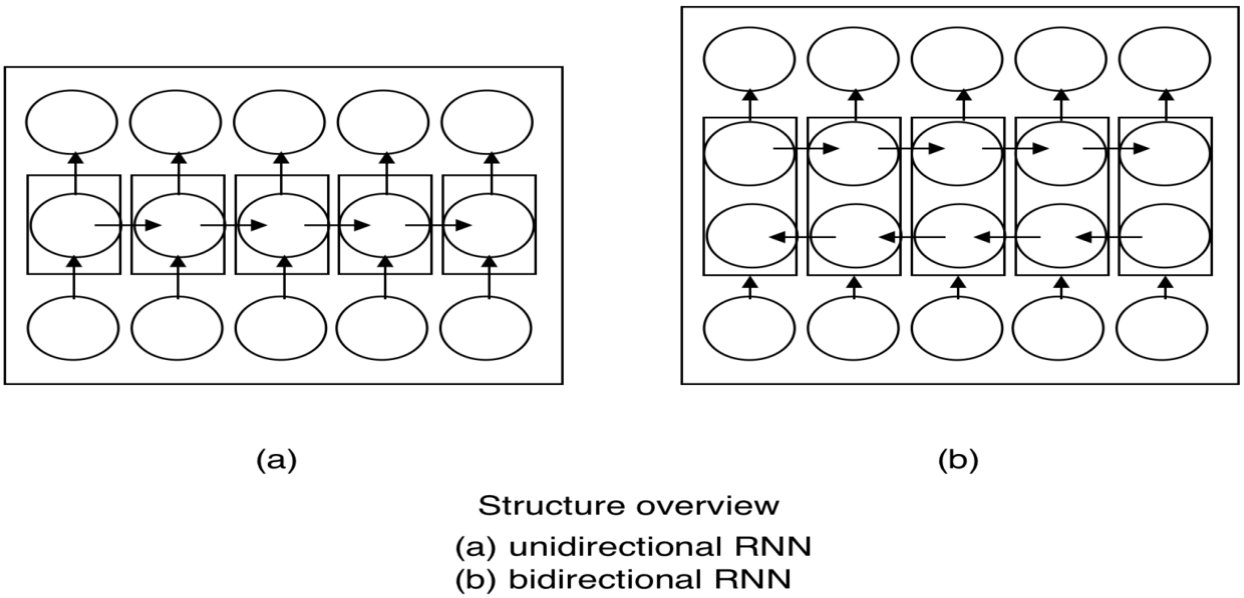

Figure 13: architecture de RNN et bidi RNN [37]

5. Évaluation

Nous avons examiné les performances des nombreux modèles d'apprentissage automatique que nous avons utilisés dans notre projet tout au long de cette phase. Nous avons utilisé de nombreux critères tels que la précision, le rappel, l'exactitude et le score f1 pour compléter cette étape. Celles-ci sont basées sur la matrice de confusion, qui est une matrice. La matrice de confusion [38], également connue sous le nom de table de contingence ou matrice d'erreur, est utilisée pour afficher le résultat de prédiction du classificateur. Il s'agit d'un tableau spécifique pour visualiser les performances du modèle, qui est présenté dans le tableau 1.

Tableau 1: Matrice de confusion

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

- Vrai Positive (TP) : nombre de tweets positifs classés correctement.
- Faux positive (FP) : nombre de tweets négatifs classés à tort comme positifs.
- Vrai Négative (TN) : nombre de tweets négatifs classés correctement.
- Faux Négative (FN) : nombre de tweets positifs classés à tort comme négatifs.

5.1. Précision

La précision [39] est également appelée valeur prédite positive, mesure la justesse du modèle. Une précision plus élevée indique moins de FP. Mathématiquement, il est défini comme :

$$\text{La précision} = TP / TP+FP$$

5.2. Rappel

Le rappel [39] est également connu sous le nom de sensibilité, mesure les cas positifs correctement classés par le modèle, une valeur de rappel élevée signifie que peu de cas positifs sont mal classés comme négatifs. Le rappel peut être calculé à l'aide de la formule suivante.

Le rappel= TP/TP+FN

5.3. Le score F1

Le score F1 ou la mesure F1 [39] est la moyenne harmonique de la précision et du rappel. Le score F peut être calculé comme suit :

Score F1= 2∗ La précision× Le rappel / La précision +Le rappel

5.4. Exactitudes

L'exactitude utilisée comme mesure pour les techniques de catégorisation. Les valeurs d'exactitudes, cependant, sont beaucoup moins réticentes aux variations du nombre de décisions correctes que la précision et le rappel. Cette technique est représentée sous la forme suivante [39]:

L'exactitude=$TP+TN/TP+FP+TN+FN$

6. Conclusion

Pour atteindre les meilleures performances potentielles, il faut utiliser une bonne méthodologie pour appliquer les processus d'analyse des sentiments.

