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

MEMOIRE

Présenté en vue de l'obtention du diplôme

Master en informatique

Spécialité : Ingénierie de l'informatique décisionnelle

THEME

Manipulation des ensembles de données multilingues pour l'analyse des sentiments

Présenté par : **GUEBLA Borhane Eddine**

TAIBI Mouloud

Soutenu publiquement le:

Devant le jury composé de:

Président : **SAIFI Lynda**

Examinateur : **ZOUAOUI Hakima**

Encadreur : **LAIFA Meriem**

2022/2023

Remerciement

Tout d’abord, nous remercions Allah le tout puissant de nous avoir donné le courage et la patience nécessaires à mener ce travail à son terme

Nos remerciements et nos profondes gratitudes vont à notre encadreur **Dr : LAIFA Meriem** pour son encadrement, son suivi et ses conseils tout au long de cette période .

Nous tenons à remercier les membres du jury pour leur précieux temps accordé à l’étude de notre mémoire.

Enfin, nous voudrions remercier toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce mémoire. Votre appui logistique, votre aide technique et vos précieux conseils ont été inestimables.

C'est grâce à vous tous que nous sommes parvenus à mener à bien ce mémoire. Votre contribution a été essentielle et nous vous en sommes profondément reconnaissants.

Dédicace

À nos familles, qui ont été présentes à chaque étape de notre vie et qui ont été une source inépuisable d'amour, de soutien et de compréhension. Votre présence constante et votre confiance en nous ont été d'une importance capitale.

À nos amis, qui ont partagé nos moments de joie, de stress et de doute. Votre amitié sincère, votre soutien indéfectible et vos encouragements nous ont donné la force de persévérer.

Ce mémoire est le fruit de nos efforts collectifs et de l'apport de nombreuses personnes qui ont croisé notre chemin. Nous leur exprimons toute notre gratitude et leur dédions ce travail avec reconnaissance et humilité.

Résumé

L'objectif principal de ce mémoire est de prévoir les manifestations publiques à l'aide d'algorithmes d'apprentissage par transfert, apprentissage automatique et d'attributs recueillis à partir des données des médias sociaux. Nous nous concentrons notamment sur l'affaire "Hirak", qui a débuté en Algérie en février 2019. L'objectif sera également de fournir un modèle de prédiction basé sur des techniques de catégorisation pour prévoir les manifestations majeures en utilisant les informations des canaux de médias sociaux publics de Twitter. Afin d'atteindre les objectifs susmentionnés, nous présentons un aperçu méthodologique de quatre techniques de classifications fondamentales d'apprentissage par transfert, apprentissage automatique et comparons leurs performances dans la catégorisation des tweets de protestation. Dans l'ensemble, nous avons obtenu des performances de classification modérées à bonnes, avec des valeurs d'exactitude allant de 0,50 à 0,75. Cela suggère que les méthodes et les algorithmes utilisés peuvent fournir des résultats significatifs dans la tâche de classification.

Abstract

The main objective of this thesis is to predict public protests using transfer learning , machine learning and attribute algorithms gathered from social media data. We focus in particular on the "Hirak" case, which began in Algeria in February 2019. The aim will also be to provide a prediction model based on categorization techniques to predict major protests using information from Twitter's public social media channels. In order to achieve the above-mentioned objectives, we present a methodological overview of four fundamental transfer learning , machine learning classification techniques and compare their performance in the categorization of protest tweets. Overall, we achieved moderate to good classification performance, with accuracy values ranging from 0.50 to 0.75. This suggests that the methods and algorithms used can provide significant results in the classification task.

الملخص

الهدف الرئيسي من هذه الأطروحة هو التنبؤ بالاحتجاجات العامة باستخدام التعلم الانتقالي والتعلم الآلي وخوارزميات السمات التي تم جمعها من بيانات وسائل التواصل الاجتماعي. نركز بشكل خاص على قضية "الحراك" التي بدأت في الجزائر في فبراير 2019. سيكون الهدف أيضًا هو توفير نموذج تنبؤ يعتمد على تقنيات التصنيف للتنبؤ بالاحتجاجات الكبرى باستخدام معلومات من قنوات التواصل الاجتماعي العامة على Twitter. من أجل تحقيق الأهداف المذكورة أعلاه ، بشكل عام حققنا أداء تصنيف متوسط إلى جيد، مع قيم دقة تتراوح من 0.50 إلى 0.75. يشير هذا إلى أن الأساليب والخوارزميات المستخدمة يمكن أن توفر نتائج مهمة في مهمة التصنيف.

Table des matière

Table des figures

Liste des tableaux

Introduction Générale

Les mouvements sociaux sont connus depuis longtemps pour mettre fin au compromis historique entre la sphère économique et la sphère sociale [1]. Par exemple : le mouvement social de la France face au système de retraite et de protection sociale [2] ou encore les manifestations en Chine pour allégées les restrictions anti-covide [3]. Cependant l'intelligence artificielle est de plus en plus utilisée pour analyser les mouvements sociaux, car elle offre un outil puissant pour mieux comprendre les phénomènes sociaux à grande échelle.

L'IA peut être utilisée pour suivre, analyser et comparer différents mouvements et leurs conversations associées, et pour comprendre la dynamique des interactions entre les participants et peut être utilisée pour extraire des données des médias sociaux afin d'identifier les participants les plus influents, d'identifier les sujets de conversation [4], les sujets les plus populaires et de détecter les tendances dans les conversations, comme elle peut également être utilisée pour identifier les relations entre différents acteurs et pour découvrir des modèles cachés dans les données ou encore pour détecter les anomalies et les valeurs aberrantes dans les données, permettant aux chercheurs d'identifier les risques ou opportunités potentiels dans le contexte des mouvements sociaux.

En tirant parti de l'IA, les chercheurs peuvent mieux comprendre la dynamique des mouvements sociaux, ce qui leur permet de développer des stratégies plus efficaces pour y répondre l'une de ces stratégies c'est l'analyse des sentiments qui est utilisée afin de mieux comprendre les opinions et les émotions des individus.

On a pris comme exemple d'étude le hirak algérien qui est un mouvement de protestation qui a débuté en février 2019. L'objectif principal du mouvement est de provoquer une réforme démocratique et la justice sociale en Algérie. Le mouvement a largement réussi à attirer l'attention sur les problèmes de corruption et de violations des droits de l'homme dans le pays et a été crédité d'avoir joué un rôle important dans la démission de l'ancien président Abdelaziz Bouteflika et d'autres membres du régime. Le mouvement a été alimenté en grande partie par les médias sociaux, les militants les utilisant pour organiser des manifestations et faire connaître leur cause [5].

Objectif et contribution

Les plateformes des médias sociaux ont été des outils d'organisation, d'information et de mobilisation cruciaux pour les manifestations algériennes tout au long de l'année 2019. À l'aide d'algorithmes d'apprentissage automatique qui exploitent les caractéristiques recueillies à partir des données des médias sociaux liées aux appels à manifester, notre objectif dans cette situation est de prévoir les manifestations publiques. Nous portons une attention particulière au cas du mouvement "Hirak", qui a débuté en février 2019. Notre proposition de méthode est basée sur l'utilisation de caractéristiques géographiques déduites du contenu des messages des utilisateurs et des communications agrégées des utilisateurs.

L'objectif principal est de mettre en place un modèle de prédiction basé sur la catégorisation pour prédire des manifestations à grande échelle en utilisant du matériel provenant des médias sociaux ouverts. La première étape consiste à rechercher les premiers indicateurs de protestation comme les hashtags. Les données seront ensuite collectées et préparées pour la formation du modèle et l'extraction des fonctionnalités. Nous insistons également sur les travaux futurs qui appellent une réflexion plus approfondie ainsi que sur les recherches en cours et les sujets connexes.

Organisation du rapport

Les deux volets de notre travail sont théoriques et pratiques. Dans la première partie, nous explorons l'analyse des sentiments de manière générale,les différents niveaux d'analyse et en examinant les approches utilisées. La deuxième partie se concentre sur l'importance de l'analyse multilingue dans l'analyse des sentiments. Nous soulignons les défis spécifiques liés à l'analyse des sentiments dans des langues différentes.

Dans la deuxième section, nous fournissons le troisième chapitre, qui décrit la méthode mise en œuvre de notre travail et justifie toutes les décisions techniques (programmation dans le langage Python). Nous créons également un large aperçu des techniques de catégorisation. Dans le quatrième chapitre, les expériences et les résultats sont couverts.

Chapitre 01: Contexte de l'analyse de sentiments

1.1. Introduction :

L'analyse de sentiments est devenue de plus en plus importante avec la montée en puissance des réseaux sociaux et des plateformes en ligne, qui génèrent un volume considérable de données textuelles en temps réel. Les entreprises, les marques et les organisations ont ainsi besoin de comprendre les opinions et les attitudes de leur public pour améliorer leur réputation, leur marketing et leur prise de décision.

1.2. Analyse de sentiments :

L'analyse de sentiments est une technique utilisée pour identifier et extraire des informations subjectives d'un texte. Il est utilisé pour mesurer l'attitude, les opinions et les émotions des individus, et est souvent utilisé pour analyser les commentaires des clients [6]. L'analyse des sentiments peut être utilisée pour mieux comprendre le sentiment des clients et pour aider les entreprises à comprendre ce que les gens pensent d'un produit ou d'un service particulier. L'analyse des sentiments peut également être utilisée pour identifier les tendances dans les commentaires des clients et pour aider les entreprises à prendre des décisions éclairées.

1.3. Les niveaux de l’analyse des sentiments :

L'analyse de sentiments peut être classée en différents niveaux en fonction de la portée et de la complexité de l'analyse. Voici les niveaux d'analyse de sentiments communément reconnus[7] :

1. **Analyse des sentiments au niveau du document** : il s'agit du niveau le plus élémentaire d'analyse des sentiments, où le sentiment d'un document entier ou d'un morceau de texte est analysé dans son ensemble. Cela implique généralement d'attribuer un score de polarité (positif, négatif ou neutre) au document en fonction du sentiment général qui y

est exprimé qui peut l'utiliser dans l'étude de marché, la gestion de la réputation, l'analyse politique, le service clients ou encore la détection de fraude.

2. **Analyse des sentiments au niveau de la phrase** : à ce niveau, l'analyse des sentiments est effectuée au niveau de la phrase, où chaque phrase est analysée individuellement pour déterminer son score de polarité. Ce niveau d'analyse peut fournir des informations plus détaillées sur le sentiment exprimé dans le texte, elle peut être utile pour les commentaires des clients, l'amélioration du produit ou du service, la surveillance des médias sociaux, la veille de marque et la publicité et marketing.
3. **Analyse des sentiments au niveau de l'aspect** : ce niveau d'analyse des sentiments va au-delà de l'analyse au niveau des phrases et des documents et se concentre sur l'identification du sentiment exprimé à l'égard d'aspects ou de caractéristiques spécifiques d'un produit, d'un service ou d'une entité. Il s'agit d'identifier et d'extraire les aspects ou caractéristiques mentionnés dans le texte et d'analyser le sentiment associé à chaque aspect principalement utilisé pour l'avis sur les produits ou services, l'analyse compétitive, le service client et l'étude de marché.
4. **Analyse des sentiments au niveau de l'entité** : il s'agit du niveau le plus avancé d'analyse des sentiments, où le sentiment est analysé au niveau de l'entité, comme le sentiment envers une marque, un produit ou une personne en particulier. Cela implique d'identifier les entités mentionnées dans le texte, d'analyser le sentiment exprimé envers chaque entité et d'agréger les scores de sentiment pour fournir un score de sentiment global pour l'entité peut être retrouvé dans les surveillance des médias sociaux, la gestion de la réputation, l'analyse des retours clients et aussi l'analyse de l'actualité.

Chaque niveau d'analyse des sentiments fournit des informations différentes et peut être utilisé à des fins différentes en fonction des besoins spécifiques de l'application ou du cas d'utilisation .

1.4. L’utilisation de l’analyse des sentiments :

L'analyse des sentiments est utilisée dans une variété d'applications, telles que l'analyse des commentaires des clients, les études de marché, l'exploration d'opinions et l'analyse politique. Dans l'analyse des commentaires des clients, l'analyse des sentiments est utilisée pour identifier le sentiment des clients à propos d'un produit ou d'un service particulier, et peut être utilisée pour obtenir des informations sur la satisfaction et la fidélité des clients [7].

