Mınistère De L'enseignement Supérieur Et De La Recherche Scientifique

Université Mohamed El-Bachir El-Ibrahimi

Bordj Bou-Arreridj

Faculté Mathématique Et Informatique

Département Informatique

Mémoire de master

Prédiction des protestations publiques à l'aide d'algorithmes de classification

---

Spécialité :
Réseaux et Multimédias

Présenté par :

- MEKHOUKH Wafa.
- TEHAMI Noura.

Proposé et dirigé par :

Dr. LAIFA Meriem.

Devant le jury composé de :

- Dr. BELAZOUG Mouhoub.
- Dr. MOHDEB Djamila.

Année universitaire : 2019-2020

Remerciement

Au début, nous remercions Allah qui nous a aidés à accomplir ce travail, et qui a été avec nous à tous les moments de notre chemin d'étude.

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

Nos remerciements et notre gratitude vont aux professeurs et enseignants de département de l'informatique ainsi que ses étudiants et son personnel côtoyés tout au long de notre cursus universitaire.

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

Que toute personne ayant œuvré de près ou de loin à la réalisation de ce projet par une quelconque forme de contribution, trouve ici le témoignage de notre plus profonde reconnaissance.

Dédicace

À nos chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de nos études.

À nos chers frères et sœurs pour leurs encouragements permanents, leur soutien moral, leur appui et leur encouragement,

À toutes nos familles et amis pour leur soutien tout au long de notre parcours universitaire.

Que ce travail soit l'accomplissement de vœux tant allégués, et le fruit de soutien infaillible.

Merci d'être toujours là pour nous.

Table des matières

Table des matières

Table des figures

Liste des tableaux

Résumé

Résumé

L’objectif primordial de ce mémoire est de prédire les manifestations publiques au moyen d'algorithmes d'apprentissage automatique utilisant les fonctionnalités extraites des données des médias sociaux. En particulier, nous considérons le cas de « Hirak » qui a commencé en février 2019 en Algérie. L’objectif sera aussi de proposer un modèle de prédiction basé sur les méthodes de classification pour prédire les manifestations de masse sur la base du contenu public des médias sociaux à partir de Twitter. Afin d’atteindre les objectifs cités, nous fournissons un aperçu méthodologique de quatre méthodes de classification essentielles de l'apprentissage automatique, et nous comparons leur efficacité dans la classification des Tweets de protestation. Nous observons une bonne précision de classification de 74%, avec la méthode de régression logistique. Et nous observons également que les autres méthodes offrent une précision raisonnable entre 60% et 72%.

Abstract

The primary objective of this memory is to predict public protests by means of machine learning algorithms using features extracted from social media data. In particular, we consider the case of « Hirak » which started in February 2019 in Algeria. The goal will also be to come up with a classification-based prediction model to predict mass protests based on public social media content from Twitter. In order to achieve the stated goals, we provide a methodological overview of four essential machine learning (ML) classification methods, and we compare their effectiveness in classifying Protest Tweets. We observe a good classification precision of 74%, with the logistic regression method. And we also observe that the other methods offer a reasonable precision between 60% and 72%.

ملخص

الهدف الأساسي لهذه المذكرة هو التنبؤ بالاحتجاجات العامة عن طريق خوارزميات التعلم الآلي باستخدام ميزات مستخرجة من بيانات وسائل التواصل الاجتماعي. على وجه الخصوص ، ننظر في قضية "الحراك" التي بدأت في فبراير 2019 في الجزائر. سيكون الهدف أيضًا هو الخروج بنموذج تنبؤ قائم على التصنيف للتنبؤ بالاحتجاجات الجماهيرية بناءً على محتوى وسيلة التواصل الاجتماعي تويتر. من أجل تحقيق الأهداف المذكورة ، نقدم نظرة عامة منهجية لأربع طرق تصنيف أساسية للتعلم الآلي ، ونقارن فعاليتها في تصنيف التغريدات الاحتجاجية. نلاحظ دقة تصنيف جيدة بنسبة 74٪ بطريقة الانحدار اللوجستي. ونلاحظ أيضًا أن الطرق الأخرى توفر دقة معقولة بين 60٪ و 72٪.

CHAPITRE I

Introduction Générale

1. Les médias sociaux, les mouvements sociaux et l'apprentissage automatique

Un réseau est une façon de réflexion sur les systèmes qui focalisent notre attention sur les relations entre les entités qui composent ce système, que nous appelons acteurs ou nœuds. Les nœuds ont des caractéristiques généralement appelées « attributs », et ils peuvent s'agir de traits catégoriels. Les relations entre les nœuds ont également des caractéristiques, et dans l'analyse de réseau, nous les considérons comme des types de liens [1].

Une question évidente à poser est pourquoi quelqu'un voudrait analyser les données des réseaux. La réponse est qu'une grande partie de la culture et de la nature semble être structurée en réseaux des cerveaux et réseaux biologiques (réseaux de neurones), et des organismes (systèmes circulatoires) aux organisations (qui rapporte qui), aux économies (qui vend à qui). Il existe aussi les réseaux de connaissances, de collaborations, et de technologies (l'internet, le Web mondial) [2].

En outre, les médias sociaux désignent un ensemble de services permettant de développer des conversations et des interactions sociales sur internet ou en situation de mobilité. Ils utilisent des techniques de communication aisément accessibles pour faciliter les interactions sociales qui se fondent sur l'idéologie et la technologie du Web 2.0. Ces technologies permettent en particulier la création et l'échange de contenus générés par les utilisateurs [3].

Les réseaux sociaux sont une des catégories des médias sociaux. Ils sont des communautés de personnes qui partagent des intérêts communs. Ces types de réseaux sont utilisés par des millions de personnes à travers le monde. Ils s'agissent d'un site Internet qui permet à ses utilisateurs de créer une page de profil personnalisable pour partager et échanger des informations, des photos et des vidéos à sa communauté d'amis et son réseau de connaissances. Exemple : Facebook, Snapchat…etc. [4].

Il convient de mentionner que les médias sociaux et les réseaux sociaux en ligne ne sont pas identiques. Les médias sociaux sont encore des médias qui sont principalement utilisés pour transmettre ou partager les informations entre les personnes, tandis que les réseaux sociaux sont un acte d'engagement, car les personnes ayant des intérêts communs s'associent et établissent des relations par l'intermédiaire de la communauté. En d'autres termes, les médias sociaux ne sont qu'un système, un canal de communication, ce n'est pas un endroit que l'on visite. En revanche, les réseaux sociaux sont une communication bidirectionnelle, où les conversations sont au cœur, et par laquelle les relations se développent [5].

Les réseaux sociaux en ligne sont utilisés pour de nombreux objectifs. Par exemple, ils sont devenus comme un outil majeur de diffusion des messages de campagne politique [6]. Un autre exemple est l'utilisation des réseaux sociaux pour le marketing.

Un mouvement social est une série de manifestations. Il peut aussi être compris comme des réseaux de personnes rassemblées par un but ou un intérêt commun. Les mouvements sociaux en tant que réseaux sociaux peuvent aussi être lus en termes d'un noyau initial composé de groupes densément connus de liens plus forts qui mobilisent ensuite des individus faiblement liés.

Les technologies numériques, et en particulier les médias sociaux, ont joué un rôle clé dans la transformation des contextes et des processus des mouvements sociaux contemporains. Leur accessibilité croissante a changé les ressources, les évolutions et les résultats des actions collectives. D'un autre côté, les questions liées à la participation progressivement compliquée des agents humains et technologiques aux mouvements sociaux méritent l'attention. Nous avons observé des changements dans la façon dont les différents acteurs abordent la protestation et la résistance. Afin de réussir dans un mouvement social, les gens essaient de planifier et d'annoncer à l'avance pour encourager et rassembler une plus grande participation pour renforcer l'influence des foules. Pour cela, les médias sociaux offrent des opportunités exceptionnelles d'organiser des masses de personnes dans des actions démocratiques avec des dépenses de participation plus faibles, et de favoriser de nouveaux répertoires d'informations et d'actions qui vont au-delà des communautés hors ligne.

L'apprentissage automatique consiste à extraire et traiter les informations à partir des données. Il facilite l'automatisation des tâches et renforcent les connaissances dans le domaine humain.[7]. Il s'agit d'un domaine de recherche à l'intersection des statistiques, de l'intelligence artificielle et de l'informatique, et il est aussi connu sous le nom d'analyse prédictive ou d'apprentissage statistique. L'application des méthodes d'apprentissage automatique est devenue partout ces dernières années dans la vie quotidienne. L'apprentissage automatique peut se catégoriser selon le type d'apprentissage qu'ils emploient : l'apprentissage supervisé, l'apprentissage non-supervisé et l'apprentissage par renforcement. [8].

L'objectif de l'apprentissage automatique est de comprendre la structure des données et de les intégrer dans des formulaires que tout le monde peut comprendre et utiliser. Il facilite l'utilisation d'ordinateurs dans la construction de modèles d'échantillonnage de données pour calculer des problèmes ou les résoudre grâce à la saisie de données et à l'utilisation d'analyses statistiques [9].

2. Objectif et contribution

Au cours de la dernière année (2019), les médias sociaux ont été largement utilisés comme outils d'organisation, d'information et de mobilisation pour les manifestations en Algérie. Dans ce contexte, nous visons à prédire les protestations publiques au moyen d'algorithmes d'apprentissage automatique utilisant les fonctionnalités extraites des données des médias sociaux appelant à la démonstration. En particulier, nous considérons le cas de « Hirak » qui a commencé en février 2019. Malgré l'importance de cet événement historique, il y a un manque critique de données fiables et préci ses ainsi que d'études scientifiques qui retracent l'action collective des manifestants sur les réseaux so ciaux et les médias. L'approche proposée devrait être basée sur des caractéristiques régionales tirées des messages d'utilisateurs agrégés et de leur contenu. L'objectif principal est de proposer un modèle de prédiction basé sur la classification pour prédire les manifestations de masse sur la base du contenu public des médias sociaux. La première phase consiste à rechercher les premiers signaux d'une manifestation tels que les hashtags. Ensuite, les données seront collectées et préparées pour l'extraction des fonctionnalités et la formation du modèle. Nous soulignons également la recherche et les questions connexes en cours, ainsi que les travaux futurs qui nécessitent une réflexion plus approfondie.

3. Organisation du rapport

Notre travail est divisé en deux parties, partie théorique et partie pratique. Dans la première partie nous parlons sur Twitter et son langage, quelques recherches connexes dans le cadre de notre travail, et aussi nous parlons brièvement du Hirak, tout ça dans le deuxième chapitre qui présente l’état de l’art.

Dans la deuxième partie, on représente le troisième chapitre qui explique la méthode d’implémentation de notre travail, en expliquant l'ensemble des choix techniques, (langage de programmation python). Et on fait une vision générale sur les méthodes de classification. Le quatrième chapitre est consacré aux expérimentations et résultats.

CHAPITRE II

Etat de l’art

1. Introduction

Avec la popularité croissante des médias sociaux et des plateformes de réseaux en tant que médias de communication et de partage, ils ont contribué de manière significative au processus de prise de décision dans divers domaines. Au cours de la dernière décennie, Twitter est devenu une source importante de données générées par les utilisateurs. En conséquence, d'importantes avancées technologiques ont été réalisées dans le traitement et l'analyse des données sur les médias sociaux à l'aide de techniques provenant de différents domaines tels que l'apprentissage automatique, le traitement des langues naturelles, les statistiques et le web sémantique.

Dans ce chapitre, nous fournirons un aperçu sur Twitter et son langage spécial de base car il sera la principale source de nos données. Après, nous soulignerons certaines recherches connexes qui ont été effectuées dans le cadre de notre travail. Enfin, nous parlerons brièvement du Hirak qui est le contexte principal de notre étude.

