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

MEMOIRE

Présente en vue de l'observation du diplôme
Master en informatique
Spécialité : Technologie de l'information et de la communication

THEME

Prédiction des manifestations publiques à l'aide de réseaux de neurones artificiels

Présenté par :
BELAID Boualem

Soutenu publiquement le :
Devant le jury composé :

Président Mr :
Examinateur Mr :
Examinateur Mr :
Encadreur Dr : LAIFA Meriem

Promotion : 2020/2021

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 et leur soutien moral ,

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.

”

- Boualem

Remerciements

Tout d'abord, je remercie Allah le tout puissant de m'avoir donné le courage et la patience nécessaires à mener ce travail à son terme.

Je tiens à remercier tout particulièrement mon encadrante **Dr. LAIFA Meriem**, pour son encadrement, pour sa patience et son encouragement. Son œil critique m'a été très précieux pour structurer le travail et pour améliorer la qualité des différentes sections.

Que les membres de jury trouvent, ici, l'expression de mes sincères remerciements pour l'honneur qu'ils me font en prenant le temps de lire et d'évaluer ce travail.

Je souhaite aussi remercier l'équipe pédagogique et administrative de département de l'informatique pour leurs efforts dans le but de nous offrir une excellente formation.

Pour finir, je souhaite remercier toute personne ayant contribué de prés ou de loin à la réalisation de ce travail.

Résumé

Le principal objectif de ce mémoire est de prédire les manifestations publiques au moyen d'algorithmes d'apprentissage automatique et l'apprentissage profond utilisant les fonctionnalités extraites des données des médias sociaux à partir de Twitter.En particulier, nous considérons le cas de « Hirak » qui a commencé en février 2019 en Algérie.
les Résultas sont réalisé grâce aux techniques de l'apprentissage automatique et l'apprentissage profond tout en respectant la méthodologie de la classification du texte liée au domaine du traitement automatique du langage naturel.

Mots clés : Apprentissage automatique,Apprentissage profond, HIRAK, Classification, Traitement du langage naturel.

Abstract

The main objective of this dissertation is to predict public protests by means of machine learning algorithms and deep learning using features extracted from social media data from Twitter.In particular, we consider the case of "Hirak" which started in February 2019 in Algeria.The The results are achieved through machine learning and deep learning techniques while respecting the methodology of text classification related to the field of automatic natural language processing..

Keywords : Machine learning, deep learning, Classification, HIRAK, Natural language proccessing.

ملخص

الهدف الرئيسي من هذه الرسالة هو التنبؤ بالاحتجاجات العامة عن طريق خوارزميات التعلم الآلي والتعلم العميق باستخدام الميزات المستخرجة من بيانات وسائل التواصل الاجتماعي من تويتر ، وعلى وجه الخصوص ، نأخذ في الاعتبار حالة "الحراك التي بدأت في فبراير 2019 في الجزائر. يتم تحقيق النتائج باستخدام تقنيات التعلم الآلي والتعلم العميق مع احترام منهجية تصنيف النص المتعلقة بمجال معالجة اللغة الطبيعية.

كلمات مفتاحية : التعلم الآلي ، التعلم العميق ، الحراك ، التصنيف ، معالجة اللغة الطبيعية.

Table des matières

Table des figures

Liste des tableaux

Liste des sigles et acronymes

NLP	*Natural Language Processing*
AS	*Analyse des Sentiments*
DL	*Deep Learning*
IE	*Extraction d'informations*
SVM	*les machines à vecteurs de support*
MNB	*Multinomial Naive Bayes*
KNN	*K-nearest neighbors*
RBF	*Radial Basic Function*
PMC	*Perceptron mono couche*
MLP	*Multi layer perceptron*
RNA	*Réseau de neurones artificiels*

Chapitre 1

Introduction générale

Contexte

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 [1]. 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.

Les réseaux sociaux sont l'une des plus grandes réussites de l'internet, puisque des sites comme Facebook et Twitter sont passés de zéro utilisateur à plus d'un milliard d'utilisateurs en moins de dix ans. Pourtant, le phénomène des médias sociaux ne fait toujours pas l'objet d'un corpus théorique cohérent[2].

Les réseaux sociaux ont joué un rôle central dans la mobilisation des citoyens algériens pour protester pacifiquement contre le régime corrompu de leur pays[3].

L'apprentissage automatique est une branche évolutive des algorithmes informatiques qui vise à imiter l'intelligence humaine en apprenant de l'environnement environnant. La technologie basée sur l'apprentissage automatique a été appliquée avec succès dans divers domaines.Les universitaires et les praticiens ont mené des recherches approfondies sur les modèles de prévision de faillite et de défaut pour la gestion du risque de crédit. La recherche universitaire révolutionnaire utilise des techniques statistiques tra-

ditionnelles (telles que l'analyse discriminante et la régression logistique) et les premiers modèles d'intelligence artificielle (tels que les réseaux de neurones artificiels) pour évaluer la faillite [4].