Dans les études de marché, l'analyse des sentiments peut être utilisée pour identifier les tendances du sentiment des clients et pour aider les entreprises à prendre des décisions éclairées. Dans l'exploration d'opinion, l'analyse des sentiments peut être utilisée pour identifier les tendances de l'opinion publique et pour identifier les acteurs les plus influents dans une conversation donnée. Dans l'analyse politique, l'analyse des sentiments peut être utilisée pour suivre, analyser et comparer différents mouvements politiques et leurs conversations associées, et pour comprendre la dynamique des interactions entre les participants.

1.5. Différentes approches et techniques :

L’analyse des sentiments utilise plusieurs approches pour analyser le langage naturel et chaqu’une de ces approches a sa propre techniques on cite parmis ces dernière :

1.5.1. Approche lexicale :

Une approche lexicale est une méthode d'analyse de texte ou de langage en examinant le sens de mots individuels et les relations entre eux. Les approches lexicales de l'analyse linguistique impliquent d'examiner le sens des mots et des phrases, d'examiner les relations entre eux et d'utiliser ces relations pour interpréter le texte. Cette approche peut être utilisée pour identifier les modèles et les tendances dans l'utilisation de la langue et pour mieux comprendre comment la langue est utilisée dans différents contextes [9].

Les approches lexicales peuvent également être utilisées pour l'analyse des sentiments, pour identifier le sentiment exprimé dans un texte ou pour identifier l'opinion

d'un individu ou d'un groupe. Cette approche est basée sur l'idée que le langage est composé de petits morceaux de sens qui peuvent être appris grâce à l'utilisation de grandes quantités d'entrées. L'approche lexicale consiste à utiliser une combinaison de techniques telles que :

1. Apprendre des éléments lexicaux en contexte - Cela implique d'utiliser des exemples de langue dans son environnement naturel, par exemple, en utilisant des conversations ou des textes réels.
2. Apprendre des mots en morceaux - Cela implique d'apprendre des mots en groupes, tels que des phrases, des collocations et des idiomes.
3. Apprentissage des familles de mots - Cela implique l'apprentissage de mots apparentés, tels que des synonymes, des antonymes et des mots apparentés.
4. Apprentissage des formes de mots - Cela implique l'apprentissage des différentes formes d'un mot, telles que les formes verbales et nominales.
5. Apprendre la signification des mots - Cela implique d'apprendre les différentes significations d'un mot, telles que ses significations littérales et figuratives.

Dans l'analyse des sentiments, l'approche lexicale utilise un lexique des sentiments, qui est une collection de mots et leurs scores de sentiments associés. Cela permet au système d'analyse des sentiments d'identifier le sentiment d'un texte donné en examinant les valeurs de sentiment des mots qu'il contient. Cette approche permet également une analyse plus complexe, car elle peut examiner le sentiment des phrases et d'autres constructions linguistiques.

1.5.2. L'apprentissage automatique (ML) :

L'apprentissage automatique (ML) est un type d'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données et de faire des prédictions sans être explicitement programmés. Il est utilisé dans une variété d'applications, y compris l'analyse des sentiments, la reconnaissance d'images, le traitement du langage naturel et le diagnostic médical [9]. Les algorithmes ML sont basés sur l'idée que les données peuvent

être utilisées pour "entraîner" un ordinateur à reconnaître des modèles et à faire des prédictions.

Pour l'analyse des sentiments, les algorithmes ML peuvent être utilisés pour identifier et classer le sentiment exprimé dans le texte, comme l'analyse des commentaires des clients pour identifier le sentiment des clients à propos d'un produit ou d'un service particulier. Les algorithmes ML peuvent également être utilisés pour identifier les tendances des commentaires des clients et aider les entreprises à prendre des décisions éclairées [10].

Il existe de nombreuses techniques différentes utilisées pour les approches d'apprentissage automatique (ML). Selon l'application, différentes techniques peuvent être appropriées. Certaines des techniques les plus courantes pour les approches ML comprennent l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage par renforcement, l'apprentissage par transfert et l'apprentissage en profondeur[10]. L'apprentissage supervisé implique la formation d'un modèle sur des données étiquetées, tandis que l'apprentissage non supervisé implique le regroupement de données en groupes sans données étiquetées.

L'apprentissage par renforcement est un processus itératif d'essais et d'erreurs, et l'apprentissage par transfert consiste à transférer des connaissances d'un modèle à un autre.

L'apprentissage en profondeur utilise des réseaux de neurones artificiels pour traiter les données et prendre des décisions. Chacune de ces techniques a ses propres forces et faiblesses, et le choix de la bonne approche dépend de l'application en question .

1.5.3. Les approches hybrides :

Les approches hybrides de l'analyse des sentiments combinent deux méthodes ou plus, telles que les approches basées sur le lexique et l'apprentissage automatique. En combinant différentes méthodes, les approches hybrides sont capables de tirer parti des points forts de chaque méthode, tels que la précision et la rapidité de l'approche basée sur le lexique et la flexibilité et l'évolutivité de l'apprentissage automatique. Les approches

hybrides sont également capables de tirer parti des faiblesses de chaque méthode et de réduire les taux d'erreur globaux.

Les approches hybrides peuvent être utilisées pour une variété de tâches, telles que l'analyse des sentiments, l'exploration d'opinions et la détection des émotions. Ces techniques utilisent généralement des règles sémantiques, des ensembles flous, des algorithmes d'apprentissage automatique non supervisés et des techniques de traitement du langage naturel pour analyser les avis et autres textes.

Cette approche est utilisée pour classer les textes en catégories de sentiments positifs, négatifs et neutres. De plus, les techniques hybrides utilisent souvent un système en cascade et des méthodes de recherche de coucou pour améliorer encore la précision et la classification [10].

1.5.4 Autres approches :

Il existe de nombreuses autres approches de l'analyse des sentiments, telles que les approches basée sur l'aspect qui est une tâche analytique qui vise à prédire les polarités des émotions est l' analyse des sentiments basée sur l' aspect , les objectifs déclarés du texte ou des mots clés , tels que "produit" ou "service" [11]. Les éléments peuvent être des attributs, des caractéristiques . Le processus de classification des sentiments en fonction d'un aspect comporte deux étapes : extraire les aspects et classer les sentiments [12] ou l'apprentissage par transfert qui est une technique qui applique les connaissances déjà acquises dans un domaine à un autre en utilisant des similitudes dans les données, la distribution des données , les tâches du modèle et d'autres facteurs [13].

Chacune de ces approches a ses propres avantages et inconvénients et peut être utilisée pour différentes tâches en fonction des besoins particuliers de l'application on les résume dans le **Tableau 1** qui suit

Tableau 1. Les avantages et inconvénient des différentes approches

Les approches	Les avantages	Les inconvénient
L'approche lexicale	● Facilité d'utilisation ● Rapidité ● Flexibilité	● Manque de contextualisation ● Manque de nuances ● Dépendance à la qualité des dictionnaires
L'apprentissage automatique (ML)	● Précision ● Adaptabilité ● Apprentissage continu	● Besoin de données ● Coût élevé ● Difficulté de compréhension ● Risque de biais
Les approches hybrides	● Précision accrue ● Flexibilité ● Compréhension améliorée	● Complexité accrue ● Coût élevé ● Temps de développement accru
L'approche basée sur l'aspect	● Compréhension fine des opinions ● Utilisation des commentaires clients ● Identification des forces et faiblesses	● Besoin de connaissances spécifiques ● Temps de développement ● Risque d'erreur
apprentissage par transfert	● Réduction du temps et des ressources nécessaires ● Amélioration de la performance ● Réutilisation des connaissances	● Adaptation nécessaire ● Risque de surapprentissage ● Dépendance au modèle préexistant

1.6. Défis dans les analyses de sentiments :

Si l'analyse de sentiments est un domaine en constante évolution, c'est dû aux nombreux défis qu'elle représente. En fait, le langage humain est souvent ambigu ou très contextuel, ce qui rend la compréhension automatique extrêmement difficile sans assistance humaine. L'aide

humaine est essentielle lors de la formation en apprentissage automatique d'une solution d'analyse de sentiment [14].

1.6.1. Ensembles de données multilingues (mixtes):

Ces ensembles de données peuvent contenir des mots de plusieurs langues, en plus des dialectes et des variations régionales. Cela peut rendre difficile l'identification précise des sentiments à partir de ces ensembles de données, car les modèles de langage utilisés pour l'analyse des sentiments peuvent ne pas être familiers avec les mots et les phrases utilisés dans ces textes.

1.6.2. Les dialectes et les variations régionales:

Ils peuvent être difficiles à identifier avec précision pour les modèles linguistiques, car ils peuvent ne pas reconnaître le dialecte ou la variation régionale en question. Pour relever ces défis, il est important d'utiliser des modèles linguistiques formés sur des ensembles de données multilingues, car cela aidera les modèles à mieux comprendre le texte et à identifier avec précision le sentiment.

1.6.3. Détection de sarcasme :

Le sarcasme est décrit par le dictionnaire anglais Macmillan comme l'acte de dire ou d'écrire l'inverse de ce que l'on a l'intention, ou de parler d'une manière destinée à rendre une autre personne stupide ou à démontrer sa colère [15] [16].

Quand quelqu'un écrit quelque chose de bien mais signifie vraiment négatif ou vice versa, le problème du sarcasme dans l'analyse des sentiments se pose, ce qui rend le processus d'analyse des sentiments plus difficile.

Dans la conversation de tous les jours, nous utilisons fréquemment des termes sarcastiques. Ainsi, il y a un intérêt croissant pour la détection du sarcasme pour résoudre le problème de l'obtention de sentiments malhonnêtes en identifiant automatiquement les déclarations sarcastiques dans un texte donné. L'identification du sarcasme est un travail NLP extrêmement difficile en raison de la complexité et de l'ambiguïté du sarcasme [17].

1.6.4. Traitement des négations:

Puisqu'ils peuvent changer la polarité d'un texte, les mots de négation comme pas, ni, etc. doivent être manipulés avec précaution dans l'analyse des sentiments. Une phrase comme "Ce film est bon" est un exemple de phrase positive, tandis que "Le film n'est pas bon" est un exemple de phrase négative.

Les mots de négation sont malheureusement parfois exclus des techniques car ils sont sur des listes de mots vides ou sont implicitement ignorés car ils ont une valeur émotionnelle neutre dans un lexique et n'affectent pas la polarité finale. Cependant, il n'est pas facile de gérer cette tâche en inversant la polarité car les mots de négation peuvent être trouvés dans une phrase sans influencer le sentiment du texte.

1.6.5. Détection de spam:

L'identification du spam est cruciale pour la discipline de l'analyse des sentiments. Alors que les choix d'achat des consommateurs sont influencés par les opinions en ligne, les spams et les avis frauduleux peuvent nuire à la réputation d'une marque et influencer artificiellement la perception que les gens en ont.

1.6.6. Résolution des anaphores et des coréférences:

Un lien de coréférence entre concepts linguistiques est une anaphore[18]. Il est utile de savoir à quoi un pronom dans une phrase fait référence dans l'analyse des sentiments, en particulier pour les aspects, car cela aide à extraire toutes les facettes d'un élément donné, malheureusement, les pronoms sont souvent abandonnés ou négligés dans une étape préalable[19].

1.6.7. Analyse des sentiments des données codées:

L'utilisation du vocabulaire et de la grammaire de plusieurs langues dans une seule phrase est connue sous le nom de code-mixing (CM) [20][21] . C'est extrêmement typique dans les communautés qui parlent de nombreuses langues et présente un obstacle important pour les tâches de la NLP comme l'analyse des sentiments. L'identification de la sémantique

compositionnelle, qui est cruciale pour mener des recherches, est entravée par l'absence d'une grammaire formelle de phrases codées.L'une des principales difficultés est l'absence de directives de mélange.

1.7. Conclusion :

Le traitement des fichiers textuels algériens est un autre défi pour l'analyse des sentiments car il présente un certain nombre de difficultés car ils sont écrits dans un dialecte arabe distinct et présentent de nombreux problèmes d'orthographe et de grammaire. De plus, les ensembles de données n'ont souvent pas suffisamment d'échantillons de formation et de test en raison de la complexité du langage et d'un manque de normalisation. Nous allons essayer de répondre à ces dilemmes dans les prochains chapitres.

Chapitre 02: Méthodologies de l’analyse de sentiment

2.1. Introduction

La méthode d'analyse de sentiment est une approche largement utilisée pour évaluer les opinions, les attitudes et les émotions des gens envers un sujet particulier en analysant les données textuelles qu'ils ont générées. Cette méthode est largement utilisée dans de nombreux domaines, notamment le marketing, la recherche en sciences sociales, la politique, la finance et bien d'autres.