2. Twitter

Au premier trimestre de 2020, le nombre d'utilisateurs actifs était de 336 millions par mois et 145 millions par jour. De plus, près de 500 millions de tweets par jour, 6 000 tweets par seconde sont partagés sur Twitter [10]. Ce dernier fournit des données multimodales contenant du texte, des images et des vidéos, ainsi que des métadonnées contextuelles et sociales telles que l'information temporelle et spatiale, et des informations sur la connectivité et les interactions des utilisateurs. Ces riches données générées par les utilisateurs jouent un rôle important dans la collecte de signaux agrégés à partir du contenu et dans le sens des opinions et des réactions du public face aux problèmes contemporains. Les données Twitter peuvent être utilisées pour des analyses prédictives dans de nombreux domaines d'application, allant de la santé personnelle et sociale à la santé publique et à la politique. L'analyse prédictive des données sur Twitter comprend une collection de techniques pour extraire des informations et des modèles des données, et prédire les tendances, les événements futurs, et les actions basées sur les données historiques.

2.1. Les caractéristiques des tweets

De nouvelles techniques de traitement et d'analyse sont nécessaires pour comprendre et obtenir des indications fiables pour prédire les tendances et les événements futurs à partir des données de Twitter en raison de leur nature unique : elles contiennent des argots, des abréviations non conventionnelles et des erreurs grammaticales, bien sûr. En outre, en raison de l'évolution de nombreux événements, qu'il s'agisse de manifestations politiques, sportives ou liées à une catastrophe, la collecte d'informations pertinentes au fur et à mesure de l'événement est cruciale. Surmonter les défis posés par le volume, la rapidité et la diversité des données sociales à l'avenir n'est pas une tâche facile. L'analyse par mots clés souffre d'une faible précision ainsi que d'un faible taux de récupération. Pour améliorer cette baisse, Sheth et al.[11] ont développé une solution qui est de suivre l'utilisateur à travers ses commentaires et son profil et les a incarnés sur un graphique de connaissances large. Dans l'analyse des données twitter, la phase de traitement comprend le traitement du langage naturel à l'aide de techniques telles que TF-IDF, Word2Vec, la fragmentation, l'élimination des mots avec une occurrence rare et le codage. D'autre part, certaines méthodes couramment utilisées, telles que l'élimination des mots d'arrêt, se sont avérées inefficaces.

2.1.1. Nature unique des tweets

Sur Twitter, des abréviations non traditionnelles, une satire d'orthographe incorrecte et des erreurs grammaticales peuvent être utilisées dans le message. Les techniques et les données descriptives identifient les meilleures caractéristiques pour améliorer les performances globales d'un modèle en construction. En raison de la nature hétérogène du contenu Twitter, des fonctionnalités telles que le texte, le langage, le visuel et la sémantique sont développées sur les métadonnées tweet utilisateur. De plus, pour gérer les données textuelles du tweet, les fonctions, techniques et outils extraits ont été personnalisés pour exploiter et être robustes en ce qui concerne les fautes d'orthographe et les abréviations. Les Tweets contiennent des hashtags, URL, émoticônes, mentions et emojis dans leur contenu qui contribuent à leur signification.

- **Hashtag** est utilisé pour catégoriser les Tweets. Il est fréquemment utilisé pour collecter et filtrer des données, des sentiments et des analyses thématiques. Wang et al.[12] ont utilisé des hashtags dans leur analyse émotionnelle et recueilli environ 2,5 millions de tweets contenant des hashtags liés aux émotions comme #happy, #annoyed, et les ont utilisés comme programme de formation.

- **URL :** (ex http://votresite.e-monsite.com). La présence d'URL dans un tweet indique généralement que le contenu est un index pour une histoire explicative plus longue pointée par l'URL. Les chercheurs ont rapporté que sa présence dans un tweet est une caractéristique majeure ou une contribution significative à l'exactitude du modèle [13-14].

- **Les émoticônes** (ex :), <3) ont été exploitées par Liu et al [15] dans leur étude d'analyse du sentiment sur Twitter (sentiment positif ou négatif). Ils ont utilisé tous les tweets contenant ces émoticônes comme un ensemble de formation auto étiqueté et les ont intégrées à l'ensemble de formation étiqueté manuellement. Ils ont réalisé une amélioration significative par rapport au modèle formé avec seulement des données étiquetées manuellement.
- **Emoji** est une représentation picturale d'expressions faciales, de lieux, de nourriture et de nombreux autres objets, utilisés très souvent sur les médias sociaux pour exprimer des opinions et des émotions sur les questions contemporaines de querelles et de discussions. L'utilisation d'emoji est similaire à l'émoticône, car elles fournissent toutes deux un moyen plus court d'expression d'une idée et d'une pensée. La différence, c'est qu'un emoji utilise une petite image pour la représentation par opposition à l'émoticône qui utilise une séquence de caractères. Kelly et al [16] ont étudié l'utilisation des emoji dans différents contextes en menant des entrevues et ont constaté que l'utilisation d'emoji va au-delà du contexte que le concepteur avait l'intention.

2.1.2. Métadonnées pour les utilisateurs et les tweets

Il existe principalement deux types de métadonnées dans un objet de tweet, les métadonnées de tweet et les métadonnées de l'utilisateur.

❖ **Les métadonnées de tweet :** contiennent des informations temporelles et spatiales, ainsi que les interactions entre les utilisateurs et d'autres informations telles que les réponses et la langue.

- ❑ **CreateAt :** contient des informations sur la date de création du tweet. Ce qui est particulièrement important lorsqu'une analyse des séries chronos est en cours.
- ❑ **FavoriCount :** Les utilisateurs de Twitter peuvent aimer un tweet, et c'est une façon d'interagir avec la plateforme. Le nombre de J'aime pour un tweet a été utilisé comme une fonctionnalité dans diverses applications qui comprennent la détection des tendances, et l'identification de l'influence et la popularité.
- ❑ **InReplyToScreenName :** Si ce champ de l'objet de tweet n'est pas nul, il s'agit d'une réponse à un autre tweet, et ce champ contiendra le nom d'utilisateur qui a écrit l'autre tweet. Ces informations sont précieuses, en particulier pour prédire l'engagement du public sur un problème auquel les tweets se rapportent.

- ❑ **GeoLocation :** associe la géolocalisation des utilisateurs au tweet. Mais c'est aux utilisateurs de le rendre public. La plupart des utilisateurs préfèrent ne pas partager leur géolocalisation.
- ❑ **Retweet_count :** Twitter permet aux utilisateurs de retweeter leur audience, et le tweet original contient ce champ pour garder le nombre de fois que le tweet a été retweeté. Cette information est utile pour intégrer la prédiction de la popularité et des sujets de tendance.

❖ **Les métadonnées de l'utilisateur :** contiennent des informations relatives à l'utilisateur qui a écrit le tweet, telles que le nom et la description de l'écran.

- ❑ **Description :** ces métadonnées contiennent des informations sur les caractéristiques de l'utilisateur, elles sont principalement utilisées dans la classification des utilisateurs.
- ❑ **Followers_count :** ce champ contient le nombre d'abonnés que possède l'utilisateur, et comme il est des informations modifiables au fil du temps, les informations situées dans un tweet spécifique peuvent ne pas être à jour.
- ❑ **Friend_count :** Twitter appelle les comptes qu'un utilisateur suit comme des "amis", ou "followers". Le nombre de followers et followees est utilisé pour déterminer la popularité des utilisateurs et des sujets.
- ❑ **Status_count :** Twitter appelle aussi les tweets "status", et dans ce cas, "status count" fait référence au nombre de tweets publiés par un utilisateur.

2.2. Les mesures de réseau et les caractéristiques statistiques

Les utilisateurs interagissent sur la plate-forme de réseau social Twitter les uns avec les autres grâce à des suivis, des réponses, des retweets, des "J'aime", et des mentions. Des mesures de centralité ont été développées pour calculer et révéler la position de l'utilisateur et leur importance en fonction de leurs connexions dans leur réseau. Ces mesures de centralité peuvent aider à identifier les utilisateurs influents. Ces mesures comprennent in-degree, out-degree, closeness, betweenness, PageRank, and eigenvector centrality. Closeness centrality est définie par Freeman [17] comme la somme des distances de tous les autres nœuds, où la distance d'un nœud à l'autre est définie comme la longueur (en liens) du chemin le plus court de l'un à l'autre. Plus la valeur de la centralité de proximité est petite, plus le nœud est central.

L'entre-deux mesures mesure la connectivité d'un nœud en calculant le nombre de chemins les plus courts qui traversent le nœud. Cet aspect fait de ce nœud, un utilisateur d'un réseau social Twitter, une partie essentielle du réseau car il contrôle le flux d'informations dans le réseau. Par conséquent, la suppression de ce nœud déconnectera le réseau. EigenVector [18-19] mesure l'importance d'un nœud en fonction de l'importance de ses connexions au sein du réseau. Par conséquent, plus un nœud se restérilisé, plus le nœud devient critique.

Ces mesures ont été utilisées dans une application de classification des utilisateurs comme caractéristiques par Wagner et al. [20] en raison de l'intuition que des utilisateurs similaires auraient des caractéristiques de connectivité réseau similaires.

Les caractéristiques statistiques telles que min, max, médiane, moyenne, déviation standard, biais, kurtosis et entropie peuvent être calculées pour plusieurs attributs de données. L'apprentissage automatique détermine un sous-ensemble de ces fonctionnalités qui ont le pouvoir discriminatif nécessaire pour des applications et des domaines particuliers, en particulier pour prédire les comportements et les types des utilisateurs. Par exemple, extraire les caractéristiques statistiques d'un utilisateur, tweet, réseau. L'analyse statistique a été faite sur des attributs tels que le nombre de suiveurs de l'expéditeur, le nombre de suivis de suiveur, le temps entre deux tweets consécutifs, et le nombre de hashtags dans un tweet. Ils ont effectué une analyse de séries chronomètres pour prédire si une même tendance est organique ou promue par un groupe. D'autre part, a utilisé des caractéristiques statistiques pour prédire le type d'utilisateurs sur les médias sociaux en fonction de leurs penchants politiques, l'ethnicité, et l'affinité pour une entreprise particulière. Au fur et à mesure qu'ils classaient les utilisateurs, ils calculaient les caractéristiques statistiques du comportement de tweeting des utilisateurs tels que le nombre moyen de messages par jour, le nombre moyen de hashtags et d'URL par tweet, ainsi que le nombre moyen et l'écart standard des tweets par jour.

2.3. Apprentissage automatique et intégration de mots

Les algorithmes d'apprentissage automatique jouent un rôle crucial dans l'analyse prédictive des relations de modélisation entre les caractéristiques pour la tâche de classification ou de régression. Une analyse comparative des études connexes a été réalisée et a abouti à plusieurs algorithmes, y compris des forêts Random, Naïve Bayes, Support Vector Machine, Artificial Neural Networks, ARIMA et Logistic Regression.

En outre, l'apprentissage en profondeur (alias l'apprentissage automatique avancé) est une stratégie visant à minimiser l'effort humain sans compromettre le rendement. C'est grâce à la capacité des réseaux neuronaux profonds à apprendre des représentations complexes à partir de données à chaque couche, il améliore l'efficacité du traitement des données volumineuses. Et aussi a utilisé pour la prédiction sur les médias sociaux.

Des études antérieures utilisent plusieurs méthodes comme TF-IDF pour obtenir des représentations textuelles de caractéristiques. Maintenant, une politique appelée Word2Vec est utilisée pour représenter numériquement un mot qui saisit sa signification contextuelle en incorporant ses mots proches dans une phrase. Ont utilisés cette méthode pour améliorer encore la prédiction des membres de gangs sur Twitter en formant leur modèle sur un corpus spécifique à un problème.