L'apprentissage profond est un domaine émergent de la recherche sur l'apprentissage machine (AM). Il comprend plusieurs couches cachées de réseaux neuronaux artificiels. La méthodologie d'apprentissage profond applique des transformations non linéaires et des abstractions de modèles de haut niveau dans de grandes bases de données. Les progrès récents des architectures d'apprentissage profond dans de nombreux domaines ont déjà apporté des contributions significatives à l'intelligence artificielle [5].

Problématique

Les questions liées à la participation de plus en plus complexe des agents humains et technologiques dans les mouvements sociaux méritent une attention particulière. Nous avons observé des changements dans la manière dont les différents acteurs de la protestation et de la résistance. L'objectif principal de ce projet est de prédire les protestations publiques à l'aide d'algorithmes de réseaux neuronaux. au moyen d'algorithmes de réseaux neuronaux utilisant des caractéristiques extraites des médias sociaux. En particulier, nous considérons le cas du " Hirak " qui a débuté en février 2019 en Algérie. L'objectif est d'utiliser un modèle de prédiction basé sur la classification pour prédire les protestations de masse en fonction du contenu des médias sociaux publics. provenant de Twitter. Les résultats et l'efficacité des algorithmes utilisés doivent être comparés à d'autres méthodes de classification basiques d'apprentissage machine (ML). En outre, les recherches connexes en cours et les travaux futurs qui nécessitent un examen plus approfondi devraient également être soulignés.

Objectifs

L'objectif primordial de ce mémoire est de prédire les manifestations publiques au moyen d'algorithmes d'apprentissage approfondi basent sur les réseaux de neurones 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 des méthodes de classification essentielles de l'apprentissage approfondie , et aussi de comparé les résultats par les réseaux de neurones avec les résultats déjà acquis avec les algorithmes de l'apprentissage automatique.

Organisation du mémoire

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), Le quatrième chapitre est consacré aux expérimentations et résultats. Et on fait une vision générale sur les méthodes de classification.

Chapitre 2

État de l'art

2.1 Introduction

Dans ce chapitre, nous décrirons différents concepts tels que le traitement du langage naturel, l'apprentissage automatique, classification des textes et la prédiction des manifestations.

2.2 Traitement du langage naturel

Les dernières avancées en matière d'apprentissage profond (DL) ont fait des percées dans de nombreux domaines, tels que la vision par ordinateur, le traitement du langage naturel (NLP) et le traitement de la parole. La recherche a montré que de nombreuses méthodes basées sur l'apprentissage profond peuvent produire des résultats de pointe dans diverses tâches, qui sont très importantes pour les réseaux sociaux en ligne et l'informatique sociale, telles que l'analyse des sentiments (AS) et la classification des sujets [6]. Les tâches de NLP sont devenues très importantes dans les réseaux sociaux en ligne.

2.2.1 Définition

Le traitement du langage naturel (NLP) est un domaine de recherche et d'applications qui explore comment utiliser les ordinateurs pour comprendre et manipuler du texte ou de la parole en langage naturel pour effectuer des tâches utiles[7]. Les chercheurs en NLP visent à recueillir des connaissances sur la façon dont les humains comprennent et utilisent le langage afin de développer des outils et des techniques appropriés pour que les systèmes informatiques comprennent et manipulent le langage naturel pour effectuer les tâches requises.

2.2.2 L'importance de l'NLP

Le traitement du langage naturel est un moyen pour les machines et les ordinateurs d'analyser, de comprendre et de déduire le sens du langage parlé humain d'une manière utile et intelligente[8]. À l'aide de techniques et de méthodes de traitement du langage naturel, les scientifiques des données peuvent facilement organiser et structurer les

connaissances acquises pour effectuer diverses tâches telles que la traduction, la synthèse automatique de texte, la reconnaissance d'entités nommées, l'analyse des sentiments, la reconnaissance vocale, l'extraction de relations et la segmentation de sujets.

2.2.3 Applications

Le traitement du langage naturel fournit la théorie et la mise en œuvre pour une série d'applications. En fait, toute application utilisant du texte est candidate à NLP. Les applications les plus courantes qui utilisent NLP sont les suivantes :

- **Recherche d'informations** [9] : compte tenu de la grande quantité de texte dans cette application, il est surprenant que peu d'implémentations utilisent le NLP. Récemment, les méthodes statistiques utilisées pour compléter le NLP ont été davantage utilisées. Les recherches de Liddy et Strzalkowski ont développé un système important basé sur le NLP.

- **Extraction d'informations (IE)** [9] : un domaine d'application plus récent, IE s'engage à identifier, marquer et extraire certaines informations clés, telles que des personnes, des entreprises, des lieux et des organisations, à partir d'un grand nombre de collections de textes sous la forme de représentations structurées. Ces extraits peuvent ensuite être utilisés dans diverses applications, y compris la réponse à des questions, la visualisation et la navigation dans les données.

- **Répondre aux questions** [9] : contrairement à la recherche d'informations, la recherche d'informations fournit une liste de documents potentiellement pertinents en fonction de la requête de l'utilisateur,la réponse à une question fournit à l'utilisateur soit le texte de la réponse elle-même, soit des éléments de réponse.