Cependant, comme pour toute méthode d'analyse, l'analyse de sentiment nécessite une méthodologie appropriée pour garantir l'exactitude et la validité des résultats. Dans cette optique, il est important de comprendre les différentes techniques utilisées pour l'analyse de sentiment, les défis associés à cette méthode et les considérations éthiques qui doivent être prises en compte pour éviter les biais et les préjugés dans l'analyse.

2.2. Types d'analyse des sentiments:

2.2.1. Analyse de polarité et l’analyse avancée

Lors de l'analyse des données textuelles pour le sentiment, deux principaux types d'approches d'analyse des sentiments sont utilisés: l'analyse de polarité et l'analyse avancée.
Le texte est classé comme ayant un sentiment positif, négatif ou neutre à l'aide de l'analyse de polarité, parfois appelée analyse de sentiment binaire. Ceci est accompli en examinant la fréquence et l'intensité d'un vocabulaire spécifié de mots et de phrases positifs et négatifs dans le texte. Bien que l'analyse de la polarité soit généralement simple et simple à utiliser, elle peut manquer des subtilités de sentiment dans le texte comme le sarcasme, l'ironie ou des émotions mitigées [22].

D'autre part, l'analyse avancée est une forme plus complexe d'analyse des sentiments qui utilise des méthodes d'apprentissage automatique pour trouver des sentiments dans des données textuelles. Le contexte du texte, y compris son ton, sa grammaire et sa sémantique, ainsi que les données démographiques de l'utilisateur, ses antécédents culturels et d'autres éléments

contextuels susceptibles d'affecter le sentiment, sont tous pris en compte dans l'analyse avancée. Une analyse avancée peut donner une évaluation plus approfondie du sentiment du texte, mais elle nécessite plus d'informations et de connaissances spécialisées à utiliser.

En fonction de l'application individuelle et de la quantité d'informations et de la complexité nécessaires à l'analyse des sentiments, l'analyse de polarité ou l'analyse avancée doit être choisie. L'analyse de la polarité peut être appropriée si une classification directe des émotions positives, négatives ou neutres est adéquate pour l'analyse. Une analyse avancée pourrait être nécessaire si une compréhension plus approfondie du sentiment est nécessaire.

2.2.2. Analyse monolingue et multilingue

Monolingue et multilingue sont des termes qui décrivent le nombre de langues impliquées dans une situation donnée. Le terme "monolingue" signifie qu'une personne ou une situation implique une seule langue, tandis que "multilingue" implique l'utilisation de plusieurs langues [22].

Dans le contexte de la technologie de la langue naturelle, la différence entre monolingue et multilingue est importante car elle peut affecter la façon dont les systèmes de traitement du langage naturel sont conçus et développés. Un système monolingue est conçu pour traiter une seule langue, tandis qu'un système multilingue est conçu pour traiter plusieurs langues.

Les avantages de l'utilisation de systèmes multilingues comprennent la capacité à traiter des données de différentes langues sans avoir besoin de systèmes distincts pour chaque langue, ainsi que la possibilité de détecter et de traduire automatiquement entre les langues cependant, les systèmes multilingues peuvent également présenter des défis, notamment la complexité de traiter plusieurs langues avec des règles grammaticales et syntaxiques différentes, ainsi que la difficulté de garantir la précision et la cohérence des résultats entre les différentes langues.

En fin de compte, la décision d'utiliser un système monolingue ou multilingue dépendra des besoins spécifiques de l'utilisateur et des exigences de la situation. Les deux options ont des avantages et des inconvénients et le choix doit être fait en tenant compte des ressources disponibles et des objectifs à atteindre.

2.3. L’importance de multilangues dans l’analyses des sentiments

L'analyse des sentiments multilingue implique une analyse des sentiments dans de nombreuses langues. Ce qui est difficile avec les sentiments, c'est que la langue et la culture ont un impact important sur la façon dont nous nous sentons et nous comportons en tant que consommateurs. Semblable au référencement et à la traduction SEO (**Optimisation pour les moteurs de recherche**), vos efforts risquent d'échouer si vous ne comprenez pas bien le contexte culturel de l'utilisateur [22].

L'analyse des sentiments ne peut pas être uniquement en anglais pour les organisations ayant une base d'utilisateurs ou de clients mondiale. Si vous souhaitez éviter les erreurs de sentiment et les mauvaises interprétations, par exemple le sentiment de vos clients portugais, devra être analysé en portugais.

2.4. Les difficultées de multilangues dans l’analyses des sentiments

L'analyse des sentiments pour les ensembles multilingues peut présenter des défis uniques en raison des différences entre les langues, des différences culturelles et des ressources limitées pour certaines langues. Voici quelques difficultés courantes de l'analyse des sentiments pour les ensembles multilingues :

2.4.1. Pré-traitement avec différents encodages

Chaque application NLP(Natural Language Processing) doit commencer par le prétraitement du texte. La norme industrielle actuelle exige que le texte ait le style de codage Unicode afin que les ordinateurs puissent le traiter. D'autres styles de codage, tels que KOI8 utilisé pour l'alphabet cyrillique, peuvent être utilisés pour les langues qui utilisent des alphabets non latins ou des sources de données textuelles anciennes. En conséquence, de nombreux langages de codage de prétraitement ne peuvent pas comprendre les langages humains, et le processus NLP est complètement arrêté [23].

2.4.2. Classificateurs pour le texte et les mots vides

Les listes de mots intégrées sont utilisées par le logiciel NLP pour aider à filtrer les mots qui n'ont pas de sens significatif. Pour garder un texte anglais intelligible pour une analyse NLP, par exemple, des termes comme "le, un et une" doivent être éliminés. Bien que les mots vides soient présents dans de nombreuses langues, ils ne peuvent toujours pas être produits automatiquement pour couvrir toutes les variétés linguistiques[23].

Les classificateurs de texte sont une deuxième liste intégrée utilisée pour l'analyse des sentiments , ils classent essentiellement chaque mot ou ensemble de mots comme positif, négatif ou neutre. Les meilleurs packages qui incluent de telles listes intégrées pour différentes langues ont été mis en évidence dans notre recherche sur les meilleurs outils open source pour l'analyse des sentiments, bien que cela ne soit pas le cas pour tous les produits sur le marché.

2.4.3. Variations dans la structure linguistique

La division du texte en minuscules morceaux significatifs, est une étape clé de l'analyse des sentiments. Dans le but d'identifier le sens fourni par les préfixes, comme la distinction entre "bon" et "pas bon", ces unités sont fréquemment des mots ou des groupes de mots. Cette procédure devrait changer pour plusieurs langues en raison de différences de structure grammaticale, telles que la quantité de mots nécessaires pour exprimer une certaine signification ou l'utilisation de différents suffixes et préfixes, selon des recherches menées à l'Université de Stanford[23].

2.5. Stratégies principales pour l'analyse des ensembles multilingues

L'analyse des ensembles multilingues implique le traitement de données provenant de plusieurs langues différentes. Cela peut être un défi pour les méthodes d'analyse, car chaque langue a ses propres caractéristiques et particularités qui doivent être prises en compte [24] .

2.5.1. Approche de traduction

L'approche de traduction pour l'analyse des ensembles multilingues consiste à traduire toutes les données dans une langue commune avant de les analyser. Cette méthode est souvent

utilisée lorsqu'il y a un grand nombre de langues différentes et que le traitement des données dans chaque langue séparément serait trop coûteux en termes de temps et de ressources [24].

Voici les principales étapes de l'approche de traduction :

1. **Collecte des données :** Les données doivent être collectées dans chaque langue séparément. Les données doivent être de qualité et représentatives de la population cible.
2. **Traduction des données :** Les données doivent être traduites dans une langue commune à l'aide d'un système de traduction automatique ou d'un traducteur humain. Il est important de s'assurer que la traduction est de haute qualité et que les nuances culturelles et linguistiques sont correctement prises en compte.
3. **Prétraitement des données :** Les données doivent être prétraitées dans la langue cible pour éliminer les erreurs et les données inutiles. Cela peut inclure le nettoyage des données, la normalisation, la tokenisation et la lemmatisation.
4. **Analyse des données :** Les données doivent être analysées dans la langue cible à l'aide des techniques d'analyse appropriées. Les résultats de l'analyse doivent être documentés soigneusement.
5. **Interprétation des résultats :** Les résultats de l'analyse doivent être interprétés en tenant compte des nuances linguistiques et culturelles qui ont pu être perdues dans la traduction. Les résultats peuvent être utilisés pour prendre des décisions éclairées dans divers domaines, tels que le marketing, la recherche d'opinion et la gestion de la réputation.

L'approche de traduction est un moyen efficace de traiter des ensembles multilingues de grande envergure. Cependant, la qualité de la traduction peut avoir un impact significatif sur la précision et la validité des résultats de l'analyse. Il est donc important de s'assurer que la traduction est de haute qualité et que les nuances linguistiques et culturelles sont correctement prises en compte avant de procéder à l'analyse des données.

2.5.2. Approche parallèle

L'approche parallèle pour l'analyse des ensembles multilingues consiste à traiter les données de chaque langue séparément et à les comparer ensuite pour identifier les similitudes et les différences entre les différentes langues [24].
Cette méthode peut être utilisée pour analyser différents types de données, y compris les textes, les images et les vidéos.
Voici les principales étapes de l'approche parallèle :

1. **Collecte des données :** Les données doivent être collectées dans chaque langue séparément. Il est important de s'assurer que les données sont de qualité et qu'elles sont représentatives de la population cible.
2. **Prétraitement des données :** Les données doivent être prétraitées dans chaque langue pour éliminer les erreurs et les données inutiles. Cela peut inclure le nettoyage des données, la normalisation, la tokenisation et la lemmatisation.
3. **Analyse des données :** Les données doivent être analysées séparément pour chaque langue à l'aide des techniques d'analyse appropriées. Les résultats de chaque analyse doivent être documentés soigneusement.
4. **Comparaison des résultats :** Les résultats de chaque analyse doivent être comparés pour identifier les similitudes et les différences entre les différentes langues. Les différences peuvent être attribuées aux caractéristiques particulières de chaque langue, telles que la grammaire, la syntaxe et les expressions idiomatiques.
5. **Interprétation des résultats :** Les résultats de l'analyse parallèle doivent être interprétés en tenant compte des différences entre les langues et des nuances culturelles. Les résultats peuvent être utilisés pour prendre des décisions éclairées dans divers domaines, tels que le marketing, la recherche d'opinion et la gestion de la réputation.

En fin de compte, l'approche parallèle est une méthode efficace pour analyser les ensembles multilingues, mais elle peut être chronophage et coûteuse en ressources. Il est donc important de considérer soigneusement les avantages et les inconvénients de cette méthode avant de l'utiliser.

2.5.3. Approche hybride :

L'approche hybride pour l'analyse des ensembles multilingues combine l'approche parallèle et l'approche de traduction pour tirer parti des avantages des deux méthodes, elle est particulièrement utile lorsque les différences entre les langues sont significatives, mais que les ressources sont limitées [24].

Voici les principales étapes de l'approche hybride :

1. **Collecte des données :** Les données doivent être collectées dans chaque langue séparément. Il est important de s'assurer que les données sont de qualité et représentatives de la population cible.
2. **Prétraitement des données :** Les données doivent être prétraitées dans chaque langue pour éliminer les erreurs et les données inutiles. Cela peut inclure le nettoyage des données, la normalisation, la tokenisation et la lemmatisation.
3. **Approche de traduction :** Les données peuvent être traduites dans une langue commune à l'aide d'un système de traduction automatique ou d'un traducteur humain. Il est important de s'assurer que la traduction est de haute qualité et que les nuances culturelles et linguistiques sont correctement prises en compte.
4. **Analyse des données :** Les données doivent être analysées séparément pour chaque langue et dans la langue cible à l'aide des techniques d'analyse appropriées. Les résultats de chaque analyse doivent être documentés soigneusement.
5. **Comparaison des résultats :** Les résultats de chaque analyse doivent être comparés pour identifier les similitudes et les différences entre les différentes langues. Les différences peuvent être attribuées aux caractéristiques particulières de chaque langue, telles que la grammaire, la syntaxe et les expressions idiomatiques.
6. **Interprétation des résultats :** Les résultats de l'analyse hybride doivent être interprétés en tenant compte des différences entre les langues et des nuances culturelles. Les résultats peuvent être utilisés pour prendre des décisions éclairées dans divers domaines, tels que le marketing, la recherche d'opinion et la gestion de la réputation.