3. Travaux connexes

De nombreuses études ont montré que twitter a joué un rôle clé dans les récentes manifestations comme celles qui ont mené au Printemps arabe [21] - [24], aux émeutes de Londres [25] - [27] et Occupy Wall Street [28] - [30]. Dans la littérature, il y a beaucoup d'analyses prédictives sur Twitter [33], certaines de ces études effectuent des analyses de contenu et de sentiment sur Twitter pendant les événements et les manifestations [28], [32], [33]. D'autres études étudient l'utilisation de Twitter en temps de crise et de catastrophe naturelle [34] - [37] ou décrivent, modélisent et interprètent les réseaux d'utilisateurs et les relations entre les réseaux sociaux ainsi que les mouvements sociaux [38] - [42]. D'un autre côté, de nombreuses études proposent des modèles pour prédire les manifestations en utilisant Twitter. Comme nos travaux relèvent de cette catégorie, nous représenterons brièvement certains travaux qui ont inspiré notre projet.

Nous observons deux approches principales de la prédiction des manifestations : la première est fondée sur les propriétés des structures des réseaux d'utilisateurs en ligne [43], [44], les interactions sur les médias sociaux, et les cascades d'activités [45], [46], tandis que la seconde est fondée sur les caractéristiques induites par les publications agrégées d'utilisateurs et leur contenu.

Dans notre étude, nous adoptons la deuxième approche pour la prédiction et nous présentons dans cette section quelques études importantes dans la littérature.

Alberto et Victor [47] ont utilisé une méthodologie mixte d'analyse de contenu et d'analyse textuelle de 784 tweets pour identifier les principaux sujets de partage de contenu liés à la dénonciation de la violence policière dans les manifestations sociales en Espagne.

Compton et al. [48] ont réalisé une analyse de contenu sur les tweets pour trouver ceux importants contenant les mentions temps et lieu de futures manifestations afin de détecter les manifestations potentielles. Muthiah et al. [49] ont élaboré un système fondé sur l'analyse du contenu et de la langue pour prévoir l'intervalle de temps et le lieu des troubles civils potentiels. En appliquant leur système à 10 pays d'Amérique Latine, ils ont montré des efforts pour détecter le moment et le lieu de manifestations importantes. Radinsky et Horvitz [50] ont utilisé une base de données de 22 ans et étudié les séquences de différents événements pour prédire si un événement d'intérêt se produira à l'avenir.

Kallus [22] a utilisé les données de plus de 300 000 sites Web différents rassemblés par Recorded Future[1] pour prédire les protestations importantes en utilisant des méthodes d'apprentissage automatique. Il a étudié le cas de l'Egypte pendant le Printemps arabe. Steinert-Threlkeld et al. [21] ont également étudié l'affaire du Printemps arabe en utilisant environ 14 millions de tweets collectés dans 16 pays et a montré qu'il y a une forte relation statistique entre les activités de protestation d'une journée donnée et le niveau de coordination de la veille. Korolov et al. [51] ont étudié les manifestations de

[1] https://www.recordedfuture.com/fr/

Baltimore en 2015. A l'aide de méthodes d'analyse du contenu des tweets, ils ont classifié quatre types de tweets de mobilisation : sympathie pour la cause, prise de conscience de la manifestation, motivation à y participer et capacité à y participer. Puis ils ont montré qu'il y a une corrélation entre la combinaison linéaire du nombre de tweets de ces quatre types et la réalité des manifestations.

Ramakrishnan et al. [52] et Doyle et al. [53] ont tous deux décrit l'architecture de conception du système EMBERS. EMBERS est un système automatisé intelligent conçu pour prévoir les actions collectives telles que les manifestations et les résultats des élections dans les pays d'Amérique latine. Le système EMBERS collecte ses données à l'aide de nombreuses sources de données ouvertes comme les agences de presse et les médias sociaux, en particulier Twitter. Ce système utilise des fonctions pilotées par le contenu des messages et les statistiques associées ainsi que des modèles de cascade d'activité, puis effectue la sélection des fonctions par régression LASSO et finalement il combine plusieurs classificateurs plutôt qu'un modèle de prévision afin d'obtenir des résultats plus fiables.

De plus, Korkmaz et al. [54] ont conçu un système de prédiction (intégré dans EMBERS), et en utilisant les données recueillies sur Twitter et les blogs dans six pays d'Amérique Latine, ils ont montré que les sources de données hétérogènes peuvent collectivement augmenter la précision dans la prédiction de futures manifestations.

4. Le Hirak

Le Hirak (en Arabe : الحراك, en Français : Mouvement) fait référence à une série de manifestations périodiques qui ont eu lieu en Algérie depuis le 16 février 2019 [55] pour exposer dans un premier temps contre la candidature d'Abdelaziz Bouteflika pour un cinquième mandat présidentiel. Bien que la grande manifestation ait eu lieu le 16 février 2019, il y avait eu des appels à des manifestations en décembre 2018 [56, 57] qui n'ont pas été entendus par la plupart des Algériens et les médias, mais ont poussé la police à mobiliser une force dissuasive importante.

Cependant, depuis le début officiel du Hirak, des manifestations ont été organisées via les plateformes de médias sociaux appelant à une énorme manifestation dans les grandes et moyennes villes le vendredi 22 février 2019. Les experts ont estimé que le nombre de manifestants se situait entre 800 000 et 1 million ce jour-là. [58, 59]. Le mouvement a duré plus d'un an jusqu'à ce que le COVID-19 atteigne l'Algérie en Mars 2020 [60], ce qui a provoqué le verrouillage de nombreuses villes et interdit les rassemblements.

5. Conclusion

Comprendre le contexte de l'étude ainsi que la littérature connexe est une étape importante pour comprendre les chapitres suivants. Ensuite, nous nous concentrerons sur la méthodologie que nous avons utilisée pour atteindre nos objectifs et nous expliquerons chaque étape en détail.

CHAPITRE III

Méthodologie et implémentation

1. Introduction

Afin d'atteindre notre objectif, qui est de prédire les protestations liées au "Hirak" à partir de Tweets, nous avons suivi 05 étapes principales telles que présentées dans l'organigramme ci-dessous (figure 01). Dans un premier temps, nous allons extraire les données associées de Twitter et les préparer pour les étapes suivantes. Étant donné que les données extraites sont textuelles, nous devons effectuer des étapes de nettoyage et de prétraitement du Traitement du Langage Naturel. Ensuite, nous nous concentrons sur l'Analyse Exploratoire des Données et l'extraction des caractéristiques basées sur les ensembles de données. Une fois les caractéristiques sont sélectionnées et extraites, nous sélectionnons, formons et évaluons les classificateurs pour la prédiction. Dans les sections suivantes, nous décrivons en détail l'approche utilisée à chaque étape. Dans ce chapitre, nous expliquerons les différentes étapes que nous avons suivies pour atteindre nos objectifs, où les résultats et l'analyse seront fournis dans le chapitre suivant.

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

2. Collecte et préparation des données

Compte tenu du fait que nous avons collecté les données après les manifestations, nous avons déjà des informations sur les événements majeurs ainsi que des mots-clés et hashtags associés en nous appuyant sur les journaux nationaux et internationaux [61-65] comme source fiable d'information. Ensuite, Les mots clés et les hashtags principales ont été choisis pour l'extraction des tweets. Comme: *22 Février*, بوتفليقة عبد العزيز, العهدة الخامسة, …etc. Nous nous concentrons sur les quatre événements principaux suivants:

- 22 Février 2019 : cet événement marque le début du Hirak.
- 08 Mars 2019 : cet événement rejoint la journée internationale de la femme. De nombreuses femmes (jeunes et vieilles) ont participé à cet événement nouveau pour la société conservatrice algérienne.

- 05 Juillet 2019 : cet événement rencontre le jour de l'indépendance algérienne.
- 1er Novembre 2019 : cet événement s'est produit à la mémoire de la guerre de révolution algérienne, il a donc une forte signification pour les Algériens.

Pour chaque événement, nous avons extrait des tweets pendant une durée d'un mois jusqu'à la survenue de l'événement (y compris le jour de l'événement) à l'aide de hashtags et de mots-clés associés.

3. Langage de programmation utilisé

Il existe un très grand nombre de langages de programmation, chacun avec ses avantages et ses inconvénients. Notre choix s'est porté sur Python avec les avantages suivants : il s'agit d'un langage gratuit interprété et orienté objet qui peut être utilisé pour différents objectifs allant du développement d'un site Web fonctionnel complet à la manipulation d'un robot. De plus, il existe plusieurs environnements de développement intégrés multi-plateformes open source tels que PyCharm, Spyder, Jupyter project. Nous avons utilisé Jupyter Notebook car il s'agit d'une application Web open source qui nous permet de créer et de partager en direct des documents contenant du code, des équations, des visualisations et du texte narratif. Il est utile notamment pour le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation de données, l'apprentissage automatique et bien plus encore.

4. Extraction des données

Nous avons utilisé la bibliothèque **`GetOldTweets3`** [66] pour extraire des tweets de twitter.com. Initialement, l'API de recherche Twitter était utilisée. Cependant, l'API Twitter officielle impose plusieurs contraintes comme la limite d'extraction des tweets aux 7 derniers jours seulement. Pour obtenir des tweets publiés à des dates antérieures, vous avez besoin des API Premium ou Enterprise Search de Twitter, qui permettent aux utilisateurs d'obtenir des tweets vieux de 30 jours ou même l'archive complète de Twitter. Comme ça nous coûte cher, nous avons opté pour l'option open-source gratuite.

`GetOldTweets3`, une amélioration du **`GetOldTweets-python`** [67] original par Jefferson Henrique, est une bibliothèque python 3 et un utilitaire de ligne de commande correspondant pour accéder aux vieux tweets. Bien qu'il ne soit pas complet, il contourne certaines contraintes de l'API officielle de Twitter en permettant à l'utilisateur d'obtenir des tweets de plus de 7 jours, ce qui est permis car les tweets à vue publique peuvent légalement être utilisés pour la recherche universitaire. **`GetOldTweets3`** est simple à utiliser et permet à l'utilisateur de spécifier des tweets par nom d'utilisateur, popularité, requête, dates de connexion, hashtags et langue. La bibliothèque est installée en utilisant : `pip install GetOldTweets3`.

Ces données extraites contiennent les informations suivantes : date et heure, texte du tweet, nom d'utilisateur, et le nombre de retweets. La géolocalisation des tweets n'a pas été extraite car elle sort du cadre de notre étude. La communauté algérienne à l'étranger (notamment au Canada, au Royaume-Uni et en France) [68-70], a également soutenu le mouvement en Algérie avec le soutien de personnalités politiques et d'opposants.

Enfin, Les tweets obtenus qui n'ont rien à voir avec le sujet (en particulier les tweets contenant des annonces, puis utiliser les hashtags des manifestations juste pour obtenir plus de vues) ont été supprimés afin de nettoyer les tweets.

Tableau 1: Résumé des hashtags et mots clés utilisées pour l'exploration de données pour chaque événement.

Evènements	Hashtags et mots clés	Nombre des tweets avant le prétraitement	Nombre des tweets après le prétraitement
Le 22/02/2019	#حراك_22_فيفري #جمهورية_ماشي_ملكية #لا_للعهدة_الخامسة #عبد_العزيز_بوتفليقة الجزائر ضد ترشح ترشح بوتفليقة	**16966**	**16143**
Le 08/03/2019	#لا_للعهدة_الخامسة #سلمية_سلمية # حراك الطلبة حراك_8_مارس حراك عيد المرأة Algeria Protests	**17809**	**17024**
Le 05/07/2019	#Yetna7aw_Ga3 #حراك_5_جويلية # يتنحاو_قاع #حراك_الطلبة Algerie_manifestation عيد_استقلالك_ياجزاير	**8528**	**7145**
Le 01/11/2019	#حراك_1_نوفمبر # نوفمبر_الاستقلال #Algerie_Libre_Democratique #تسقط_انتخابات_العصابات تحيا الجزائر مسيرة أول نوفمبر	**8457**	**7928**

5. Prétraitement

Avant de passer directement à la classification, nous devons prétraiter les données textuelles. Le texte peut prendre diverses formes, allant d'une liste de mots individuels à des phrases en plusieurs paragraphes avec des caractères spéciaux (comme des tweets et d'autres ponctuations).