Chapitre 03 : Résultats et Discussion

1. Introduction

Nous passerons en revue les outils matériels et logiciels dont nous avions besoin pour mener à bien ce projet dans ce chapitre. De plus, nous passerons en revue une analyse exploratoire des données sur les résultats que nous avons recueillies en profondeur.

2. L'environnement de travail et les outils utilisés

2.1. L'environnement Matériel

On a utilisé un pc condor équipé d'un processeur Intel(R) Core (TM) i5-3320M avec RAM de 8 GB.

2.2. L'environnement Logiciel

Nous avons utilisé le langage de programmation Python, version 3.10, pour atteindre notre objectif. Python est un langage raisonnablement simple à apprendre, open source, gratuit et interprété qui est récemment devenu le langage de choix des informaticiens. Guido Var Rossum travaille sur ce dernier depuis 1989.

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

- **Package pandas** pour la manipulation et l'analyse des données.
- **Package Numpy** pour manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions.
- **Package Matplot** pour tracer et visualiser des données sous formes de graphiques.
- **Package CSV** pour enregistrer les données dans un format tabulaire.
- **Packages sys, os, re, csv, codecs.**
- **Package NLTK** pour le traitement automatique des langues.
- **Package Seaborn** pour dessiner des statistiques attrayantes et informatives.
- **Package Sklearn** pour l'apprentissage automatique.
- **Package Keras** pour fournit une interface Python pour les réseaux de neurones artificiels. Keras agit comme une interface pour la bibliothèque TensorFlow.

- **Package Tensorflow** pour l'apprentissage automatique et l'intelligence artificielle.

3. Présentation des données

Nous avons utilisé une base de données avec quatre caractéristiques et des enregistrements d'internautes sur le mouvement social algérien pour faire l'analyse des sentiments.

Après, on a ajouté deux colonnes «Annotation » et «type». Les deux colonnes qu'on a concentré c'est «Text» et « Annotation». Le tableau 2 présente une description sur les attributs dans notre base de données.

Tableau 2:Description des attributs de la base de données utilisée.

Attributs	Description
Utilisateur	Il représente l'identificateur de chaque utilisateur.
Date	C'est la date et l'heure auxquelles le tweet a été tweeté par l'utilisateur.
Texte	Il se compose des avis donnés par chaque utilisateur individuel.
Retweets	Il présente le nombre de fois que ce Tweet a été retweeté par l'utilisateur.
Annotation	Il contient des sentiments positifs, négatifs ou neutres.
type	Il sépare le texte en arabe ou bien en français.

Selon les idées exprimées par chaque utilisateur individuel, la caractéristique "Annotation" est classée en trois classes pour notre projet. Cette division produit trois types d'attributs de sentiment : positif, négatif et neutre.

Selon les idées exprimées par chaque utilisateur individuel, la caractéristique "Annotation" est classée en trois classes pour notre projet. Cette division produit trois types d'attributs de sentiment : positif, négatif et neutre. Il y a 1793 avis positifs, 232 avis négatifs et 1672 avis neutres dans l'ensemble de donnéescomme illustré dans la figure 13.

```
<AxesSubplot:xlabel='Annotation', ylabel='count'>
```

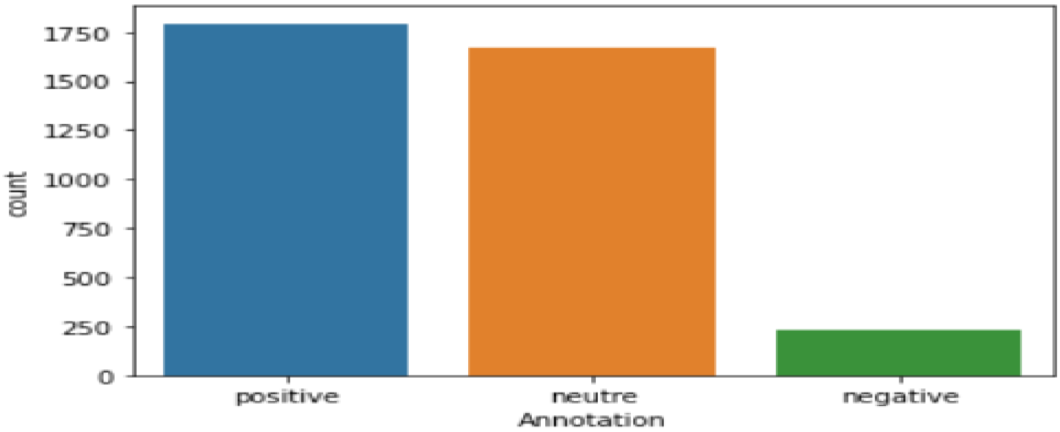

Figure 14:La répartition des avis positifs, négatifs et neutres.