2.6. Conclusion :

En conclusion, l'analyse de sentiment est une méthode puissante et largement utilisée pour évaluer les opinions, les attitudes et les émotions des gens envers un sujet particulier en analysant les données textuelles. Cependant, pour obtenir des résultats précis et fiables, il est important de suivre une méthodologie appropriée qui prend en compte les techniques d'analyse, les défis associés à l'analyse et les considérations éthiques.

En fin de compte, lorsque la méthode est utilisée avec soin et prudence, l'analyse de sentiment peut fournir des informations précieuses pour aider les entreprises et les organisations à mieux comprendre leur public cible et à prendre des décisions stratégiques plus éclairées.

Chapitre 03: [Méthodologie]

3.1. Introduction :

Les méthodologies d'analyse de sentiments peuvent varier en fonction des données et des objectifs visés. En générale, c'est un processus qui se décompose en plusieurs étapes ces étapes sont présentés dans l'organigramme qui suit :

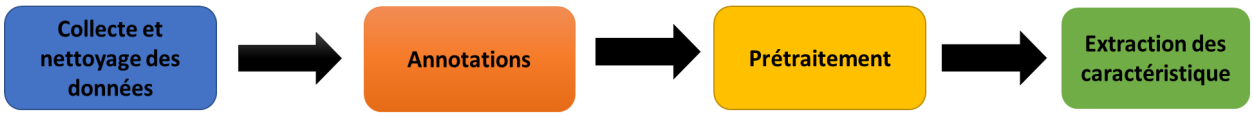

Figure 1. Les étapes principales de prédire les protestations liées au "Hirak".

3.2. Description des étapes du projet :

3.2.1. Collecte de données :

La collecte de données a été effectuée par des étudiantes qui ont eu le même thème de mémoire, elles ont utilisé la bibliothèque **GetOldTweets3** [25] pour extraire des tweets de twitter.com.

L'API de recherche Twitter a été initialement utilisée. L'API Twitter officielle impose cependant un certain nombre de restrictions, comme une fenêtre d'extraction de 7 jours pour les tweets. Les API Premium ou Enterprise Search de Twitter, qui permettent aux utilisateurs d'accéder à des tweets vieux de 30 jours ou même à l'intégralité des archives Twitter, sont nécessaires pour obtenir des tweets envoyés à des dates antérieures.Elles ont choisi l'alternative open-source gratuite car l'autre option était chère pour eux.

GetOldTweets3 [26], une amélioration du **GetOldTweets-python** original par Jefferson Henrique, est un outil de ligne de commande et un module Python 3 associé pour récupérer les tweets passés. Même s'il est incomplet, il contourne les limites de l'API officielle de Twitter en

permettant aux utilisateurs d'accéder aux tweets de plus de sept jours. Cela est permis puisque les tweets accessibles au public peuvent être utilisés légalement pour des études universitaires.

GetOldTweets3 est simple à utiliser et permet aux utilisateurs de filtrer les tweets par langue, hashtags, popularité, recherche, dates de connexion et noms d'utilisateur. L'installation de la bibliothèque nécessite l'utilisation de **pip install GetOldTweets3**. La date, l'heure, le contenu du tweet, le nom d'utilisateur et la quantité de retweets sont inclus dans ces données récupérées. Parce qu'elle sort du cadre de notre étude, la géolocalisation des tweets n'a pas été extraite.

Afin de nettoyer les tweets, les tweets qui ont été collectés mais qui n'ont rien à voir avec le sujet ont été supprimés (en particulier les tweets qui contiennent des publicités et utilisent des hashtags de protestation pour obtenir plus de vues).

3.2.2. Nettoyage des données :

Le nettoyage des données est le premier pas vers la transformation des données. Cette tâche consiste en trois tâches subordonnées pour accomplir ce processus

- **Supprimer les doublons :** Il est probable que certains tweets apparaissent plus d'une fois dans notre ensemble de données fusionnées, car ils ont été extraites à l'aide de différents hashtags et mots-clés. C'est pour cette raison que les tweets en double ont dû être supprimés. La commande **drop_duplicates** a été utilisée pour effectuer cela.
- **Supprimer les hashtags et les mettre dans une nouvelle colonne :** On a supprimer les hashtags pour diminuer le processus de traitement tout en gardant trace de ces dernier dans une nouvelle colonne pour cela nous avons écrit une procédure qui extrait les hashtags de notre champ d'exploration et les remplace par une chaîne de caractère vide et crée une nouvelle colonne nommée hashtags pour déposer tout les hashtags extrait auparavant dans cette nouvelle colonne.
- **Supprimer les liens et les mentions :** Les URLs et mentions qui sont incluses dans les tweets doivent être supprimées car elles sont inutiles pour l'analyse. Pour cela, nous avons utilisé la fonction re.sub() de Python, qui insère des espaces à la place de toutes les phrases et premières sections de phrases commençant par HTTP ou @.

3.2.3. Annotations de données :

Suivant le plan décrit par Mohammad (2016), les arabophones natifs algériens ont marqué chaque phrase deux fois en utilisant un système prédéterminé. Chaque tweet du premier tour a été rigoureusement classé dans l'une des catégories suivantes après avoir été tagué par deux annotateurs différents[16] :

- Négatif : un tweet négatif reflète des sentiments ou des opinions négatifs concernant Hirak, ses différents acteurs ou concernant la réaction du gouvernement face à ce dernier. Un sentiment négatif peut se traduire par des indicateurs explicites ou implicites et peut dénoter tristesse, colère, anxiété, etc.
- Positif : un tweet est étiqueté comme positif s'il reflète un sentiment ou une opinion positive à propos de la SM, ses acteurs ou la situation générale en Algérie à l'époque. C'est-à-dire qu'il y a une indication claire dans le tweet - explicite ou implicite - d'un sentiment positif, comme le bonheur, l'admiration, la détente, la joie, etc.
- Neutre : Si le tweet n'avait aucune indication explicite ou implicite du sentiment de l'utilisateur, alors il a été classé comme neutre. Tous les tweets qui partageaient des faits ou des nouvelles impartiales et les informations sur le sujet ont été qualifiés de neutres. Tweets écrits comme les questions neutres étaient également qualifiés de neutres.

Les rédacteurs ont passé en revue les tweets marqués lors de la deuxième série d'annotations. La balise d'un tweet restait inchangée si elle avait été appliquée par les deux annotateurs précédents.

Cependant, s'ils annotaient chacun un seul tweet différemment, une troisième personne serait impliquée.Les auteurs ont ajouté une annotation. L'étiquette finale était celle sur laquelle au moins deux annotateurs pouvaient s'entendre. Une illustration de la répartition des annotations attribuées peut être vue sur le **Tableau 2** ci-dessous.

Tableau 2. Des exemples sur les annotations de données

Utilisateurs	Tweets	Annotateur 1	Annotateur 2
Histoir6	انت رائع واصل تحيا الجزائر	Positive	Positive
Houari_Boukar	Extraordinaire 36e mardi de mobilisation des étudiant-es & des citoyen-nes à Oran. De très bon augure pour le 1er novembre. Tahya El Ahrars, Tahya El Jazair ! ط_انتخابات_العصاباتّ	Positive	Positive
Nedjidz	A moi, il me donne envie de vomir, car si je n'étais pas une femme de principes, j'aurais peut-être servi ce sale SYST pourri avec plus d'intelligence et de panache. Tfouuu !! _Libre_Democratique n	Negative	Negative
Jazzandbloom	Exhorte que dalle ! روحوا تقودوا بالانتخابات تاعكم	Negative	Negative
wiamlb	Alger	Neutre	Neutre
TweetRetweetDz	إضراب عام لمدة ثلاث أيام طبقا للمادة 71 من الدستور إبتداء من الثلاثاء 05 نوفمبر إلى غاية 07 نوفمبر 2019	Neutre	Neutre

3.2.4. Prétraitement des données :

Le prétraitement, qui vise à transformer les données brutes en un format adapté, est une étape cruciale dans toute activité de NLP. Cela implique beaucoup d'étapes. Chaque étape sera expliquée séparément dans cette section.

- **Suppression des émoticônes et des chiffres :** Nous avons éliminé les émoticônes et les chiffres du texte de nos tweets en utilisant replace() pour remplacer tous les chiffres et les valeurs d'émoticônes par un espace vide, rendant les données plus lisibles.
- **Suppression des caractères spéciaux :** les caractères tels que ~, %, *, !, +, ", {] ont également été supprimés
- **Suppression des signes diacritiques :** Les pointeurs de consonnes et le tashkil, ou marques de voyelles, sont des exemples de signes diacritiques arabes. Les signes

diacritiques ont tous été éliminés. Par exemple, “ إِنَّ السَّعِيدَ لَمَن جُنِّبَ الفِتَنَ ” devient إن السعيد لمن جنب الفتن

- **Suppression de la ponctuation :** cette étape consiste à supprimer la ponctuation car elle ne fournit aucune information supplémentaire lors du traitement des données textuelles. De ce fait, les supprimer permet de condenser les données d'entraînement.
- **Suppression des mots d’arrêt :** Les données textuelles doivent être exemptes de mots vides. Les mots vides sont des mots répandus et souvent utilisés mais qui n'affectent pas de manière significative le sens de la phrase. Nous ne voulons pas que ces termes prennent de la place dans notre base de données ou prennent trop de temps au traitement. Pour cela, il nous est simple de les éliminer. La bibliothèque **NLTK.Corpus** est utilisée dans cette phase pour importer notre collection de mots vides.
- **Suppression des caractères répétés :** À des fins d'accentuation ou d'exagération, les lettres peuvent parfois être répétées. Les répétitions de lettres arabes ont toutes été éliminées. Une illustration ”الداااااااار“ devient ”الدار“.
- **Normalisation :** Une forme unique a été créée à partir de plusieurs formes de lettres arabes différentes. Par exemple, Alef “أ” , la lettre initiale de l'alphabet arabe, a une forme unique. telles que ”آ“ , “ا“ ,”إ” Pour rendre le texte plus cohérent, toutes les lettres Alef ont été changées en ”ا”.
- **Tokenisation :** La décomposition du texte en plus petits morceaux significatifs possibles est appelée tokenisation. Il existe des tokenizers de mots et des tokenizers de phrases. Dans cette étape, qui est essentielle pour le prétraitement du texte avant tout type d'analyse, nous avons utilisé un word tokenizer. De nombreuses bibliothèques, notamment NLTK, SpaCy et TextBlob, effectuent la tokenisation.

3.2.5. Extraction des caractéristiques :

- **Sac de mot :** Le texte est divisé en une séquence de mots ou de phrases à l'aide de la tokenisation. Dans notre illustration, nous séparons nos tweets en une chaîne de mots à l'aide de la fonction split. Le soi-disant sac de mots est utilisé pour symboliser ces termes.Pour créer une matrice de nombres de jetons à partir de ces mots, nous avons utilisé la technique **CountVectorizer**. L'utilisation d'un sac de mots a un inconvénient.

Disons qu'un certain terme apparaît dans chaque document du corpus. alors il deviendra plus significatif dans nos techniques antérieures. Cela nuit à notre analyse.

- **TF-IDF :** L'une des méthodes les plus efficaces pour calculer le poids des mots est TF-IDF. Le but de TF-IDF est de considérer la signification d'un mot pour un document d'une collection, et de normaliser les termes qui sont souvent utilisés dans tous les documents. Nous avons déterminé la fréquence des termes, abrégée TF, en utilisant le modèle Bag of Words de la section précédente. La valeur de fréquence brute d'une phrase dans un certain document sert de descripteur pour la fréquence de ce terme sur l'ensemble du vecteur de document. En divisant le nombre total de documents dans notre corpus par la fréquence des documents pour chaque terme, puis en appliquant une mise à l'échelle logarithmique au résultat, nous pouvons obtenir la fréquence inverse des documents fournie par IDF, qui est l'inverse de la fréquence des documents pour chaque terme. Nous avons utilisé la méthode **TfidfVectorizer**.

3.2.6. Analyse des sentiments :

Nous avons choisi l'approche parallèle pour l'analyse des ensembles multilingues car elle est particulièrement utile lorsque les données sont très hétérogènes et que les différences entre les langues sont significatives.

Nous avons aussi utilisé l'apprentissage automatique et l'apprentissage par transfert pour résoudre une grande variété de problèmes de traitement de données.