Le prétraitement consiste à transformer les données de texte brut en un format compréhensible. Les données du monde réel sont très souvent incomplètes, incohérentes et remplies de beaucoup de bruit et sont susceptibles de contenir de nombreuses erreurs. Le prétraitement est une méthode éprouvée pour résoudre de tels problèmes. Le prétraitement des données prépare les données de texte brut pour un traitement ultérieur. Il comprend de nombreuses étapes. Dans ce qui suit, nous expliquons chaque étape séparément.

5.1. Nettoyage des données

Le nettoyage des données est le premier pas vers la transformation des données. Cette tâche consiste en six tâches subordonnées pour accomplir ce processus. La vue d'ensemble des tâches subordonnées est donnée ci-après.

5.1.1. Suppression des tweets répétés

Étant donné que nous avons extrait les données à l'aide de hashtags et de mots-clés distincts, il était possible que certains tweets puissent apparaître plusieurs fois dans notre ensemble de données combiné. Par conséquent, les tweets en double ont dû être supprimés.

5.1.2. Suppression des émoticônes et des mentions

Comme les émoticônes et les mentions ne sont pas nécessaires dans notre champ d'analyse, nous les avons supprimées de notre texte de tweets en utilisant **`replace ()`** pour remplacer tous les symboles, valeurs d'émoticônes et les mentions par un espace vide, rendant les données claires.

5.1.3. Suppression de la ponctuation et les nombres

L'étape suivante consiste à supprimer la ponctuation et les nombres, car ces derniers n'ajoutent aucune information supplémentaire lors du traitement des données textuelles. Par conséquent, leur suppression nous aidera à réduire la taille des données d'apprentissage.

5.1.4. Suppression de L'URL

L'étape suivante consiste à supprimer l'URL intégrée dans les tweets, car elle ne fournit aucune information lors de l'analyse. Pour cela, nous avons utilisé **`re.split()`** en Python, qui remplace toutes les phrases et sous-parties de phrases commençant par HTTP par des espaces.

5.1.5. Conversion en minuscules

Nous avons dû convertir tous les tweets en minuscules afin d'apporter les tweets sous une forme cohérente. Ce faisant, nous pouvons effectuer d'autres transformations et classifications sans avoir à nous préoccuper de la non-cohérence des données. Cette tâche est effectuée à l'aide de la fonction **`lower()`**. Cette fonction a converti tous les alphabets en minuscules et résolu le problème de sensibilité à la casse en rendant toutes les données cohérentes dans les minuscules (les données en français ou anglais). Cela évite d'avoir plusieurs copies des mêmes mots. Par exemple, lors du calcul du nombre de mots, les termes "Analytique" et "analytique" seront considérés comme des mots différents.

5.1.6. Suppression des mots d'arrêt

Les mots d'arrêt doivent être supprimés des données de texte. Le retrait des mots d'arrêt élimine les mots courants et fréquents qui n'ont pas d'influence significative dans la phrase. Nous ne voudrions pas que ces mots occupent de l'espace dans notre base de données, ou qu'ils prennent le temps de traiter. Pour cela, nous pouvons les supprimer facilement. Dans cette tâche de prétraitement. Cette étape est effectuée en important notre liste de mots stop depuis la bibliothèque **`NLTK.Corpus`**. Et nous avons aussi créé une liste supplémentaire des mots d'arrêt algériens, par exemple [واش، لاه، هك، هذو، ماشي، كاين، ياو، بصح، راح، راكم……etc.].

6. Extraction des Caractéristiques

6.1. La segmentation de texte

La segmentation fait référence à la division du texte en unités minimales significatives. Il y a une segmentation de phrase et une segmentation de mot. Nous avons utilisé une segmentation de mots dans cette étape, qui est une étape obligatoire du prétraitement de texte pour tout type d'analyse. Il existe de nombreuses bibliothèques pour effectuer la segmentation comme **`NLTK, SpaCy`** et **`TextBlob`**.

6.1.1. La représentation des tweets par la méthode Sac des mots

Dans notre exemple, nous avons utilisé la fonction split pour transformer nos tweets en une série de mots. Ces mots sont représentés par ce qu'on appelle le sac de mots. Nous avons utilisé la méthode **`CountVectorizer`**, qui convertit ces mots en une matrice de nombres de jetons.

La méthode de sac de mots a un inconvénient. Supposons qu'un mot particulier apparaisse dans tous les documents du corpus, puis il gagnera en importance dans nos méthodes précédentes. C'est mauvais pour notre analyse.

6.1.2. La pondération des termes par la méthode TF-IDF

TF-IDF est l'un des moyens les plus efficaces de calculer le poids à terme. L'idée d'avoir TF-IDF est de réfléchir à l'importance d'un mot pour un document dans une collection, et donc de normaliser les mots qui apparaissent fréquemment dans tous les documents. La fréquence des termes, notée TF, est ce que nous avons calculé dans le modèle Sac de mots dans la section précédente. La fréquence des termes dans tout vecteur de document est désignée par la valeur de fréquence brute de ce terme dans un document particulier.

La fréquence inverse des documents indiquée par IDF est l'inverse de la fréquence des documents pour chaque terme et est calculée en divisant le nombre total de documents de notre corpus par la fréquence des documents pour chaque terme, puis en appliquant une mise à l'échelle logarithmique au résultat. Nous avons utilisé la méthode **`TfidfVectorizer`**.

7. La classification automatique

La catégorisation de texte peut être effectuée de plusieurs manières. Nous nous concentrons explicitement sur l'utilisation d'une approche supervisée utilisant la classification. Le processus de classification n'est pas limité au texte seul et est utilisé assez fréquemment dans d'autres domaines tels que la science, la santé, la météo et la technologie.

Un point important à retenir est que certaines données d'apprentissage étiquetées manuellement sont nécessaires pour la classification de texte supervisée, donc même si nous parlons de classification de texte automatisée, pour lancer le processus, nous avons besoin d'un effort manuel. Bien sûr, les avantages sont multiples car une fois que nous avons un classificateur formé, nous pouvons continuer à l'utiliser pour prédire et classer de nouveaux documents avec un minimum d'effort et une supervision manuelle.

Compte tenu du fait qu'il n'y a pas d'ensembles de données publiques disponibles sur les statistiques et les informations des manifestations algériennes dont nous avons connaissance, nous avons dû nous appuyer sur d'autres sources pour créer des ensembles de données utilisables pour notre classification. Pour cela, nous avons utilisé les sources suivantes [71, 72, 73, 74, 75] pour évaluer le nombre de manifestations qui se sont déroulées en Algérie à l'époque des événements choisis au début de l'étude. Si le nombre de manifestations a dépassé 35 (dans différentes wilayas) un jour spécifique, les tweets liés à cette date sont étiquetés comme **1** (indiquant qu'une manifestation a eu lieu ce jour-là), sinon, ils seraient étiquetés comme **0**. L'ensemble de données étiqueté sera utilisé plus tard pour l'étape d'apprentissage des modèles de classification.

Les recherches en cours ont utilisé diverses méthodes de classification textuelle pour évaluer les données sur les médias sociaux. Ces classificateurs sont regroupés en plusieurs catégories en fonction de leurs similarités. La section qui suit traite des détails sur quatre classificateurs essentiels que nous avons utilisés, y compris la régression logistique (LR), Naïve Bayes (NB), Arborescence des décisions (DT) et machine vectorielle de support (SVM).

7.1. L'algorithme de régression logistique (RL)

Cet algorithme a été nommé d'après la fonction centrale utilisée dans celui-ci qui est la fonction logistique. La fonction logistique est aussi connue sous le nom de fonction sigmoïde. Il s'agit d'une courbe en forme de S qui prend des valeurs réelles en entrée et la convertit en une plage comprise entre 0 et 1. L'une des caractéristiques de cette classification est qu'elle est fondée sur la probabilité d'un résultat est fondée sur une fonction logistique. La régression logistique est transparente et facile à comprendre, Régularisé pour éviter le surajustement, mais il est une phase d'apprentissage coûteuse [76, 77].

7.2. Classificateur de Naïve Bayes (NB)

Naïve Bayes est une technique d'apprentissage supervisée, efficacement utilisée dans la classification de texte. Il est basé sur le théorème bayésien avec l'indépendance. Le classifieur bayésien naïf est très simple et efficace. L'étude des paramètres d'évaluation de la fonction est donc très nécessaire. Étant donné que de nombreux ensembles de données textuelles sont multiclasses, il est naturel d'adapter le ratio des côtes aux problèmes multi classes, malgré la simplicité du modèle et la restriction des hypothèses d'indépendance qu'il formule. Et il a été démontré dans la pratique que des attributs fonctionnels dépendants peuvent effectivement améliorer l'exactitude de la classification dans certains cas [76, 78].

7.3. L'arbre de décision (DT)

Le classifieur de l'arbre de décision est un algorithme simple et utilisé couramment pour classifier les données. L'arbre de décision représente une structure arborescente avec des nœuds internes représentant les conditions de test et les nœuds feuilles comme étiquettes de classe. Cette approche de classification peut être appliqué à tous les types de données tels que nominal, ordinal, numérique. Les données de test sont classifiées très rapidement à l'aide d'un algorithme d'arbre de décision [77, 79].

7.4. Machine à support vectorielle (SVM)

Cet algorithme fonctionne sur une stratégie simple de séparation des hyperplans. Compte tenu des données d'apprentissage, l'algorithme classe les données de test en un hyperplan optimal. Les points de données sont tracés dans un espace vectoriel *N* dimension (*N* dépend des caractéristiques des points de données). L'algorithme SVM est utilisé pour les tâches de classification et de régression binaires [77].

8. Conclusion

Dans ce chapitre, nous avons introduit la méthodologie de préparation et de collection des données pour les prétraiter, et après nous avons extrait les caractéristiques et nous avons donné une vision générale sur quelques méthodes de classification. Nous allons voir l'analyse exploratoire des données et les résultats de classification dans le prochain chapitre.

CHAPITRE IV

Expérimentations et Résultats

1. Introduction

Après avoir les différentes étapes de préparer les données pour la classification, nous allons concentrer sur l'analyse exploratoire des données, et nous allons appliquer les méthodes de classification que nous avons expliquées dans le précédent chapitre.

Dans ce chapitre, nous décrivons en détail l'analyse exploratoire des données, nous discussions les résultats de classification et obtenir la meilleure.

2. Analyse exploratoire des données

7.1. Générer un nuage de mots

Pour analyser les tendances des tweets aux dates précises que nous avons choisies, nous nous appuyons sur la visualisation Wordcloud qui peut immédiatement nous montrer quels sont les mots les plus utilisés dans les tweets relatifs à la manifestation. Nous utilisons la bibliothèque **`wordcloud`** pour créer un nuage de mots. Nous avons généré un nuage de mots pour les 100 premiers mots, où plus un mot est populaire, plus le mot sera gros dans le nuage de mots.

Figure 2: Nuage de mots pour le 22-02-2019.

On peut facilement remarquer que les 100 mots les plus utilisés dans les tweets collectés reflètent les principales préoccupations et objectifs du peuple algérien pendant ces périodes. Dans la **figure 02** par exemple, on remarque que la plupart des mots sont tournés vers l'ancien président Bouteflika et sa nomination pour un 5e mandat, lorsque les premiers appels à manifester ont commencé à se répandre. D'autre part, dans la **figure 03**, Les tweets ont souligné l'importance de maintenir les manifestations pacifiques et respectueuses, en particulier que les femmes ont commencé à se joindre dans les rues. Il est également à noter que les femmes sont clairement invitées à rejoindre, le 8 mars 2019 étant un vendredi.

Figure 3: Nuage de mots pour le 08-03-2019.

Figure 4 : Nuage de mots pour le 05-07-2019.

Figure 5 : Nuage de mots pour le 01-11-2019.