Nous avons utilisé de nombreux packages dans cette partie, comme nous l'avons dit au chapitre 3, pour supprimer tous les bruits tels que les hashtags, les mentions et les ponctuations. De l'autre côté, nous avons supprimé les mots vides afin d'avoir un texte simple. Nous allons nous concentrer sur le Word Cloud dans cette étape. Ce dernier est une représentation graphique des termes les plus fréquemment utilisés dans les tweets. En général, lorsque des mots sont utilisés ou populaires, ils sont affichés dans les tailles et poids de police les plus significatifs.

Figure 15: Le nuage de mots.

Le nuage de mots est une technique de visualisation de données utilisée pour représenter des données textuelles dans lesquelles la taille de chaque mot indique sa fréquence ou son importance. Les points de données textuels significatifs peuvent être mis en évidence à l'aide d'un nuage de mots.

Comme le montre la figure 14, elle nous montre des mots comme(تحيا الجزائر/ تسقط انتخابات/ انتخابات العصابات/ الاستقلال) dans de grandes tailles, ce qui montre que la majorité des tweets contenaient ces mots, qui à leur tour étaient avec le mouvement, nous constatons donc que la plupart des tweets sont positifs dans l'annotation.

4. Classification

Les réseaux de neurones fonctionnent avec des couches et les manipuler en fait un bon modèle.Et après plusieurs tentatives sur les modèles et manipulation de couches, nous avons choisi les meilleurs modèles pour les réseaux de neurones convolutifs (RNN) et les réseaux de neurones récurrents (CNN).Nous avons représenté également les résultats les méthodes classiques de classification SVM, RL, DT et NB avec TF-IDF et BOW

Les résultats des CNN et RNNsont ordonnés dans le tableau 3:

Tableau 3: Le résultat d'exactitude des classificateurs

	Positive			Neutre			Négative			Accurecy
	Précision	Recall	F1	Précision	Recall	F1	Précision	Recall	F1	
Bidi LSTM	0.65	0.77	0.71	0.73	0.65	**0.68**	0.44	0.18	0.28	0.68
LSTM	0.67	**0.73**	**0.70**	**0.74**	**0.64**	**0.68**	**0.36**	**0.42**	**0.39**	0.67
CNN	0.62	0.66	0.64	0.62	0.63	0.63	0.25	0.09	0.13	0.61

Pour les résultats qui nous avons obtenunous notons que la précision est élevée dans la classe neutre avec le classificateur LSTM par rapport les autres classes et les autres classificateurs

avec une valeur égale à 0.74.D'une autre part nous notons que le recall est élevée dans la classe positive avec le classificateur bidirectionnel LTSM avec une valeur égale à 0.77.Pour le F1-score, le F1-score est élevé dans la classepositive aussi avec le classificateur bidirectionnel LTSM avec une valeur égale 0.71.

Les résultats ont obtenu à l'aide des classificateurs SVM, RL, DT et NB pour classer les sentiments en classes positive, négative et neutre. Pour chaque classificateur nous avons appliqués les deux méthodes d'extraction des attributs qui sont : le sac de mots (BOW) et le TF-IDF. Les Tableaux 4 et 5 représentent les résultats d'exactitude des classificateurs utilisant le TF-IDF et le BOW respectivement.

Tableau 4: Le résultat d'exactitude des classificateurs avec le TF-IDF.

classificateur	accuracy	Positif			Neutre			Négatif		
		Précision	Recall	F1	Précision	Recall	F1	Précision	Recall	F1
SVM	**0.68**	0.68	0.76	0.72	0.70	0.69	0.70	0.55	0.14	0.23
NB	**0.69**	0.69	0.76	0.72	0.70	0.71	0.72	0	0	0
LR	**0.69**	0.68	0.81	0.73	0.73	0.67	0.70	0.50	0.02	0.05
DT	**0.64**	0.66	0.68	0.67	0.63	0.68	0.65	0.47	0.19	0.27

Pour les résultats qui nous avons obtenu avec le TF-IDF, nous notons que la précision est élevée dans la classe neutre avec le classificateur LR par rapport les autres classes et les autres classificateurs avec une valeur égale à 0,73.D'une autre part nous notons que le recall est élevée dans la classepositive avec le même classificateur LR avec une valeur égale à 0,81.Pour le F1-score, le F1-score est élevé dans la classe positive avec le même classificateur LR aussi avec une valeur égale à 0,73.