3.2.6.1. Apprentissage automatique :

L’apprentissage automatique[27] est une méthode de l'intelligence artificielle permettant aux machines d'apprendre à partir de données sans être explicitement programmées. Voici quelques avantages et inconvénients d'apprentissage automatique :

Avantages :

- L'apprentissage automatique permet de détecter des tendances et des modèles dans les données qui pourraient être difficiles à percevoir pour un humain.
- Les modèles d'apprentissage automatique peuvent être utilisés pour prédire des résultats futurs avec une précision élevée, en fonction des données d'entraînement fournies.

- L'apprentissage automatique peut être utilisé dans une variété de domaines, tels que la reconnaissance de la parole, la vision par ordinateur, l'analyse des sentiments, la classification de textes, etc.

Inconvénients :

- Les modèles d'apprentissage automatique nécessitent des ensembles de données d'entraînement de haute qualité pour fonctionner efficacement, ce qui peut être difficile à obtenir dans certains domaines.
- Les modèles d'apprentissage automatique peuvent être sujets à des biais dans les données d'entraînement, qui peuvent se propager à travers les prédictions du modèle.
- Les modèles d'apprentissage automatique peuvent être complexes et nécessiter des ressources de calcul importantes pour fonctionner efficacement, en particulier pour les modèles de grande taille.

Il existe de nombreux types de modèles différents utilisés dans l'apprentissage automatique, et le modèle spécifique utilisé dépendra du problème à résoudre et du type de données utilisé. Voici les types de modèles que nous avons utilisés :

3.2.6.1.1. Logistic Regression :

Logistic Regression [29] est une méthode statistique utilisée en apprentissage automatique pour les problèmes de classification binaire. Elle utilise une fonction logistique pour transformer une combinaison linéaire des caractéristiques d'entrée en une probabilité comprise entre 0 et 1, représentant la probabilité d'appartenance à la classe positive.

Avantages de la régression logistique :

- C'est un modèle simple et facile à interpréter.
- Il est rapide à entraîner et à prédire sur de grandes quantités de données.
- Il peut gérer des données non linéaires en utilisant des fonctions de base polynomiales ou des transformations non linéaires des caractéristiques.

Inconvénients de la régression logistique :

- Elle peut avoir des performances inférieures à celles d'autres modèles plus complexes pour des problèmes plus complexes.
- Elle ne fonctionne pas bien pour des problèmes de classification multiclasse, sauf s'il est combiné avec d'autres techniques telles que l'analyse discriminante.

- Elle est sensible aux données manquantes et aux valeurs aberrantes, qui peuvent affecter les performances du modèle.

En résumé, la régression logistique est un modèle simple et rapide qui peut être utile pour les problèmes de classification binaire, mais qui peut avoir des performances inférieures à celles d'autres modèles plus complexes pour des problèmes plus complexes. Elle est également sensible aux données manquantes et aux valeurs aberrantes.

3.2.6.1.2. Support Vector Machines (SVM) :

Les machines à vecteurs de support (SVM) [29] sont une méthode d'apprentissage supervisée utilisée pour la classification et la régression. Le but des SVM est de trouver un hyperplan qui sépare les données de différentes classes de manière optimale.

Avantages des SVM :

- Les SVM peuvent gérer des données non linéaires en utilisant une technique appelée noyau, qui transforme les données d'entrée dans un espace de dimension supérieure où les classes peuvent être séparées linéairement.
- Les SVM peuvent être utilisées pour des problèmes de classification binaire et multiclasse.
- Les SVM ont des performances élevées pour des ensembles de données de petite et moyenne taille, même dans des espaces de grande dimension.
- Les SVM sont peu sensibles aux valeurs aberrantes et aux données manquantes.

Inconvénients des SVM :

- Les SVM peuvent être sensibles au choix du noyau et à ses paramètres, qui doivent être choisis soigneusement pour obtenir de bonnes performances.
- Les SVM sont moins adaptées aux ensembles de données de grande taille, car elles peuvent être lentes à entraîner et à prédire.
- Les SVM ne fournissent pas de probabilités directement, mais doivent être utilisées avec des techniques de calibration de probabilité pour estimer les probabilités de chaque classe.

En résumé, les SVM sont une méthode puissante et flexible pour la classification et la régression. Ils peuvent gérer des données non linéaires et sont peu sensibles aux valeurs

aberrantes et aux données manquantes. Cependant, ils peuvent être lents pour de grandes quantités de données et sont sensibles au choix du noyau et de ses paramètres.

3.2.6.2. Apprentissage par transfert :

Le transfert d'apprentissage [24] est une technique d'apprentissage automatique qui permet de transférer des connaissances acquises par un modèle pré-entraîné à une nouvelle tâche. Au lieu d'entraîner un nouveau modèle à partir de zéro, on utilise un modèle existant qui a été pré-entraîné sur une tâche similaire ou connexe, puis on adapte le modèle aux nouvelles données de la tâche cible.

Cette technique présente plusieurs avantages :

- Le transfert d'apprentissage peut réduire le temps et les ressources nécessaires pour entraîner un nouveau modèle à partir de zéro.
- Les modèles pré-entraînés sont souvent très performants, car ils ont été entraînés sur de grandes quantités de données.
- Le transfert d'apprentissage peut aider à résoudre des problèmes de sur apprentissage, car les modèles pré-entraînés ont déjà été régularisés.
- Le transfert d'apprentissage peut aider à résoudre des problèmes de sous-ajustement, car les modèles pré-entraînés ont déjà appris des représentations utiles.

Cependant, le transfert d'apprentissage peut présenter des limites en termes de généralisation à de nouvelles tâches. Les modèles pré-entraînés peuvent avoir été entraînés sur des données différentes de celles de la tâche cible, ce qui peut entraîner une perte de précision. Il est donc important de choisir un modèle pré-entraîné approprié et d'effectuer une adaptation fine pour optimiser les performances sur la nouvelle tâche.

Nous avons utilisée les modèles d'apprentissage par transfert qui suives :

3.2.6.2.1. BERT (Bidirectional Encoder Representations from Transformers) :

C’est un modèle de traitement de langage naturel pré-entraîné sur une grande quantité de données textuelles. Voici quelques avantages et inconvénients de BERT [30] :

Avantages :

- BERT est capable de capturer des informations sémantiques et contextuelles plus complexes que les modèles de traitement de langage naturel précédents, en utilisant une approche bidirectionnelle.
- BERT peut être utilisé pour résoudre une grande variété de tâches de traitement de langage naturel, telles que la classification de texte, la génération de texte et la réponse à des questions.
- BERT est disponible en plusieurs tailles, avec des modèles plus grands ayant généralement une meilleure précision.

Inconvénients :

- Comme tous les modèles de traitement de langage naturel, BERT nécessite une grande quantité de données d'entraînement et un temps d'entraînement considérable pour atteindre sa pleine efficacité.
- BERT est un modèle très complexe avec des millions de paramètres, ce qui peut rendre son utilisation coûteuse en termes de temps et de ressources de calcul.
- La taille des modèles BERT peut également entraîner des problèmes de stockage, qui peuvent être difficiles à gérer sur des appareils avec des ressources limitées.

3.2.6.2.2. XLM-Roberta-Base :

C'est un modèle de traitement de langage naturel pré-entraîné basé sur la famille de modèles Roberta. Voici quelques avantages et inconvénients de XLM-Roberta-Base [31] :

Avantages :

- XLM-Roberta-Base a été entraîné sur une grande variété de langues, ce qui le rend efficace pour des tâches de traitement de langage naturel multilingues.
- Le modèle utilise une architecture de transformer similaire à celle de BERT, qui permet une compréhension sémantique et contextuelle approfondie du texte.
- XLM-Roberta-Base est disponible en plusieurs tailles, ce qui permet une utilisation adaptée à différents types de tâches de traitement de langage naturel.

Inconvénients :

- Comme pour tous les modèles de traitement de langage naturel pré-entraînés, XLM-Roberta-Base peut nécessiter des ressources de calcul importantes pour fonctionner

efficacement, en particulier pour les tâches qui nécessitent une grande quantité de données d'entraînement.

- La complexité du modèle peut également rendre son utilisation difficile pour les utilisateurs novices en traitement de langage naturel.
- XLM-Roberta-Base a une taille de modèle relativement importante, ce qui peut entraîner des problèmes de stockage sur des appareils avec des ressources limitées.

3.3. Conclusion :

La méthodologie est un aspect crucial pour obtenir des résultats précis et fiables dans ce domaine. Les étapes clés de la méthodologie de l'analyse des sentiments, telles que la collecte et le prétraitement des données, la sélection d'un algorithme approprié, l'entraînement de l'algorithme et l'analyse des résultats, doivent être effectuées avec soin pour garantir la qualité et la fiabilité des résultats obtenus.

Chapitre 04: Expérimentations et Résultats

4.1. Introduction :

Les environnements d'apprentissage par transfert et l'apprentissage automatique sont extrêmement puissants et contribuent à faciliter la révolution de l'IA. Il serait extrêmement difficile pour les scientifiques de travailler sur des défis d'apprentissage sans ces outils.

4.2. L'environnement de travail et les outils utilisés :

4.2.1. L'environnement Matériel :

Le matériel utilisé est représenté dans le tableau suivant :

	POSTE DE TRAVAIL
Pc	HP
Système d'exploitation	Windows 10 Professionnel
Processeur	Intel(R) Core(TM) i3-3110M CPU @2.40GHz 2.40 GHz
RAM	4 GB

4.2.2. L'environnement Logiciel:

Nous avons utilisé le langage de programmation Python, version 3.10, pour atteindre notre objectif, ce dernier est un langage de programmation adaptable, polyvalent , gratuit et c'est un langage raisonnablement simple à apprendre et très efficace qui permet aux développeurs de logiciels de fournir des solutions informatiques. Le langage python a été créé par Guido van

Rossum et il est rendu public en 1991. C'est un langage open source qui ne cesse d'évoluer depuis sa création .

4.2.3. Editeur de code:

On a utilisé Google Collab ce qui nous a permis de travailler en équipe à distance, qui est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles d'apprentissage automatique directement dans le cloud. Sans donc avoir besoin d'installer quoi que ce soit sur votre ordinateur à l'exception d'un navigateur.

Nous avons aussi utilisé de diverse package comme :

- Package pandas
- Package Numpy
- Package Matplot
- Packages re
- Package NLTK
- Package Scikit-learn
- Tensorflow

4.3. Analyse exploratoire :

Tableau 3. Nombre des tweets avant et après le prétraitement pour chaque événement

Evénements	Nombre des tweets avant le prétraitement	Nombre des tweets après le prétraitement
LE 22/02/2019 ARABE	**10000**	**8476**
LE 22/02/2019 MULTILINGUE	**1053**	**937**
LE 01/11/2019 ARABE	**3698**	**2933**
LE 01/11/2019 MULTILINGUE	**1047**	**757**

4.3.1. Générer un nuage de mots

Pour analyser les tendances dans les tweets aux dates précises que nous avons choisies, nous appuyons sur la visualisation **Wordcloud** qui peut immédiatement nous montrer quels mots sont les plus utilisés dans les tweets sur la manifestation. Un nuage de mots est produit à l'aide de la bibliothèque Wordcloud. Nous avons créé un nuage de mots pour les 30 principaux termes, où plus un mot est courant, plus il apparaît dans le nuage.

Figure 2. Nuage de mots du 22/02/2019 ARABE

Figure 3. Nuage de mots du 22/02/2019 MULTILINGUE

Les 30 mots les plus fréquemment utilisés dans les tweets collectés reflètent les principales préoccupations et objectifs du peuple algérien pendant cette période. Dans la figure 2 et 3, on peut observer que la plupart des mots sont liés à l'ancien président Bouteflika et à sa candidature pour un cinquième mandat, ce qui a suscité les premiers appels à manifester.

Figure 4. Nuage de mots du 01/11/2019 ARABE

La figure 4 reflète un profond attachement à l'Algérie en mettant en avant des mots patriotiques et des références historiques. Démontre un fort sentiment de fierté nationale et un engagement envers la préservation.

Figure 5. Nuage de mots du 01/11/2019 MULTILINGUE

Dans la figure 5 certains mots révèlent également une certaine déception et frustration face aux défis liés à la trahison et à la confiance politique et aux changements limités dans cette période.