2.2. Popularité des hashtags

Nous avons extrait les tweets à l'aide d'un terme de recherche prédéfini qui contient une liste de hashtags spécifiques, relatifs aux manifestations. En outre, les tweets peuvent aussi contenir d'autres hashtags qui ne sont pas définis dans le terme de recherche, à condition que le tweet contienne des hashtags qui sont définis par le terme de recherche. Dans cette section, nous voulons découvrir quels sont les hashtags les plus populaires utilisés par les utilisateurs de Twitter durant les périodes choisi pour cette étude.

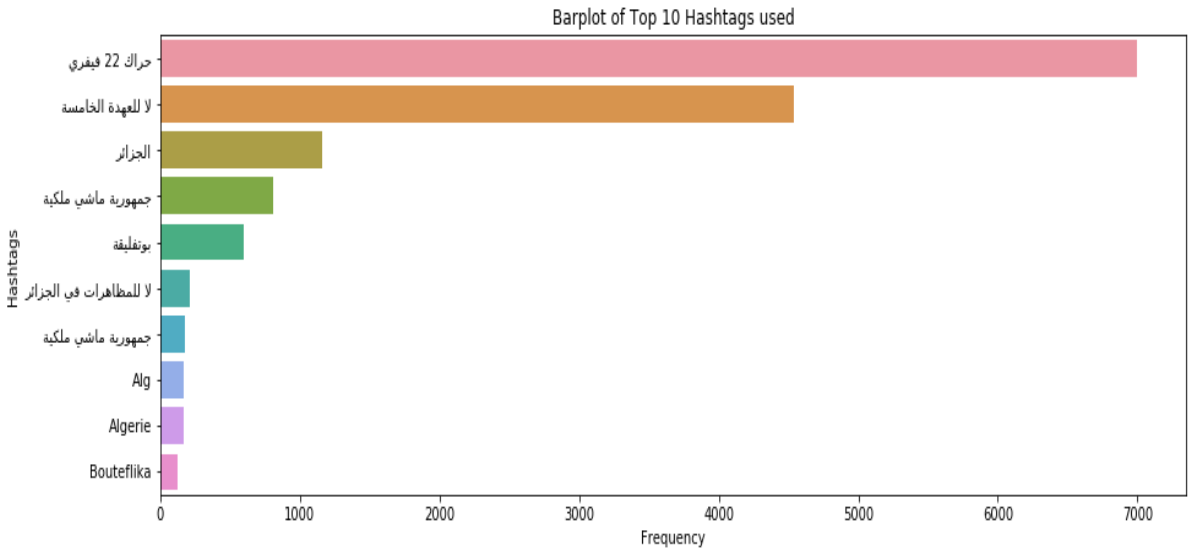

Figure 6 : Top 10 Hashtags avant et durant le 22-02-2019.

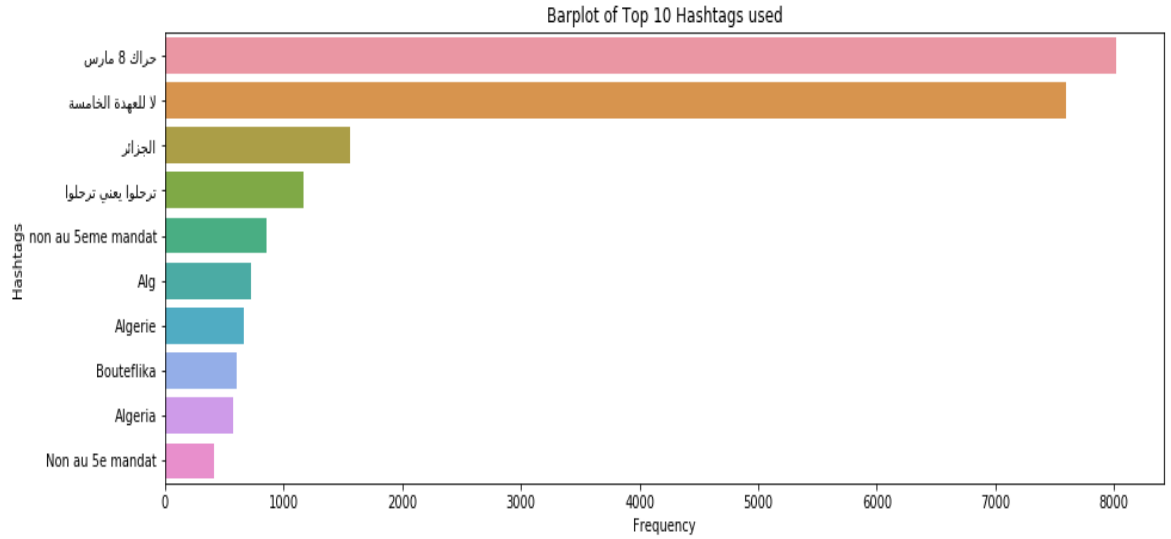

Figure 7 : Top 10 Hashtags avant et durant le 08-03-2019.

Notamment, il est possible d'identifier des périodes et les événements importants dans les manifestations en surveillant la popularité et la tendance quotidiennes des hashtags utilisés par les utilisateurs. En théorie, cela signifie que l'on pourrait simplement surveiller les hashtags et se passer de lire les nouvelles ou de faire défiler les médias sociaux pour se tenir au courant du mouvement de protestation.

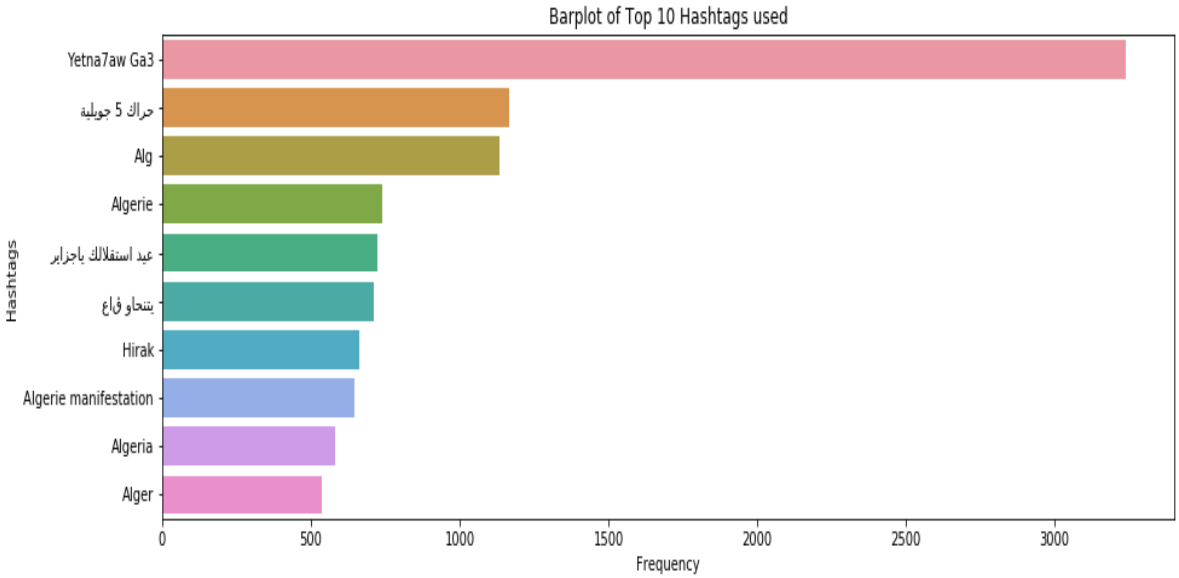

Figure 8 : Top 10 Hashtags avant et durant le 05-07-2019.

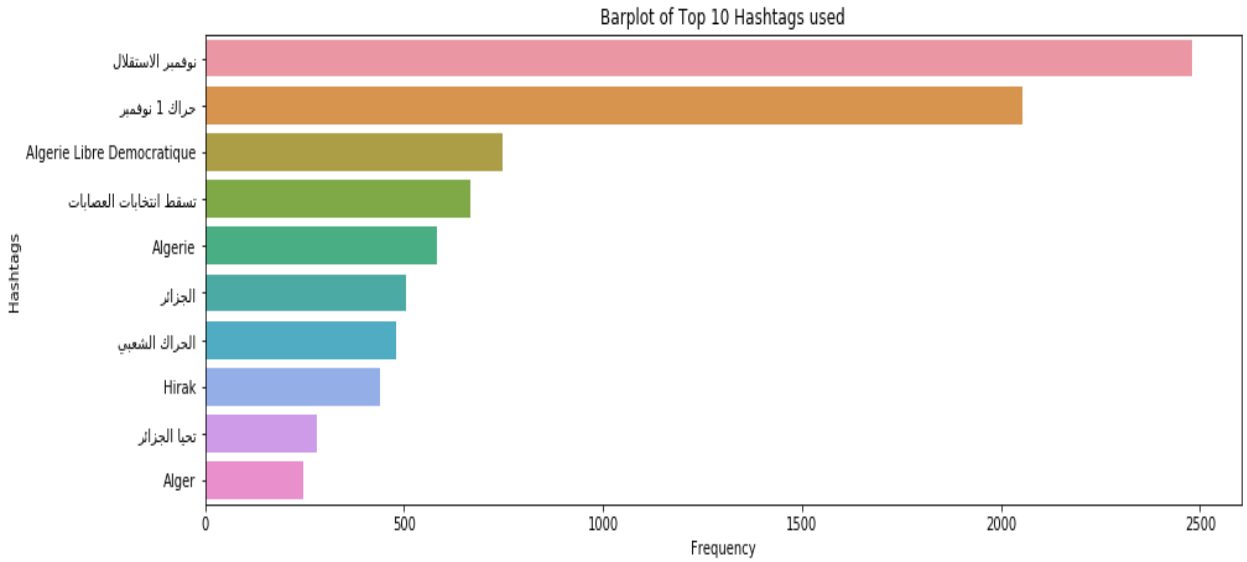

Figure 9 : Top 10 Hashtags avant et durant le 01-11-2019.

2.3. Popularité des tweets

Nous avons également analysé les tweets les plus populaires. Il y a deux indicateurs qui peuvent nous aider à y parvenir : le nombre de retweets et le nombre favori. Malheureusement, nous n'avons pu extraire que le nombre de retweets parce qu'il y a eu quelques difficultés à récupérer le nombre favori. Les figures suivantes illustrent les top 10 tweets pour chaque période.

retweets	Text
1075.0	ترشح بوتفليقة بالمراسلة مشاهدته سماع صوته سابقة تاريخية مثيل تكرر حياة البشرية لذلك السكوت وإلا سنصبح متواطئين الجماعة احتيال الوطن دخل الاحتضار ترشح رجل يقر بمرضه الرسالة المنسوبة
775.0	اجزائر مرفت مرفت رأي شباب السودان
577.0	peuple algérien done leçon courage dignité civisme monde entier
511.0	مليونية عنابة
466.0	jeunes préféré recouvrir image bouteflika afin détruire biens publics
431.0	quand manifestants trolent flics balancent lacrymogènes excelent
403.0	يون مسيرات حاشده بالعاصمة لإنقاذ وطنهم ويصرخون
305.0	القهر العربي هاشتاگات
242.0	بجاية الحضارة والثقافة الاحتجاج تعلموا وسيرو الدرب
229.0	وهران اليوم

Figure 10 : Top 10 tweets du 22-02-2019.