Tableau 5: Le résultat d'exactitude des classificateurs avec le BOW.

classificateur	accuracy	Positif			Neutre			Négatif		
		Précision	Recall	F1	Précision	Recall	F1	Précision	Recall	F1
SVM	**0.66**	0.70	0.64	0.67	0.63	0.76	0.69	0.58	0.17	0.26
NB	**0.69**	0.69	0.77	0.73	0.71	0.71	0.71	1.00	0.01	0.02
LR	**0.67**	0.71	0.67	0.69	0.64	0.76	0.69	0.78	0.17	0.27
DT	**0.65**	0.70	0.63	0.66	0.63	0.75	0.68	0.54	0.25	0.34

Ce tableau représente le résultat qui nous avons obtenue avec le BOW, nous notons que la précision est élevée dans la classe négative avec LR et valeur égale 0.78.D'une autre part nous notons que le recall est élevée dans la classe positive avec NB et valeur égale 0.77.Pour le F1-score, le F1-score est élevé dans la classe positive avec le même classificateur NB et valeur égale 0.73.

5. Comparaison

5.1.Pour les réseaux neurones

Nous avons découvert que le bidirectionnelLSTM fournissait la meilleure catégorisation sur la base des résultats que nous avons recueillis. Cependant, comme nous avons sélectionné les meilleurs modèles pour chaque approche, les résultats sont clairement proches. De plus, les côtes de précision sont médiocres, ce qui indique que la catégorisation n'est pas assez précise. Une explication pourrait être que nos données sont déséquilibrées. Les tweets positifs et neutres sont plus courants que les tweets négatifs. Un autre facteur qui peut avoir contribué au faible taux de précision est les difficultés que nous avons rencontrées lors du prétraitement du dialecte algérien. La manipulation du dialecte algérien est un domaine académique nouveau et en plein essor.

Les cellules LSTM (Long Short Term Memory), qui possèdent une mémoire interne appelée cellule. La cellule permet de maintenir un état aussi longtemps que nécessaire. Cette cellule consiste en une valeur numérique que le réseau peut piloter en fonction des situations [40].

Et La mémoire bidirectionnelle à long-court terme (LSTM bidirectionnelle) est la méthode permettant à tout réseau de neurones de stocker des informations de séquence dans les directions arrière (futur vers passé) et avant (passé vers futur) comme on dit dans la définition de LSTM bidirectionnelle dans deux couches Comme le montre la figure 12 c'est la cause que ce dernier mieux performé que LSTM et CNN.

5.2.Pour les méthodes classiques de classification

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

6. Conclusion

Dans le monde d'aujourd'hui, l'analyse des sentiments est un sujet fascinant et précieux. Cependant, peu d'études ont été réalisées sur les textes mixtes algériens, qui, selon nous, nécessitent des méthodologies uniques pour bien fonctionner dans notre environnement. Afin d'améliorer les résultats de la catégorisation des sentiments, des études supplémentaires sont nécessaires.

Conclusion générale

Depuis le début de son mouvement social en 2019, l'Algérie a connu beaucoup de changements. L'importance de ce mouvement social (surnommé le Hirak) nous a obligés à l'enquêter à l'aide des médias sociaux. L'objectif de cette étude, en particulier, était d'utiliser des algorithmes de notation pour évaluer le sentiment des tweets algériens liés au Hirak. Le langage de programmation Python a été utilisé pour mener à bien le projet.

Cette étude comprenait des étapes spécifiques de l'approche que nous avons utilisée, ainsi que des projets de recherche associés et des explications sur les éléments essentiels de traitement du langage naturel et d'analyse des sentiments.

Nous avons fait face à plusieurs défis et limites. Pour commencer, l'annotation manuelle des tweets était un processus chronophage, laborieux et plutôt subjectif influencé par nos propres préférences. Cela pourrait avoir influencé le taux de précision des algorithmes de catégorisation.