4.3.2. Diagramme à bandes

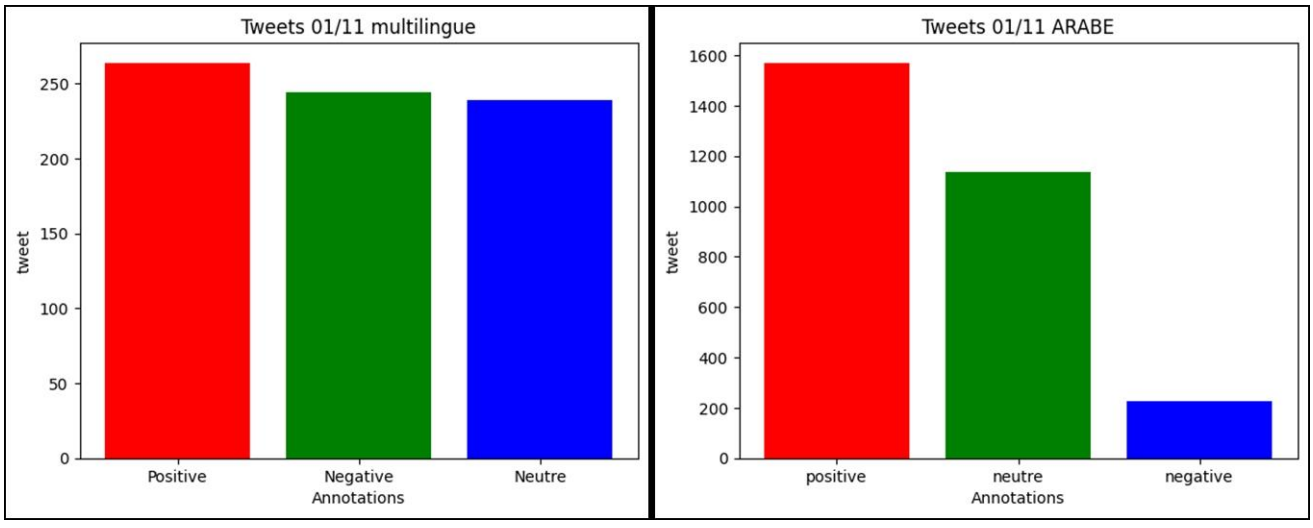

Figure 6. Diagramme à bandes de l'événement 01/11/2019

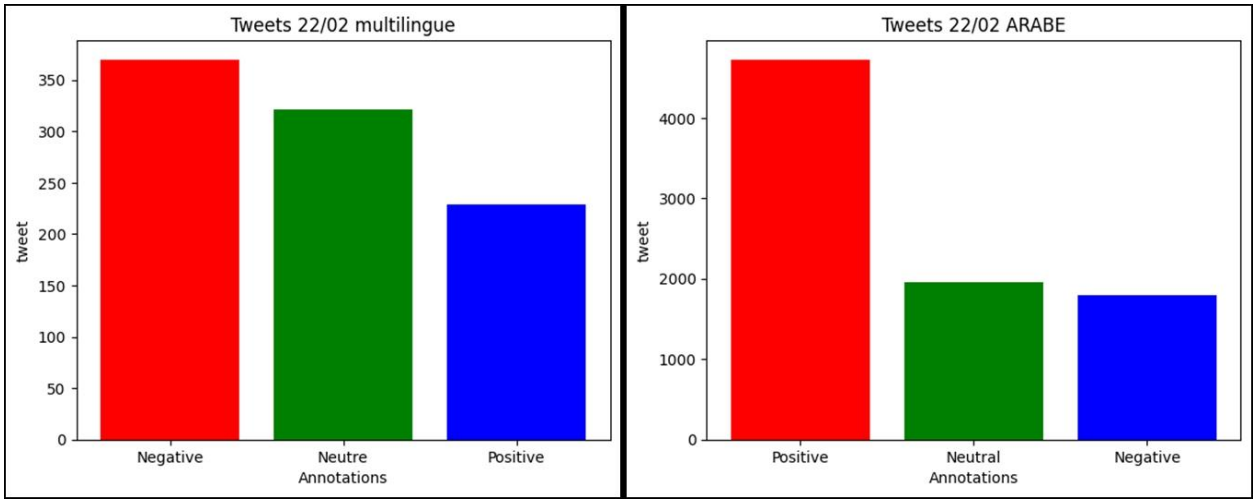

Figure 7. Diagramme à bandes de l'événement 22/02/2019

4.4. Résultats et évaluation :

Une stratégie fréquemment utilisée dans l'analyse des sentiments pour évaluer l'efficacité des modèles d'apprentissage automatique est la méthode de répartition 80/20. Les données doivent être divisées en deux ensembles distincts, un ensemble d'entraînement 20% et un ensemble de test 80% , afin d'utiliser cette technique. Le modèle est formé à l'aide de l'ensemble d'apprentissage et ses performances sont évaluées à l'aide de l'ensemble de test.

Nous avons évalué les performances des multiples modèles d'apprentissage automatique utilisés dans notre projet pendant cette phase. Nous avons pris en compte plusieurs critères tels que la précision, le rappel et le score F1 pour cette évaluation. Ces mesures sont basées sur la matrice de confusion, qui est un tableau représentant les résultats de prédiction du classificateur. Il s'agit d'un outil visuel spécifique pour évaluer les performances du modèle, présenté dans le tableau 4.

Tableau 4. Matrice de confusion

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

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

Pour évaluer les résultats nous utiliserons plusieurs paramètres tels que (Accuracy,Precision, Recall , f1 mesure) qui seront décrits ci-dessous.

- **Accuracy (ACC) :** c'est le pourcentage de prédictions correctes qui correspondent à la valeur réelle[32].

$$ACC = (T P + T N) /(T P + T N + F P + F N)$$

- **Précision (P)** : c'est le pourcentage de l'exactitude des situations positives[33].

$$P =T P / (T P + F P)$$

- **Recall (R)** : c'est le pourcentage des situations réelles qui ont été déterminées avec précision [33].

$$R =T P / (T P + T N)$$

- **F1-score :** évalue la capacité d'un modèle de classification à prédire efficacement les individus positifs, en faisant un compromis entre la précision et le recall [33].

$$F1\ Score = 2(recall.precision / (recall+precision))$$

Tableau 5. Le résultat d'exactitude avec **tf-idf** et **bow**

date		Méthode de classification	ACCURACY	Positif			Neutre			Negative		
				Précision	Recall	F1-score	Précision	Recall	F1-score	Précision	Recall	F1-score
LE 22/02/2019 Arabe	TF IDF	SVM	**0,64**	**0,69**	**0,83**	**0,75**	**0,47**	**0,33**	**0,39**	**0,57**	**0,43**	**0,49**
		LR	**0,63**	**0,64**	**0,93**	**0,76**	**0,51**	**0,20**	**0,29**	**0,68**	**0,28**	**0,40**
	BOW	SVM	**0,61**	**0,68**	**0,78**	**0,73**	**0,40**	**0,78**	**0,73**	**0,55**	**0,42**	**0,48**
		LR	**0,63**	**0,67**	**0,85**	**0,75**	**0,45**	**0,31**	**0,37**	**0,59**	**0,37**	**0,46**
LE 22/02/2019 Multilingue	TF IDF	SVM	**0,58**	**0,72**	**0,61**	**0,66**	**0,47**	**0,48**	**0,47**	**0,60**	**0,63**	**0,62**
		LR	**0,58**	**0,82**	**0,50**	**0,62**	**0,46**	**0,43**	**0,44**	**0,57**	**0,72**	**0,63**
	BOW	SVM	**0,52**	**0,63**	**0,57**	**0,60**	**0,39**	**0,55**	**0,46**	**0,59**	**0,46**	**0,52**
		LR	**0,51**	**0,71**	**0,48**	**0,57**	**0,39**	**0,62**	**0,48**	**0,58**	**0,45**	**0,51**
LE 01/11/2019 arabe	TF IDF	SVM	**0,66**	**0,66**	**0,79**	**0,72**	**0,67**	**0,61**	**0,64**	**0,50**	**0,07**	**0,12**
		LR	**0,66**	**0,64**	**0,85**	**0,73**	**0,71**	**0,54**	**0,61**	**0,00**	**0,00**	**0,00**
	BOW	SVM	**0,61**	**0,66**	**0,65**	**0,65**	**0,60**	**0,67**	**0,63**	**0,11**	**0,05**	**0,07**
		LR	**0,66**	**0,67**	**0,75**	**0,71**	**0,65**	**0,66**	**0,65**	**0,25**	**0,05**	**0,08**
LE 01/11/2019 Multilingue	TF IDF	SVM	**0,75**	**0,66**	**0,61**	**0,63**	**0,59**	**0,46**	**0,52**	**0,50**	**0,66**	**0,59**
		LR	**0,56**	**0,63**	**0,61**	**0,62**	**0,58**	**0,42**	**0,49**	**0,49**	**0,66**	**0,56**
	BOW	SVM	**0,55**	**0,64**	**0,55**	**0,59**	**0,49**	**0,63**	**0,55**	**0,56**	**0,47**	**0,51**

		LR	**0,56**	**0,62**	**0,55**	**0,58**	**0,52**	**0,62**	**0,57**	**0,55**	**0,51**	**0,53**

Voici une présentation des résultats obtenus dans le **tableau 5** avec les différentes méthodes de classification (TF-IDF et BoW) en utilisant les algorithmes SVM et LR.Nous avons comparé les performances de classification en termes d'exactitude (Accuracy) ainsi que les mesures de précision, rappel (recall) et score F1 pour chaque catégorie : Positif, Neutre et Négatif.

L'événement du 22/02/2019 (arabe):

Méthode de classification : TF-IDF

Pour le modèle SVM, nous avons une exactitude de 0,64. En ce qui concerne la catégorie "Positif", nous obtenons une précision de 0,69, un rappel de 0,83 et un score F1 de 0,75. Pour la catégorie "Neutre", la précision est de 0,47, le rappel est de 0,33 et le score F1 est de 0,39. Enfin, pour la catégorie "Négatif", nous avons une précision de 0,57, un rappel de 0,43 et un score F1 de 0,49, En ce qui concerne le modèle LR, l'exactitude obtenue est de 0,63. Pour la catégorie "Positif", nous avons une précision de 0,64, un rappel de 0,93 et un score F1 de 0,76. En ce qui concerne la catégorie "Neutre", la précision est de 0,51, le rappel est de 0,20 et le score F1 est de 0,29. Pour la catégorie "Négatif", nous obtenons une précision de 0,68, un rappel de 0,28 et un score F1 de 0,40.

Méthode de classification : BoW (sac de mots)

Pour le modèle SVM, nous avons une exactitude de 0,61. En ce qui concerne la catégorie "Positif", nous obtenons une précision de 0,68, un rappel de 0,78 et un score F1 de 0,73. Pour la catégorie "Neutre", la précision est de 0,40, le rappel est de 0,78 et le score F1 est de 0,73. Enfin, pour la catégorie "Négatif", nous avons une précision de 0,55, un rappel de 0,42 et un score F1 de 0,48, le modèle LR nous donne une exactitude de 0,63. Pour la catégorie "Positif", nous avons une précision de 0,67, un rappel de 0,85 et un score F1 de 0,75. Pour la catégorie "Neutre", la précision est de 0,45, le rappel est de 0,31 et le score F1 est de 0,37. En ce qui concerne la

catégorie "Négatif", nous obtenons une précision de 0,59, un rappel de 0,37 et un score F1 de 0,46.

L'événement du 22/02/2019 (multilingue):

Méthode de classification : TF-IDF

Pour le modèle SVM, nous avons une exactitude de 0,58. Pour la catégorie "Positif", la précision est de 0,72, le rappel est de 0,61 et le score F1 est de 0,66. Pour la catégorie "Neutre", la précision est de 0,47, le rappel est de 0,48 et le score F1 est de 0,47. Enfin, pour la catégorie "Négatif", nous avons une précision de 0,60, un rappel de 0,63 et un score F1 de 0,62, pour le modèle LR, l'exactitude obtenue est de 0,58. Pour la catégorie "Positif", nous avons une précision de 0,82, un rappel de 0,50 et un score F1 de 0,62. Pour la catégorie "Neutre", la précision est de 0,46, le rappel est de 0,43 et le score F1 est de 0,44. En ce qui concerne la catégorie "Négatif", nous obtenons une précision de 0,57, un rappel de 0,72 et un score F1 de 0,63.