retweets	Text
1066	اليوم عيد بلغكم مرادكم ياشعب البطولات اعلموا رسالتكم يكن صداها داخل أسوار بلادكم ومردودها يثلج صدوركم أمة واحدة والحدود تراب
890	impresionant grandiose historique
862	اجزائريون يقاطعون مراسل الشعب اجزائري والمذيع كالعادة بقطع الإتصال مذيعو أصبحوا فنيو قطع الإتصال كثرة المواقف المشتقة السياسة التحريرية لقناة يتعرضون
833	مشهد حضاري يتقسم آلاف المتظاهرين صفين بشكل تلقائي لفسح الطريق أمام سيارات مكافحة الشغب والشرطة تمر بينهم وهم يردون شعارات جيش شعب خاوة خاوة
717	NaN
682	إستفزاز الثوار وإستخدام الألفاظ الدنيئة المبتذلة ضدهم ضباط أمن نظاميين أقسموا اداء المهنة بشرف أكبر دليل أنهم رهقوا وتخبطوا السلمية وسياسة النفس الطويل محاولة يائسة لتوجيه الثورة خارج حدود السلمية سلمية سلمية الحرامية
654	أرى قائد الجيش اجزائري أحمد قايد صالح أضع يدي قلبي أتمنى الشعب اجزائري واعيا يدار ستار أتمنى يبتعد صالح مواطن الشر ومدابح الثورة المضادة فاجزائر دولة قوية وعتية طيق وصاية سفهاء أبو ظبي
605	أوقفوا خدمة الميترو ليمنعوا تجمع الطلبة فجاء الطلبة الأقدام حدب صوب ليقولوا اجزائرين
602	chair poule casa del mouradia entonée miliers manifestants dont suporteurs usm alger
577	quel peuple tremble régime maudit fin proche

Figure 11 : Top 10 tweets du 08-03-2019.

retweets	Text
889.0	شاهدوا واسمعوا الشارع يهتز إيقاع وكلمات النشيد الوطني فاشهدوا فاشهدوا فاشهدوا
314.0	بلادي فخري بيك اليوم وانا نحكي رجالك صحاو بقاع عندهم بدمهم سقاو ترابك خلاوك ليذا أمانة واجبي اليوم فالسما نحلي علامتك ربي يبعد الحساد ينصرك عنداتك يديم عليذا سما العافية عام نحتفلوا باستقلالك
255.0	قتل الرئيس محمد بوضياف
153.0	chose promise chose due hirak fait floter drapeau somet mont shasta californie altitude somes peuple aspirons somets lent dur épuisant finisons toujours ariver
141.0	mesage jfk john fitzgerald kenedy président étatsunis amérique peuple algérien juilet _corsica
138.0	voilà pase quand laise chiens sans leses policiers haine cœur non désolé merde khouya ailent diable source image chaîne youtube rachid nekaz
137.0	centre balcon
121.0	tsunami humain alger marée humaine pouvoir doit maintenant ceser faire sourde oreile écouter peuple exprime
116.0	maroc algérie tunisie libye emblème brandi chacun pays frères celui divise celui unit contre celui remet cause divise
105.0	grandeur besoin aucun mot décrite aucun superlatif comparée

Figure 12 : Top 10 tweets du 05-07-2019.

retweets	Text
397.0	déclaration er novembre audio chair poule
293.0	alger présent aime telement pays chaire poule
288.0	er novembre début guere algérie gloire martyrs morts patrie dignité libertéoublions jamais histoire somes nouvele génération libérer algérie mafia عقدنا العزم تحيا اجزائر
205.0	rasemblement nocturne devant grande poste veile er novembre
190.0	سيول بشرية تلتحق بالعاصمة صباح امس فغدا جمعة اول يوم عظيم بتظاهر ويحتج المستضعفون الطغاة البغاة نهبوا البلاد وأخضعوها لقوى القهر والنهب العالمية الجمعة ثورة الشعب اجزائري يوم تتكسر قيود يهتز العالم
186.0	المجد الخلود لشهدائنا الابرار
183.0	qasaman force alger veile premier novembre jeunes chantent pays pays tous pays tous tahya djazaïr majd chouhadas
165.0	sofiane lance nouvele tendance sauce er novembre yahatou meftah yatalbou smah tahya djazaïr
142.0	عقدنا العزم تحيا اجزائر فاشهدوا فاشهدوا شهدوا المجد الخلود لشهداء ثورتنا العظيمة ثورة
140.0	atention ancienes videos afrontements policiers manifestants circulent

Figure 13 : Top 10 tweets du 05-07-2019.

Les tweets les plus populaires peuvent aider à révéler les sentiments continus des utilisateurs généraux de Twitter concernant le mouvement. Nous pouvons le comprendre comme suit. Lorsqu'un tweet sur un certain événement ou contenu est retweeté par de nombreuses personnes, cela peut signifier que ces personnes résonnent avec le message et veulent le partager avec autant de personnes que possibles. Les tweets les plus populaires peuvent également fournir plus de détails et de précision sur les principaux sujets / événements d'une journée.

2.4. Fréquence des tweets

La dernière analyse exploratoire que nous avons effectuée s'est concentrée sur le nombre de tweets pendant les périodes que nous avons choisies. À mesure que le temps approche du jour de la manifestation, une augmentation de l'activité de tweet est perceptible, en particulier le jour des manifestations (**figures 14,15,16,17)**. Des études montrent que le volume d'activités en ligne sur Twitter est corrélé aux occurrences réelles de manifestations hors ligne. L'une des raisons importantes de l'augmentation du nombre de tweets réside dans les dernières tentatives des utilisateurs de mobiliser des foules plus importantes pour protester. Les gens sont plus encouragés / disposés à participer à une manifestation de protestation lorsqu'ils sont informés du grand nombre de participants potentiels, surtout si leur famille, leurs amis et leurs voisins font partie de ceux-ci. De plus, les manifestants tweetent beaucoup pendant l'événement pour garder la motivation élevée et pour partager des photos et des vidéos sur les manifestations afin de documenter comment l'événement s'est déroulé dans leur région.

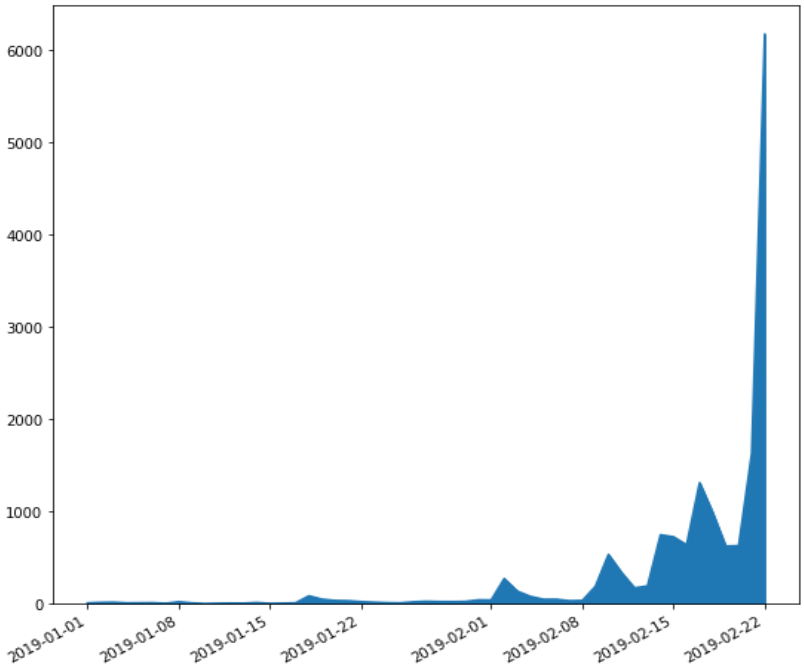

Figure 14 : Fréquence des tweets par jour (avant et durant 22-02-2019).

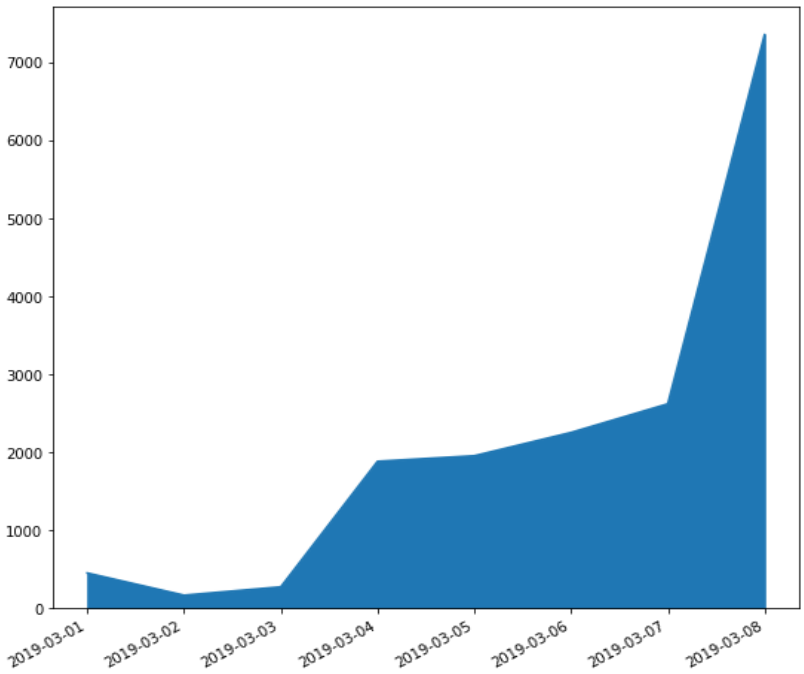

Figure 15: Fréquence des tweets par jour (avant et durant 08-03-2019).

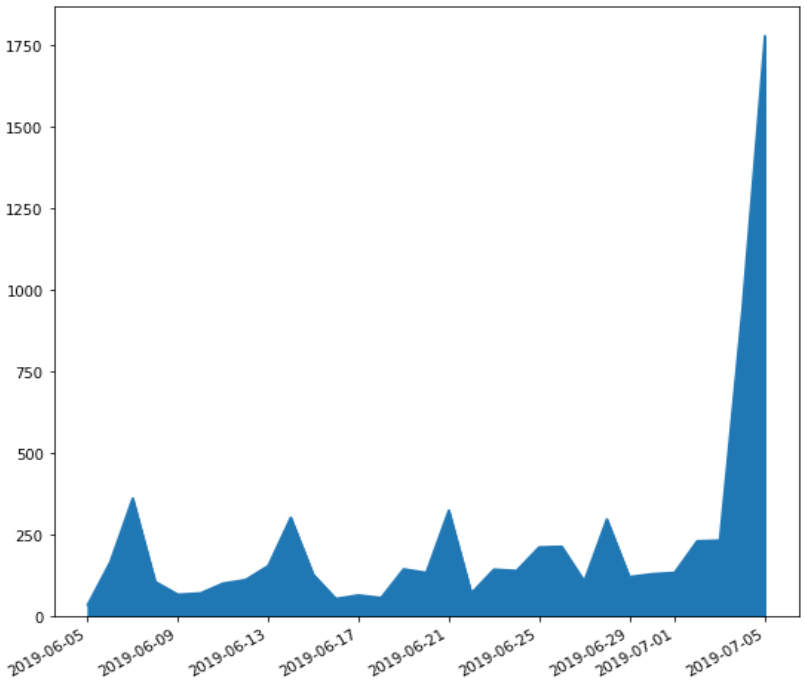

Figure 16: Fréquence des tweets par jour (avant et durant 05-07-2019).

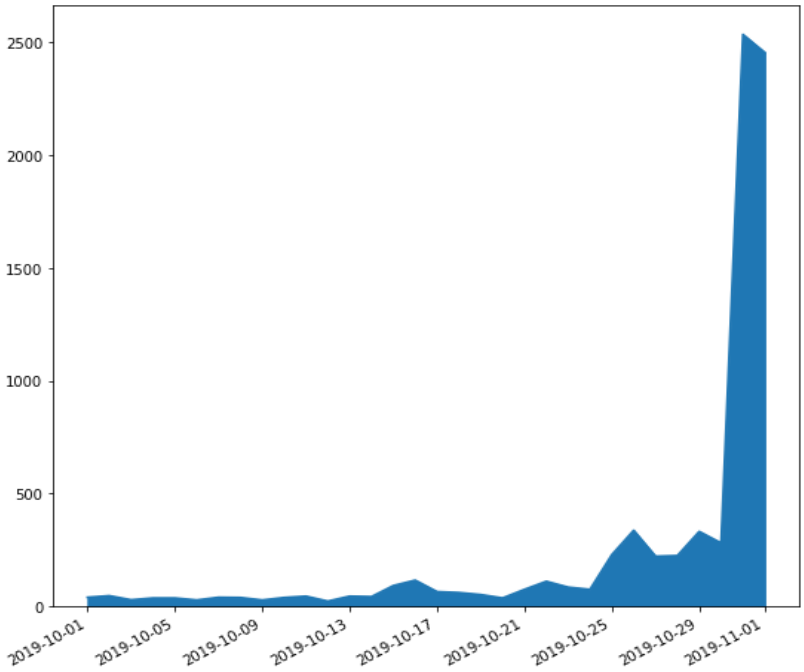

Figure 17: Fréquence des tweets par jour (avant et durant 01-11-2019).