Nous proposons les méthodes réalisables suivantes pour améliorer ce travail et surmonter les contraintes décrites ci-dessus.

- Utilisez un autre type d'annotation, comme des annotations automatisées ou externalisées.

- Utiliser d'autres méthodes de classification plus adaptées au traitement des langues et dialectes compliqués.

- Introduire un nouveau mécanisme de prétraitement spécifique au dialecte algérien, qui est combiné avec d'autres langues fréquemment parlées en Algérie, telles que l'anglais et le français.

Références

[1] "collective action problem - collective action". Encyclopædia Britannica.

[2] Scott, John; Marshall, Gordon (2009), "Social movements", A Dictionary of Sociology, Oxford University Press, doi:10.1093/acref/9780199533008.001.0001, ISBN 978-0-19-953300-8, retrieved 2020-03-06

[3] «La Révolution Twitter» des Moldaves » , sur slate.fr, 10 avril 2009

[4] « The Role of Digital Networked Technologies in the Ukrainian Orange Revolution » ,sur harvard.edu, 1er décembre 2007

[5] RasmusAleniusBoserup; Luis Martinez, eds. (2016). Algeria Modern: From opacity to complexity. CERI/Sciences Po. London: Hurst. ISBN 9781849045872.

[6] Hamborg, Felix; Donnay, Karsten (2021). "NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles". "Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume"

[7] Mehdi Hadji : Analyse des sentiments : Généralités. Medium

[8] Wankhade, M., Rao, A.C.S. &Kulkarni, C. A survey on sentiment analysis methods, Applications, and challenges.ArtifIntell Rev (2022).

[8] Kasmuri E, Basiron H (2017) Subjectivity analysis in opinion mining—a systematic literature review. Int J Adv Soft ComputAppl 9(3):133–159

[9] Kamal A (2013) Subjectivity classification using machine learning techniques for mining feature-opinion pairs from web opinion sources. arXiv preprint arXiv:13126962

[10] Wang G, Sun J, Ma J, Xu K, Gu J (2014) Sentiment classification: the contribution of ensemble learning. Decis Support Syst 57:77–93

[11] Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H (2015) Survey of review spam detection using machine learning techniques. J Big Data 2(1):1–24

[12] Fang Z, Zhang Q, Tang X, Wang A, Baron C (2020) An implicit opinion analysis model based on feature-based implicit opinion patterns. ArtifIntell Rev 53(6):4547–4574

[13] Kanapala A, Pal S, Pamula R (2019) Text summarization from legal documents: a survey. ArtifIntell Rev 51(3):371–402

[14] Bai X, Liu P, Zhang Y (2020) Investigating typed syntactic dependencies for targeted sentiment classification using graph attention neural network. IEEE/ACM Trans Audio Speech Lang Process 29:503–514

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

[16] Jiménez-Zafra SM, Martín-Valdivia MT, Molina-González MD, Ureña-López LA (2019) How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain.ArtifIntell Med 93:50–57

[17] Clark EM, James T, Jones CA, Alapati A, Ukandu P, Danforth CM, Dodds PS (2018) A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across twitter. arXiv preprint arXiv:180509959

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

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

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

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

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

[23] Mirbabaie, M., Brünker, F., Wischnewski, M., and Meiner, J., : The Development of Connective Action during Social Movements on Social Media. ACM Transactions on Social Computing, 4(1), pp. 1–21, (2021).

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

[25] A. Drif and K. Hadjoudj,: An Opinion Spread Prediction Model With Twitter Emotion Analysis During Algeria's Hirak. The Computer Journal, 64(3), pp. 358–368, (2021).

[26] A. Al-Hasan, D. Yim, and H. C. Lucas,: A Tale of Two Movements: Egypt during the Arab Spring and Occupy Wall Street. IEEE Transactions on Engineering Management, 66(1) pp. 84–97, (2019).

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

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

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

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

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

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

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

[34] Ian Goodfellow and YoshuaBengio and Aaron Courville (2016).Deep Learning.MIT Press.p. 326.

[35] ShreyaGhelani :Text Classification — RNN’s or CNN’s?. towardsdatascience

[36] Schmidhuber, Jürgen (2021). "The most cited neural networks all build on work done in my labs". AI Blog. IDSIA, Switzerland. Retrieved 2022-04-30.

[37] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." Signal Processing, IEEE Transactions on 45.11 (1997): 2673-2681.2. AwniHannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan

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

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

[40] RomainHerault,Clement Chatelain. Découvrez les cellules à mémoire interne : les LSTM.Openclassrooms.