Méthode de classification : BoW (sac de mots)

Pour le modèle SVM, nous avons une exactitude de 0,52. Pour la catégorie "Positif", nous obtenons une précision de 0,63, un rappel de 0,57 et un score F1 de 0,60. Pour la catégorie "Neutre", la précision est de 0,39, le rappel est de 0,55 et le score F1 est de 0,46. En ce qui concerne la catégorie "Négatif", nous avons une précision de 0,59, un rappel de 0,46 et un score F1 de 0,52. Le modèle LR, a obtenu une exactitude de 0,51. Pour la catégorie "Positif", nous avons une précision de 0,71, un rappel de 0,48 et un score F1 de 0,57. Pour la catégorie "Neutre", la précision est de 0,39, le rappel est de 0,62 et le score F1 est de 0,48. En ce qui concerne la catégorie "Négatif", nous obtenons une précision de 0,58, un rappel de 0,45 et un score F1 de 0,51.

L'événement du 01/11/2019 (arabe):

Méthode de classification : TF-IDF

En ce qui concerne SVM l'exactitude obtenue est de 0,66. Pour la catégorie Positif, la précision est de 0,66, le rappel est de 0,79 et le score F1 est de 0,72. Pour la catégorie Neutre, la précision est de 0,67, le rappel est de 0,61 et le score F1 est de 0,64. En ce qui concerne la catégorie Négatif, nous avons une précision de 0,50, un rappel de 0,07 et un score F1 de 0,12. Cependant le modèle LR (Régression Logistique) a obtenu une exactitude de 0,66. Pour la catégorie Positif, nous avons une précision de 0,64, un rappel de 0,85 et un score F1 de 0,73. Pour la catégorie Neutre, la précision est de 0,71, le rappel est de 0,54 et le score F1 est de 0,61. En ce qui concerne la catégorie Négatif, nous obtenons une précision de 0,00, un rappel de 0,00 et un score F1 de 0,00.

Méthode de classification : BoW (sac de mots)

Avec SVM L'exactitude obtenue est de 0,61. Pour la catégorie Positif, nous avons une précision de 0,66, un rappel de 0,65 et un score F1 de 0,65. Pour la catégorie Neutre, la précision est de 0,60, le rappel est de 0,67 et le score F1 est de 0,63. En ce qui concerne la catégorie Négatif, nous avons une précision de 0,11, un rappel de 0,05 et un score F1 de 0,07. Pour le modèle LR (Régression Logistique) L'exactitude obtenue est de 0,66. Pour la catégorie Positif, nous avons une précision de 0,67, un rappel de 0,75 et un score F1 de 0,71. Pour la catégorie Neutre, la précision est de 0,65, le rappel est de 0,66 et le score F1 est de 0,65. En ce qui concerne la catégorie Négatif, nous obtenons une précision de 0,25, un rappel de 0,05 et un score F1 de 0,08.

L'événement du 01/11/2019 (multilingue):

Méthode de classification : TF-IDF

En utilisant SVM l'exactitude obtenue est de 0,75. Pour la catégorie Positif, la précision est de 0,66, le rappel est de 0,61 et le score F1 est de 0,63. Pour la catégorie Neutre, la précision est de 0,59, le rappel est de 0,46 et le score F1 est de 0,52. En ce qui concerne la catégorie

Négatif, nous avons une précision de 0,50, un rappel de 0,66 et un score F1 de 0,59. Pour le modèle LR (Régression Logistique) : L'exactitude obtenue est de 0,56. Pour la catégorie Positif, nous avons une précision de 0,63, un rappel de 0,61 et un score F1 de 0,62. Pour la catégorie Neutre, la précision est de 0,58, le rappel est de 0,42 et le score F1 est de 0,49. En ce qui concerne la catégorie Négatif, nous obtenons une précision de 0,49, un rappel de 0,66 et un score F1 de 0,56.

Méthode de classification : BoW (sac de mots)

Pour le modèle SVM L'exactitude obtenue est de 0,55. Pour la catégorie Positif, nous avons une précision de 0,64, un rappel de 0,55 et un score F1 de 0,59. Pour la catégorie Neutre, la précision est de 0,49, le rappel est de 0,63 et le score F1 est de 0,55. En ce qui concerne la catégorie Négatif, nous avons une précision de 0,56, un rappel de 0,47 et un score F1 de 0,51. Pour le modèle LR (Régression Logistique) L'exactitude obtenue est de 0,56. Pour la catégorie Positif, nous avons une précision de 0,62, un rappel de 0,55 et un score F1 de 0,58. Pour la catégorie Neutre, la précision est de 0,52, le rappel est de 0,62 et le score F1 est de 0,57. En ce qui concerne la catégorie Négatif, nous obtenons une précision de 0,55, un rappel de 0,51 et un score F1 de 0,53.

Tableau 6. Le résultat d'exactitude avec **XML** et **BERT**

date	Méthode de classification	ACCURACY	Positif			Neutre			Negative		
			Précision	Recall	F1-score	Précision	Recall	F1-score	Précision	Recall	F1-score
LE 22/02/2019 Arabe	XML	**0,56**	**0,56**	**1,00**	**0,72**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**
	BERT	**0,61**	**0,70**	**0,76**	**0,73**	**0,38**	**0,42**	**0,40**	**0,56**	**0,39**	**0,46**
LE 22/02/2019 Multilingue	XML	**0,51**	**0,48**	**0,35**	**0,41**	**0,48**	**0,54**	**0,51**	**0,54**	**0,59**	**0,56**
	BERT	**0,58**	**0,55**	**0,65**	**0,59**	**0,53**	**0,62**	**0,57**	**0,65**	**0,50**	**0,57**

LE 01/11/ 2019 Arabe	XML	**0,65**	**0,62**	**0,84**	**0,72**	**0,71**	**0,52**	**0,60**	**0,00**	**0,00**	**0,00**
	BERT	**0,66**	**0,69**	**0,74**	**0,71**	**0,71**	**0,62**	**0,66**	**0,29**	**0,36**	**0,32**
LE 01/11/ 2019 Multil ingue	XML	**0,50**	**0,38**	**0,54**	**0,45**	**0,57**	**0,74**	**0,64**	**0,57**	**0,28**	**0,37**
	BERT	**0,57**	**0,66**	**0,61**	**0,63**	**0,59**	**0,46**	**0,52**	**0,50**	**0,66**	**0,57**

Pour les résultats que nous avons obtenu dans le **Tableau 6** avec le **XML** et **BERT**

L'événement du 22/02/2019 (arabe):

Méthode de classification : XML

L'exactitude obtenue est de 0,56. Pour la catégorie "Positif", nous avons une précision de 0,56, un rappel de 1,00 et un score F1 de 0,72. Cela indique que le modèle est capable de prédire avec précision les exemples positifs, mais il est important de noter que le rappel élevé indique qu'il y a peut-être des exemples positifs non détectés.

Cependant, pour la catégorie "Neutre", nous obtenons une précision, un rappel et un score F1 de 0,00, ce qui indique une performance insuffisante du modèle pour cette catégorie. Cela signifie que le modèle ne parvient pas à prédire correctement les exemples neutres.

De même, pour la catégorie "Négatif", nous obtenons également une précision, un rappel et un score F1 de 0,00, indiquant une performance insuffisante du modèle pour cette catégorie. Le modèle ne parvient pas à prédire correctement les exemples négatifs.

Méthode de classification : BERT

L'exactitude obtenue est de 0,61. Pour la catégorie "Positif", nous avons une précision de 0,70, un rappel de 0,76 et un score F1 de 0,73. Cela indique que le modèle parvient à prédire avec précision les exemples positifs, en obtenant à la fois une précision et un rappel élevés.

En ce qui concerne la catégorie "Neutre", nous obtenons une précision de 0,38, un rappel de 0,42 et un score F1 de 0,40. Ces scores indiquent que le modèle a du mal à prédire correctement les exemples neutres, avec une précision et un rappel relativement faibles.

Pour la catégorie "Négatif", nous avons une précision de 0,56, un rappel de 0,39 et un score F1 de 0,46. Cela signifie que le modèle est capable de prédire les exemples négatifs avec une précision raisonnable, mais le rappel est relativement faible, ce qui suggère que certains exemples négatifs ne sont pas correctement détectés.

L'événement du 22/02/2019 (multilingue):

Méthode de classification : XML

L'exactitude obtenue est de 0,51. Pour la catégorie "Positif", nous avons une précision de 0,48, un rappel de 0,35 et un score F1 de 0,41. Cela indique que le modèle a du mal à prédire correctement les exemples positifs, avec une précision relativement faible et un rappel encore plus bas.

Pour la catégorie "Neutre", nous obtenons une précision de 0,48, un rappel de 0,54 et un score F1 de 0,51. Ces résultats montrent que le modèle a une performance équilibrée pour les exemples neutres, avec une précision et un rappel assez proches.

En ce qui concerne la catégorie "Négatif", nous avons une précision de 0,54, un rappel de 0,59 et un score F1 de 0,56. Cela suggère que le modèle parvient à prédire les exemples négatifs avec une précision raisonnable et un rappel légèrement plus élevé que pour les autres catégories.

Il est important de noter que les performances globales du modèle sont relativement faibles, avec une exactitude de 0,51.

Méthode de classification : BERT

L'exactitude obtenue est de 0,58. Pour la catégorie "Positif", nous avons une précision de 0,55, un rappel de 0,65 et un score F1 de 0,59. Cela signifie que le modèle parvient à prédire avec une précision raisonnable les exemples positifs, mais le rappel pourrait être amélioré pour détecter davantage d'exemples positifs.

Pour la catégorie "Neutre", nous obtenons une précision de 0,53, un rappel de 0,62 et un score F1 de 0,57. Ces résultats indiquent que le modèle parvient à prédire les exemples neutres avec une précision raisonnable et un rappel correct, bien qu'il y ait encore de la marge pour améliorer ces valeurs.

En ce qui concerne la catégorie "Négatif", nous avons une précision de 0,65, un rappel de 0,50 et un score F1 de 0,57. Cela suggère que le modèle parvient à prédire les exemples négatifs avec une précision raisonnable, mais le rappel pourrait être amélioré pour capturer davantage d'exemples négatifs.

Dans l'ensemble, les performances du modèle sont modérées, avec une exactitude de 0,58.

L'événement du 01/11/2019 (arabe):

Méthode de classification : XML

L'exactitude obtenue est de 0,65. Pour la catégorie "Positif", nous avons une précision de 0,62, un rappel de 0,84 et un score F1 de 0,72. Cela indique que le modèle parvient à prédire avec une précision raisonnable les exemples positifs, avec un rappel élevé qui capture la plupart des exemples positifs.

En ce qui concerne la catégorie "Neutre", nous obtenons une précision de 0,71, un rappel de 0,52 et un score F1 de 0,60. Ces résultats indiquent que le modèle a une précision élevée pour prédire les exemples neutres, mais le rappel est relativement faible, ce qui signifie que certains exemples neutres sont manquants.

Cependant, pour la catégorie "Négatif", nous obtenons une précision, un rappel et un score F1 de 0,00, indiquant une performance insuffisante du modèle pour cette catégorie. Cela signifie que le modèle ne parvient pas à prédire correctement les exemples négatifs.

Méthode de classification : BERT

L'exactitude obtenue est de 0,66. Pour la catégorie "Positif", nous avons une précision de 0,69, un rappel de 0,74 et un score F1 de 0,71. Cela indique que le modèle parvient à prédire avec une précision raisonnable les exemples positifs, avec un rappel qui capture une grande partie des exemples positifs.

Pour la catégorie "Neutre", nous obtenons une précision de 0,71, un rappel de 0,62 et un score F1 de 0,66. Ces résultats montrent que le modèle parvient à prédire les exemples neutres avec une précision relativement élevée et un rappel correct, ce qui indique une bonne performance pour cette catégorie.

En ce qui concerne la catégorie "Négatif", nous avons une précision de 0,29, un rappel de 0,36 et un score F1 de 0,32. Ces résultats suggèrent que le modèle a du mal à prédire correctement les exemples négatifs, avec une précision faible et un rappel modéré..

L'événement du 01/11/2019 (multilingue):

Méthode de classification : XML

L'exactitude obtenue est de 0,50. Pour la catégorie "Positif", nous avons une précision de 0,38, un rappel de 0,54 et un score F1 de 0,45. Ces résultats indiquent que le modèle a du mal à prédire correctement les exemples positifs, avec une précision relativement faible et un rappel modéré.

En ce qui concerne la catégorie "Neutre", nous obtenons une précision de 0,57, un rappel de 0,74 et un score F1 de 0,64. Cela suggère que le modèle parvient à prédire les exemples neutres avec une précision raisonnable et un rappel élevé, ce qui indique une meilleure performance par rapport à la catégorie "Positif".