3. Résultats de classification

Tableau 2 : Résultats de classification pour chaque évènement.

L'évènement \ Méthodes de classification		SVM		NB		DT		LR	
		Sac des mots	TF-IDF	Sac des mots	TF-IDF	Sac des mots	TF-IDF	Sac des mots	TF-IDF
Le 22-02-2019	Accuracy	0.695	**0.719**	0.7	**0.713**	0.664	**0.670**	0.715	**0.725**
	Precision	**0.583**	0.533	**0.624**	0.514	0.664	**0.687**	**0.580**	0.524
	Recall	0.647	**0.717**	0.642	**0.713**	0.584	**0.587**	0.685	**0.739**
Le 08-03-2019	Accuracy	0.613	**0.632**	0.639	**0.641**	**0.603**	0.593	0.629	**0.638**
	Precision	**0.429**	0.374	**0.483**	0.405	0.677	**0.685**	**0.422**	0.373
	Recall	0.589	**0.648**	0.620	**0.656**	**0.543**	0.533	0.622	**0.666**
Le 05-07-2019	Accuracy	0.625	**0.648**	0.629	**0.643**	**0.611**	0.608	0.643	**0.656**
	Precision	**0.454**	0.390	**0.456**	0.394	**0.410**	0.401	**0.441**	0.347
	Recall	0.591	**0.663**	0.597	**0.649**	**0.576**	0.572	0.629	**0.720**
Le 01-11-2019	Accuracy	0.626	**0.644**	0.646	**0.650**	0.612	**0.625**	**0.645**	0.644
	Precision	**0.320**	0.253	**0.470**	0.229	**0.375**	0.341	**0.330**	0.212
	Recall	0.530	**0.601**	0.552	**0.640**	0.501	**0.527**	0.576	**0.626**

Le tableau ci-dessus représente le résumé de nos résultats de classification et d'évaluation à l'aide des ensembles de données collectés. Nous avons utilisé 4 méthodes de classification qui sont les suivantes SVM, NB, DT et LR appliquées aux données de notre événements. Pour chaque évènement, nous avons appliqués chaque méthode avec les deux caractéristiques que l'on a choisi : sac des mots et TF-IDF. Les résultats sont à la fois : l'exactitude (accuracy), la précision (precision) et le Rappel (recall).

4. Comparaison des résultats

En premier lieu, nous avons fait une comparaison entre les résultats de classification avec le sac de mots et avec TF-IDF de chaque classificateur, en termes d'exactitude, de précision et de rappel, nous avons fait cette comparaison pour chaque événement, et les résultats sont les suivants :

- **Le 22-02-2019** : pour chaque classificateur nous observons que :
  - L'exactitude avec TF-IDF est meilleure que l'exactitude avec sac des mots.
  - La précision avec sac des mots est meilleure que la précision avec TF-IDF**.** Sauf le classificateur DT.
  - Le rappel avec TF-IDF est meilleure que le rappel avec sac des mots.
  - ➢ Ce comparatif montre que la classification avec TF-IDF est meilleure qu'avec le sac des mots.

- **Le 08-03-2019** : pour chaque classificateur nous observons que :
  - L'exactitude avec TF-IDF est meilleure que l'exactitude avec sac des mots.
  - La précision avec sac des mots est meilleure que la précision avec TF-IDF.
  - Le rappel avec TF-IDF est meilleure que le rappel avec sac des mots. Sauf pour le classificateur DT, c'est le contraire.
  - ➢ Ce comparatif montre que la classification avec TF-IDF est meilleure qu'avec le sac des mots.
- **Le 05-07-2019** : pour chaque classificateur nous observons que :
  - L'exactitude avec TF-IDF est meilleure que l'exactitude avec sac des mots. Sauf pour le classificateur DT
  - La précision avec sac des mots est meilleure que la précision avec TF-IDF.
  - Le rappel avec TF-IDF est meilleure que le rappel avec sac des mots. Sauf pour le classificateur DT
  - ➢ Ce comparatif montre que la classification avec TF-IDF est meilleure qu'avec le sac des mots.

- **Le 01-11-2019** : pour chaque classificateur nous observons que :

  - L'exactitude avec TF-IDF est meilleure que l'exactitude avec sac des mots. Sauf pour le classificateur LR
  - La précision avec sac des mots est meilleure que la précision avec TF-IDF.
  - Le rappel avec TF-IDF est meilleure que le rappel avec sac des mots.

  ➢ Ce comparatif montre que la classification avec TF-IDF est meilleure qu'avec le sac des mots.

En raison des résultats que nous avons obtenus, la classification avec TF-IDF est meilleure que la classification par sac des mots, et c'est à propos de tous les événements. L'étape suivante consiste à comparer les résultats de la classification par TF-IDF pour chaque classification pour obtenir la meilleure classification.

- Pour le 22-02-2019 : nous observons que la classification LR avec TF-IDF est meilleure par rapport les autres classifications.
- Pour le 08-03-2019 : nous observons que la classification LR avec TF-IDF est meilleure par rapport les autres classifications.
- Pour le 05-07-2019 : nous observons que la classification LR avec TF-IDF est meilleure par rapport les autres classifications.
- Pour le 01-11-2019 : nous observons que la classification NB avec TF-IDF est meilleure par rapport les autres classifications.

  ✓ À partir de cette comparaison, nous concluons que la meilleure classification est **la classification régression logistique (RL) avec TF-IDF.**

5. Conclusion

Ici, nous avons formé différents classificateurs et évalué leurs performances. Le résultat est satisfaisant, compte tenu des données et de l'étendue du prétraitement effectué. Les algorithmes peuvent être améliorés en se concentrant sur les étapes de prétraitement. Outre la méthode des mots pondérés comme `CountVectorizer` et `tf-idfVectorizer` utilisée ici, les méthodes d'incorporation de mots comme `Word2Vec` peuvent également être appliquées pour obtenir de meilleurs résultats.

CHAPITRE V

Conclusion générale et implications

La connaissance du traitement du langage associée aux concepts de l'intelligence artificielle et l'apprentissage automatique aident à construire des systèmes intelligents, qui peuvent exploiter les données textuelles et aider à résoudre des problèmes pratiques du monde réel. L'avantage de l'apprentissage automatique est qu'une fois qu'un modèle est formé, nous pouvons directement utiliser ce modèle sur des données nouvelles et inédites pour commencer à voir des informations utiles et les résultats souhaités.

Dans cette étude, nous avons présenté un modèle de prédiction basé sur la classification pour prédire les manifestations de masse sur la base des données Twitter. L'analyse exploratoire des Tweets a révélé des aspects intéressants concernant l'importance d'analyse des données textuelles. En outre, les résultats de notre étude soulignent le rôle clé des médias sociaux, en particulier Twitter, dans les récentes manifestations en tant qu'outil d'organisation et d'information, et que Twitter a le pouvoir de révéler des réponses à de nombreuses questions de recherche.

Cette étude peut aider à la fois le gouvernement et les manifestants. Pour le gouvernement, prévoir et comprendre les tendances présentes sur les réseaux sociaux peut aider les responsables à se préparer à des événements à grande échelle et à s'assurer qu'ils ne deviennent pas violents. Les méthodes de classification peuvent être utilisées pour détecter les appels à l'action sur les réseaux sociaux et, par conséquent, elles peuvent être soit arrêtées, soit censurées. Pour les manifestants, les médias sociaux sont un outil puissant pour engager un plus grand nombre de manifestants pour une cause spécifique. Prédire le comportement et les sentiments des manifestants peut assurer leur sécurité.

Dans les travaux futurs, nous considérons et incluons en outre des caractéristiques spécifiques aux événements (par exemple, la fréquence des tweets, la popularité des hashtags) dans notre modèle pour améliorer les performances de prédiction. On peut aussi essayer de valider les modèles proposés avec différents événements et rechercher des autres caractéristiques spécifiques à un événement peuvent être générées pour divers événements.

Il peut également être possible d'utiliser différents algorithmes d'apprentissage automatique pour améliorer la précision de la prédiction, compte tenu des hashtags et du contenu des tweets ainsi que des attributs spécifiques à l'événement.

Lors de ce travail, nous avons rencontré quelques difficultés, parmi ces difficultés :

- L’extraction des données de Twitter via API est limitée.
- Le traitement des données après leur extraction, du fait de la diversité des langues dans un tweet, et il contient également le dialecte algérien qui est difficile à comprendre pour la machine.
- La suppression manuelle des données non liées aux protestations, du fait que certains utilisateurs publient des hashtags pour juste obtenir des vues.
- Le travail à distance et le mauvais flux d'Internet pendant cette période avec le stress que nous avons vécu avec l'émergence du virus Covid-19.

Références

[1] Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). Analyzing social networks. Sage.2e

[2] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12), 7821-7826.

[3] Dewing, M. (2010). Les médias sociaux Introduction. Bibliothèque du parlement.

[4] Baden, R., Bender, A., Spring, N., Bhattacharjee, B., and Starin, D. (2009, August). Persona: an online social network with user-defined privacy. In Proceedings of the ACM SIGCOMM 2009 conference on Data communication (pp. 135-146).

[5] Edosomwan, S., Prakasan, S. K., Kouame, D., Watson, J., & Seymour, T. (2011). The history of social media and its impact on business. Journal of Applied Management and entrepreneurship, 16(3), 79-91.

[6] Kim, T., Atkin, D. J., & Lin, C. A. (2016). The influence of social networking sites on political behavior: Modeling political involvement via online and offline activity. *Journal of Broadcasting & Electronic Media*, *60*(1), 23-39.

[7] Brunton, S. L., Noack, B. R., & Koumoutsakos, P. (2020). Machine learning for fluid mechanics. Annual Review of Fluid Mechanics, 52, 477-508.

[8] Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".

[9] Guillaume Saint-Cirgue. (2019). Apprendre le Machine Learning en une semaine - machinelearnia.com

[10] Matt Ahlgren, "40+ TWITTER STATISTICS & FACTS FOR 2020"**,** Twitter by the numbers: facts, usage stats and demographics you should know, Last Updated: Apr 16, 2020.

[11] Sheth, A., Kapanipathi, P.: Semantic filtering for social data. IEEE Internet Comput. (2016).

[12] Wang, W., Chen, L., Thirunarayan, K., Sheth, A.P.: Harnessing twitter 'Big Data' for automatic emotion intification. In: IEEE International Conference on Social Computing
(SocialCom) (2012)

[13] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Technical Report (2009)
[14] Agarwal, A., Xie, B., Vovsha, I.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38 (2011)
[15] Liu, K.-L., Li, W.-J., Guo, M.: Emoticon smoothed language models for twitter sentiment analysis. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012).
[16] Kelly, R., Watts, L.: Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. In: Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design (2015)
[17] Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Networks 179, 215–239 (1978)
[18] Bonacich, P.: Power and centrality : a family of measures. Am. J. Sociol. 92(5), 1170–1182 (1987)
[19] Lawyer, G.: Understanding the influence of all nodes in a network. Nat. Sci. Rep. (2015)
[20] Wagner, C., Asur, S., Hailpern, J.: Religious politicians and creative photographers: automatic user categorization in twitter. In: SocialCom (2013)
[21] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler, "Online social networks and offline protest," *EPJ Data Sci.*, vol. 4, p. 19, 2015.
[22] N. Kallus, "Predicting Crowd Behavior with Big Public Data."
[23] K. Clarke and K. Kocak, "Launching Revolution: Social Media and the Egyptian Uprising's First Movers," *Br. J. Polit. Sci.*, pp. 1–21, Dec. 2018.
[24] A. Bruns, T. Highfield, and J. Burgess, "The Arab Spring and Social Media Audiences: English and Arabic Twitter Users and Their Networks," *Am. Behav. Sci.*, vol. 57, no. 7, pp. 871–898, Jul. 2013.
[25] P. Panagiotopoulos, A. Ziaee Bigdeli, and S. Sams, "'5 Days in August' - How London local authorities used twitter during the 2011 riots," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7443 LNCS, pp. 102–113, 2012.
[26] A. Gupta, A. Joshi, and P. Kumaraguru, "Identifying and characterizing user communities on twitter during crisis events," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 23–26, 2012.
[27] M. Cheong, S. Ray, and D. Green, "Interpreting the 2011 London riots from Twitter metadata," *Int. Conf. Intell. Syst. Des. Appl. ISDA*, pp. 915–920, 2012.