Cependant, pour la catégorie "Négatif", nous avons une précision de 0,57, un rappel de 0,28 et un score F1 de 0,37. Ces résultats montrent que le modèle a des difficultés à prédire correctement les exemples négatifs, avec une précision relativement élevée mais un rappel faible.

Méthode de classification : BERT

L'exactitude obtenue est de 0,57. Pour la catégorie "Positif", nous avons une précision de 0,66, un rappel de 0,61 et un score F1 de 0,63. Ces résultats indiquent que le modèle parvient à prédire les exemples positifs avec une précision raisonnable et un rappel correct, ce qui suggère une performance acceptable pour cette catégorie.

En ce qui concerne la catégorie "Neutre", nous obtenons une précision de 0,59, un rappel de 0,46 et un score F1 de 0,52. Ces résultats montrent que le modèle a des difficultés à prédire les exemples neutres, avec une précision relativement élevée mais un rappel modéré, indiquant une performance mitigée pour cette catégorie.

Pour la catégorie "Négatif", nous avons une précision de 0,50, un rappel de 0,66 et un score F1 de 0,57. Cela suggère que le modèle parvient à prédire les exemples négatifs avec une précision raisonnable, mais le rappel plus élevé indique qu'il manque certains exemples négatifs, ce qui peut affecter la performance globale pour cette catégorie.

4.5. Comparaison et discussion des résultats

En analysant les résultats, voici quelques observations et discussions :

4.5.1. Méthode de classification apprentissage par transfert (XML,BERT) :

La méthode BERT semble avoir de meilleures performances globales que la méthode XML en termes d'exactitude, de précision, de rappel et de score F1 pour les catégories Positif et Neutre cependant la méthode XML semble avoir des difficultés à détecter la catégorie Négatif, avec des performances insuffisantes dans la plupart des cas mais les performances des différentes méthodes peuvent varier en fonction de la langue et de la date d'évaluation.

En résumé, les résultats suggèrent que les méthodes BERT et XML montrent des performances variables selon les langues et les dates d'évaluation. Cependant, la méthode BERT semble généralement mieux performer que la méthode XML, en particulier pour les catégories Positif et Neutre.

4.5.2. Méthode de classification apprentissage automatique (SVM,LR) :

La comparaison montre que les performances des différentes méthodes de classification varient en fonction de la langue et de la méthode utilisée. Les résultats obtenus en utilisant le modèle SVM avec la méthode TF-IDF ont généralement été légèrement meilleurs que ceux obtenus avec la régression logistique (LR). Cependant, la performance globale reste modérée avec des scores de précision, de rappel et de F1 qui varient d'une classe à l'autre et d'une langue à l'autre.

4.5.3. Les deux méthodes :

Les résultats montrent que SVM avec TFIDF a tendance à obtenir de meilleures performances pour la classe positive par rapport aux méthodes XML et BERT. Cela peut indiquer que SVM est plus efficace dans la capture des caractéristiques discriminantes pour cette tâche de classification spécifique.

Dans l'ensemble, les deux tableaux montrent des performances de classification modérées à bonnes, avec des valeurs d'exactitude allant de 0,50 à 0,75. Cela suggère que les méthodes et les algorithmes utilisés peuvent fournir des résultats significatifs dans la tâche de classification.

4.5. Conclusion :

L'étude de l'analyse des sentiments est intéressante et bénéfique. Cependant, il n'y a pas eu beaucoup de recherches sur les textes mixtes en Algérie, qui à notre avis ont besoin d'approches particulières pour être efficaces dans notre contexte. Pouvoir des recherches supplémentaires sont nécessaires pour améliorer les résultats de la classification des sentiments.

Conclusion générale

Compréhension du traitement du langage, associée aux idées de l'intelligence artificielle , apprentissage par transfert et l'apprentissage automatique, aide au développement des systèmes intelligents qui peuvent utiliser les données textuelles à leur avantage et aider à résoudre les défis du monde réel.

L'avantage de ces leçons est qu'après les avoir appliquées,une fois le modèle formé, nous pouvons l'appliquer immédiatement à des données fraîches et non publiées pour voir les informations nécessaires.

Notre travail a de nombreux objectifs dans ce contexte. Il nous a d'abord offert la liberté d'enquêter sur le domaine de l'échange d'informations humaines sur Internet et les plateformes de médias sociaux et de comprendre toute sa complexité en matière de traitement, de contrôle et de direction. De plus, nous avons eu la chance de mettre en valeur et de mettre à profit toute notre expertise dans les domaines d'apprentissage par transfert et d'apprentissage automatique sur un sujet aussi crucial. Nous avons également eu l'opportunité de travailler dans le domaine du traitement du langage naturel, qui se développe rapidement, très prometteur et de plus en plus important pour la société moderne.

Le dernier chapitre montre comment les méthodes décrites dans ce mémoire ont produit des résultats raisonnables. Pour évaluer l'efficacité de notre extracteur d'événements, nous recueillons plus de données et faisons plus d'expériences. Nous suggérons d'utiliser une variété d'algorithmes d'apprentissage automatique pour améliorer la précision des prédictions.

Nous avons rencontré quelques obstacles en travaillant sur ce projet, notamment les obstacles suivants :

- Le traitement des données après leur extraction, car un tweet comprend une variété de langues, y compris le dialecte algérien, que l'algorithme trouve difficile à interpréter.
- La rédaction avec Word en ligne (Google Docs) à cause des coupures d’internet et le mauvais flux d’Internet.
- Le travaille à distance et la disponibilité de chacun d’entre nous.
- Annoter les tweets manuellement.

Afin d'améliorer ce travail et d'aller au-delà des limites mentionnées ci-dessus, nous proposerons les solutions réalisables suivantes:

- Utilisez une autre forme d'annotation, telle qu'une forme automatisée ou externalisée.
- Appliquer différentes techniques de catégorisation plus adaptées à la gestion des langues et dialectes difficiles.
- Introduire une nouvelle méthode de prétraitement unique au dialecte algérien et la combiner avec le français et l'anglais, deux langues supplémentaires régulièrement parlées en Algérie.

Références

[1] I. Qu’est-CE Qu’un Mouvement social ? | cairn.info . https://www.cairn.info/sociologie-des-mouvements-sociaux--9782707169358-page-5.htm.

[2] “Mouvement social contre le projet de réforme des retraites en France de 2023,” Wikipedia https://fr.wikipedia.org/wiki/Mouvement_social_contre_le_projet_de_r%C3%A9forme_des_retraites_en_France_de_2023.

[3] France 24, “La Chine annonce un allègement général des restrictions contre le covid-19,”.

[4] A. Farrokhi, F. Shirazi, N. Hajli, and M. Tajvidi, “Using artificial intelligence to detect crisis related to events: Decision making in B2B by Artificial Intelligence,” *Industrial Marketing Management*, vol. 91, pp. 257–273, 2020. doi:10.1016/j.indmarman.2020.09.015 .

[5] A. Mazari and A. Djeffal, “Deep learning-based sentiment analysis of Algerian dialect during Hirak 2019: Semantic scholar,” 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH). https://www.semanticscholar.org/paper/Deep-Learning-Based-Sentiment-Analysis-of-Algerian-Mazari-Djeffal/ae779bd58fe0b0135abc4bd15e41a43853807bd9 .

[6] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," in Knowledge and Information Systems, vol. 60, no. 2, pp. 617-663, Aug. 2019. doi: 10.1007/s10115-018-1236-4.

[7] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis*, vol. 2, 1–2 Vols. Boston: Now Publishers, 2008.

[8] N. Nigam and D. Yadav, "Lexicon-based approach to sentiment analysis of tweets using R language," in Advances in Computing and Data Sciences: Second International Conference, ICACDS 2018, Dehradun, India, April 20-21, 2018, Revised Selected Papers, Part I 2, Springer, 2018, pp. 154-164.

[9] L. A. Williams, “ Pushing the Envelope of Sentiment Analysis Beyond Words and Polarities ”.

[10] M. A. Hassonah, R. Al-Sayyed, A. Rodan, A. M. Al-Zoubi, I. Aljarah, and H. Faris, "An

efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter," Knowledge-Based Systems, vol. 192, pp. 105353, Mar. 2020. doi: 10.1016/j.knosys.2019.105353.

[11] "Improving aspect-based sentiment analysis via aligning aspect embedding," ScienceDirect.

[12] Q. Xu and L. Zhu, "Dai T., and Yan C., 'Asp.-Based Sentim. Classif. Multi-Atten. Network'," Neurocomputing, vol. 388, pp. 135-143, 2020.

[13] R. Liu, Y. Shi, C. Ji, and M. Jia, "A Survey of Sentiment Analysis Based on transfer learning ," IEEE Access, vol. 7, pp. 85401-85412, 2019. doi: 10.1109/ACCESS.2019.2925059.

[14] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," Journal of King Saud University - Engineering Sciences, vol. 30, no. 4, pp. 330-338, Oct. 2018. doi: 10.1016/j.jksues.2016.04.002.

[15] M. Mayor and M. Rundell, "Macmillan English Dictionary: For Advanced Learners," Macmillan Education, 2003.

[16] M. Birjali, A. Beni-Hssane, and M. Erritali, "A Method Proposed for Estimating Depressed Feeling Tendencies of Social Media Users Utilizing Their Data," in Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016), A. Abraham, A. Haqiq, A. M. Alimi, G. Mezzour, N. Rokbani, and A. K. Muda, Eds., in Advances in Intelligent Systems and Computing, Cham, Springer International Publishing, 2017, pp. 413-420. doi: 10.1007/978-3-319-52941-7_41.

[17] L. Ren, B. Xu, H. Lin, X. Liu, and L. Yang, "Sarcasm Detection with Sentiment Semantics Enhanced Multi-level Memory Network," Neurocomputing, vol. 401, pp. 320-326, Aug. 2020. doi: 10.1016/j.neucom.2020.03.081.

[18] I. Toledo-Gómez, E. Valtierra-Romero, A. Guzman-Arenas, A. Cuevas-Rasgado, and L. Méndez-Segundo, "AnaPro, tool for identification and resolution of direct anaphora in Spanish," Journal of Applied Research and Technology, vol. 12, no. 1, pp. 14-40, 2014.

[19] "Anaphora and coreference resolution: A review," ScienceDirect.

[20] "Common and uncommon ground: Social and structural factors in codeswitching," Language in Society.

[21] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg:

Current challenges and new directions in sentiment analysis research," IEEE Transactions on Affective Computing, 2020.

[22] "Multilingual Sentiment Analysis: How to Do It," Crisol, 5 octobre 2021.

[23] "Challenges & Methods for Multilingual Sentiment Analysis in 2023.

[24] "A survey on sentiment analysis methods, applications, and challenges," SpringerLink.

[25] "GetOldTweets3: Get old tweets from Twitter,". https://github.com/Mottl/GetOldTweets3.

[26] J. Henrique, "Get Old Tweets Programatically," 24 mai 2023. https://github.com/Jefferson-Henrique/GetOldTweets-python.

[27] S. Brachemi-Meftah and F. Barigou, "Algerian Dialect Sentiment Analysis: State of Art," in 2020 21st International Arab Conference on Information Technology (ACIT), Nov. 2020, pp. 1-7, doi: 10.1109/ACIT50332.2020.9300060.

[28] X. Tan, Y. Cai, J. Xu, H.-F. Leung, W. Chen, and Q. Li, "Improving aspect-based sentiment analysis via aligning aspect embedding," Neurocomputing, vol. 383, pp. 336-347, Mar. 2020. doi: 10.1016/j.neucom.2019.12.035.

[29] "Sentiment Classification System of Twitter Data for US Airline Service Analysis" https://ieeexplore.ieee.org/abstract/document/8377739/.

[30] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?" arXiv, 5 février 2020. [En ligne]. Disponible : http://arxiv.org/abs/1905.05583.

[31] "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond," https://aclanthology.org/2022.lrec-1.27/.

[32] N. A. S. Abdullah and N. I. A. Rusli, "Multilingual Sentiment Analysis: A Systematic Literature Review," Pertanika J. Sci. Technol., vol. 29, no. 1, 2021, doi: 10.47836/pjst.29.1.25.

[33] F. M. Kundi, A. Khan, S. Ahmad, and M. Z. Asghar, "Lexicon-based sentiment analysis in the social web," J. Basic Appl. Sci. Res., vol. 4, no. 6, pp. 238-248, 2014.