[28] M. Tremayne, “Anatomy of Protest in the Digital Era: A Network Analysis of Twitter and Occupy Wall Street,” *Soc. Mov. Stud.*, vol. 13, no. 1, pp. 110–126, 2014.
[29] Y. Theocharis, W. Lowe, J. W. van Deth, and G. García-Albacete, “Using Twitter to mobilize protest action: online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements,” *Inf. Commun. Soc.*, vol. 18, no. 2, pp. 202–220, 2015.
[30] M. D. Conover, E. Ferrara, F. Menczer, and A. Flammini, “The Digital Evolution of Occupy Wall Street,” *PLoS One*, vol. 8, no. 5, 2013.
[31] U. Kursuncu, M. Gaur, U. Lokala, K. Thirunarayan, A. Sheth, and I. B. Arpinar, “Predictive Analysis on Twitter: Techniques and Applications,” Springer, Cham, 2019, pp. 67–104.
[32] Y. Hu, F. Wang, and S. Kambhampati, “Listening to the crowd: Automated analysis of events via aggregated twitter sentiment,” *IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 2640–2646, 2013.
[33] K. Bajpai and A. Jaiswal, “A framework for analyzing collective action events on Twitter,” in *8th International Conference on Information Systems for Crisis Response and Management: From Early-Warning Systems to Preparedness and Training, ISCRAM 2011*, 2011.
[34] T. Sakaki, M. Okazaki, and Y. Matsuo, “Tweet analysis for real-time event detection and earthquake reporting system development,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 919–931, 2013.
[35] C. W. Woo, M. P. Brigham, and M. Gulotta, “Twitter Talk and Twitter Sharing in Times of Crisis: Exploring Rhetorical Motive and Agenda-Setting in the Ray Rice Scandal,” *Commun. Stud.*, vol. 71, no. 1, pp. 40–58, 2020.
[36] S. Brown, “Twitter usage in times of crisis,” *J. Digit. Res. Publ.*, vol. 12, 2011.
[37] T. Sakaki, “Earthquake shakes Twitter users: real-time event detection by social sensors,” in *International World Wide Web Conference Com- mittee (IW3C2)*, 2010, pp. 851–860.
[38] K. Starbird, G. Muzny, and L. Palen, “Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions,” *ISCRAM 2012 Conf. Proc. - 9th Int. Conf. Inf. Syst. Cris. Response Manag.*, vol. 2011, no. April, pp. 1–10, 2012.
[39] M. Li, N. Turki, C. R. Izaguirre, C. DeMahy, B. L. Thibodeaux, and T. Gage, “Twitter as a tool for social movement: An analysis of feminist activism on social media communities,” *J. Community Psychol.*, no. April 2019, pp. 1–15, 2020.
[40] D. Ray and M. Tarafdar, “How does twitter influence a social movement?,” *Proc. 25th Eur. Conf. Inf. Syst. ECIS 2017*, vol. 2017, pp. 3123–3132, 2017.

[41] K. Hunt, “Twitter, social movements, and claiming allies in abortion debates,” *J. Inf. Technol. Polit.*, vol. 16, no. 4, pp. 394–410, 2019.

[42] D. Isa and I. Himelboim, “A Social Networks Approach to Online Social Movement: Social Mediators and Mediated Content in #FreeAJStaff Twitter Network,” *Soc. Media Soc.*, vol. 4, no. 1, Jan. 2018.
[43] J. M. Larson, J. Nagler, J. Ronen, and J. A. Tucker, “Social Networks and Protest Participation: Evidence from 130 Million Twitter Users,” *Am. J. Pol. Sci.*, vol. 63, no. 3, pp. 690–705, Jul. 2019.
[44] C. A. D. L. Salge and E. Karahanna, “Protesting Corruption on Twitter: Is It a Bot or Is It a Person?,” *Acad. Manag. Discov.*, vol. 4, no. 1, pp. 32–49, Mar. 2018.
[45] J. Cadena, G. Korkmaz, C. J. Kuhlman, A. Marathe, N. Ramakrishnan, and A. Vullikanti, “Forecasting social unrest using activity cascades,” *PLoS One*, vol. 10, no. 6, Jun. 2015.
[46] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, “The dynamics of protest recruitment through an online network,” *Sci. Rep.*, vol. 1, no. 1, pp. 1–7, Dec. 2011.
[47] A. Hermida and V. Hernández-Santaolalla, “Twitter and video activism as tools for counter-surveillance: the case of social protests in Spain,” *Inf. Commun. Soc.*, vol. 21, no. 3, pp. 416–433, Mar. 2018.
[48] R. Compton, C. Lee, T. C. Lu, L. De Silva, and M. Macy, “Detecting future social unrest in unprocessed Twitter data: ‘Emerging phenomena and big data,’” in *IEEE ISI 2013 - 2013 IEEE International Conference on Intelligence and Security Informatics: Big Data, Emergent Threats, and Decision-Making in Security Informatics*, 2013, pp. 56–60.
[49] S. Muthiah *et al.*, “Planned Protest Modeling in News and Social Media.”
[50] K. Radinsky and E. Horvitz, “Mining the web to predict future events,” in *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, 2013, pp. 255–264.
[51] R. Korolov *et al.*, “On predicting social unrest using social media,” in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 2016, pp. 89–95.
[52] N. Ramakrishnan *et al.*, “‘Beating the news’ with EMBERS: Forecasting civil unrest using open source indicators,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 1799–1808.
[53] A. Doyle *et al.*, “Forecasting significant societal events using the embers streaming predictive analytics system,” *Big Data*, vol. 2, no. 4, pp. 185–195, Dec. 2014.

[54] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, “Combining heterogeneous data sources for civil unrest forecasting,” in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, 2015, pp. 258–265.

[55] Mellah, S. révolte populaire contre le régime algérien.

[56] Mohammed, M., & Mucchielli, L. (2005). La police dans les «quartiers sensibles»: un profond malaise. *Quand les banlieues brûlent… Retour sur les émeutes de novembre*, 98-119.

[57] Stora, B. (2020). *La gangrène et l'oubli: la mémoire de la guerre d'Algérie*. La Découverte.

[68] Belakhdar, N. (2019). «L’Indépendance, c’est maintenant!» Réflexion sur le soulèvement populaire en Algérie. *L’Année du Maghreb*, (21), 105-116.

[59] Stora, B. (2020). *La gangrène et l'oubli: la mémoire de la guerre d'Algérie*. La Découverte.

[60] mondiale de la Santé, O. (2020). *Eau, assainissement, hygiène et gestion des déchets en rapport avec le virus responsable de la COVID-19: orientations provisoires, 23 avril 2020* (No. WHO/2019-nCoV/IPC_WASH/2020.3). Organisation mondiale de la Santé.

[61] Rawaa kacem le 26-02-2019 : الاحتجاجات ضد العهدة الخامسة للرئيس عبد العزيز بوتفليقة الجزائر بين تحدي الاستقرار والمطالبة بالتغيير Dernier accès le 24/04/2020 https://ar.lemaghreb.tn/item/34939--الاحتجاجات-ضد-العهدة-الخامسة-للرئيس-شؤون-عربية-و-دولية/عبد-العزيز-بوتفليقة-الجزائر-بين-تحدي-الاستقرار-والمطالبة-بالتغيير

[62] B. Mohammed le 23/02/2019 “ .... تجمعات و مسيرات ضد العهدة الـ5 المعارضة تؤكد ضعفها في الشارع " **elhayatonline** Dernier accès le 24/04/2020 https://www.elhayatonline.com/العدد-1627-ليوم-23-فيفري-2019/

[63] Zohra Bensemra le 9 mars 2019 " Des manifestations « historiques » en Algérie, selon la presse" **le monde** Dernier accès le 27/04/2020 www.lemonde.fr/afrique/article/2019/03/09/des-manifestations-historiques-en-algerie-selon-la-presse_5433802_3212.html

[64] Nadia Ben Taher le 23/03/2019 " مسيرة الجزائر لـ”22 مارس” بعيون الصحافة العربية والعالمية " **ennaharonline.com** Dernier accès le 27/04/2020 https://www.ennaharonline.com/مسيرة-الجزائر-لـ22-مارس-بعيون-الصحافة-ا/

[65] Nabia Lahchi Publié le 3 mai 2019 à 18 :10, mis à jour le 3 mai 2019 à 19 :31 " Bordj Bou Arreridj, l’autre capitale de la contestation en Algérie " **lefigaro.fr** Dernier accès le 27/04/2020 https://www.lefigaro.fr/international/bordj-bou-arreridj-l-autre-capitale-de-la-contestation-en-algerie-20190503

[66] https://pypi.org/project/GetOldTweets3/

[67] https://github.com/Jefferson-Henrique/GetOldTweets-python

[68] Samir Bendjafer le 30/09/2019 جزائريون يطلبون من جوستان ترودو دعم الحراك الشعبي في راديو كندا الدولي الجزائر RCI Dernier accès le 06/06/2020 [https://www.rcinet.ca/ar/2019/09/30/جزائريون-يطلبون-من-جوستان-ترودو-دعم-ال/]

[69] مسيرة للجالية الجزائرية في بريطانيا تطالب بتغيير كامل...Nov 9, 2019 - Uploaded by Alghad TV - قناة الغد [https://www.youtube.com/watch?v=YAboVyTy_A]

[70] France24 Le 17/03/2019 فيديو: الجزائريون يتظاهرون في باريس تضامنا مع الحراك المناهض لبوتفليقة https://www.france24.com/ar/20190317--الجزائر-مظاهرات-باريس-تضامن-الحراك-الشعب-بوتفليقة

[71] Lebovich, A. (2015). *Deciphering Algeria: the Stirrings of Reform?*. European Council on Foreign Relations (ECFR).

[72] ROUDABEH KISHI (4 June 2020). CDT SPOTLIGHT: ALGERIA & THE HIRAK MOVEMENT.

[73] ACLED 2019 Armed Conflict Location & Event Data Project (ACLED).FROM THE STREETS TO THE ELITES: DEMONSTRATIONS IN A POST-BOUTEFLIKA ALGERIA.

[74] HILARY MATFESS (13 June 2019). STEADY AS SHE GOES: ALGERIA'S "SMILE REVOLUTION" CONTINUES.

[75] HILLARY TANOFF (8 March 2019). "THE POUVOIR" AND THE PEOPLE: DEMONSTRATIONS IN ALGERIA

[76] Samuel, J., Ali, G. G., Rahman, M., Esawi, E., & Samuel, Y. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, *11*(6), 314.

[77] Rane, A., & Kumar, A. (2018, July). Sentiment classification system of Twitter data for US airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 1, pp. 769-773). IEEE.

[78] Ali, A. R., & Ijaz, M. (2009, December). Urdu text classification. In *Proceedings of the 7th international conference on frontiers of information technology* (pp. 1-7).

[79] Rathi, M., Malik, A., Varshney, D., Sharma, R., & Mendiratta, S. (2018, August). Sentiment analysis of tweets using machine learning approach. In *2018 Eleventh international conference on contemporary computing (IC3)* (pp. 1-3). IEEE.