- **Résumé** [9] : un niveau plus élevé de NLP, en particulier le niveau de la voix, peut permettre la réalisation d'une forme narrative courte mais richement structurée du document original qui est pliée en un texte plus long.

- **Traduction automatique** [9] : peut-être la plus ancienne de toutes les applications NLP. Le niveau NLP a été utilisé dans les systèmes de traduction automatique,

allant des méthodes « basées sur les mots » aux applications impliquant une analyse de plus haut niveau.

- **Les systèmes de dialogue** [9] : intégrés aux systèmes envisagés par les grands fournisseurs d'applications d'utilisateurs finaux pourraient devenir une application courante à l'avenir. Les systèmes de dialogue se concentrent généralement sur des applications étroitement définies (par exemple, les réfrigérateurs ou les systèmes audio à domicile) et utilisent actuellement des langages de niveau de parole et de vocabulaire.

- **L'analyse des sentiments** [10] : est l'interprétation et la classification des émotions (positives, négatives ou neutres). À partir des critiques de produits ou des publications sur les réseaux sociaux, la tâche consiste à déterminer si le sentiment est positif, neutre ou négatif.

- **La détection de spam** [11] : est utilisée pour détecter les e-mails indésirables arrivant dans la boîte de réception de l'utilisateur.

2.3 L'apprentissage automatique

Les algorithmes d'apprentissage automatique utilisés aujourd'hui sont axés sur les performances et basés sur des attributs connus appris à partir d'échantillons d'apprentissage, en se concentrant sur la précision de la classification et/ou prédiction.

2.3.1 Types de l'apprentissage automatique

L'apprentissage supervisé est généralement une tâche d'apprentissage automatique pour apprendre une fonction qui met en correspondance une entrée et une sortie sur la base d'échantillons de paires d'entrée-sortie. Il utilise des données d'apprentissage étiquetées et une collection d'exemples d'entraînement pour déduire des fonctionnalités[12]. Lorsqu'il est déterminé que certains objectifs doivent être atteints à partir d'un ensemble spécifique d'entrées, un apprentissage supervisé, tel qu'une approche basée sur les tâches, est réalisé.

Les tâches supervisées les plus courantes sont la « classification » et la « régression », la « classification » est utilisée pour séparer les données et la « régression » est utilisée pour ajuster les données.

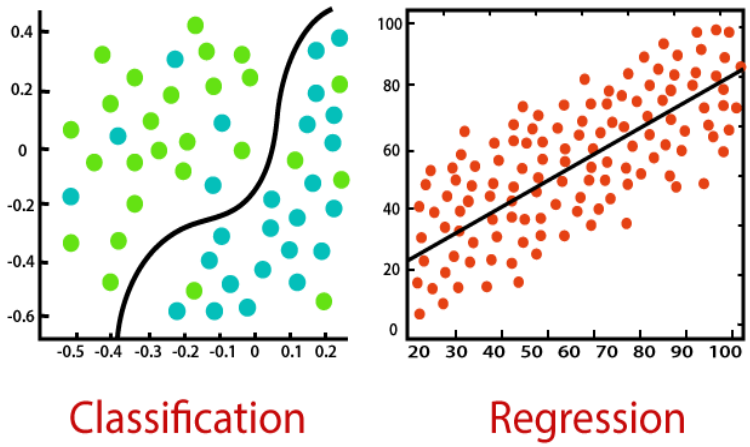

FIG. 2.1 : La Différence entre les tâches supervisées

L'apprentissage non supervisé peut analyser des ensembles de données non étiquetés sans intervention humaine, il s'agit d'un processus géré par les données. Il est largement utilisé pour extraire des caractéristiques génératives, identifier les tendances et structures importantes, regrouper les résultats et à des fins exploratoires [12]. Les tâches d'apprentissage non supervisé les plus courantes sont le regroupement, l'estimation de la densité, l'apprentissage des caractéristiques, la réduction de la dimensionnalité, la recherche de règles d'association et la détection d'anomalies.

2.4 La classification des textes

2.4.1 Définition

La classification de texte implique l'attribution d'un ensemble de catégories prédéfinies pour ouvrir le texte. Les classificateurs de texte peuvent être utilisés pour organiser, structurer et classer presque tous les types de texte, qu'il s'agisse de documents, de dossiers de recherche et médicaux ou de réseaux entiers.

La classification manuelle du texte nécessite un annotateur humain, qui peut interpréter le contenu du texte et le classer en conséquence. Cette méthode peut produire

de bons résultats, mais elle est longue et coûteuse.

La classification automatique du texte utilise l'apprentissage automatique, le traitement du langage naturel (NLP) et d'autres technologies basées sur l'IA pour classer automatiquement le texte plus rapidement, plus efficacement et avec plus de précision.

2.4.2 Méthodes

Il existe de nombreuses méthodes de classification automatique de texte, mais elles appartiennent toutes à trois types de systèmes :

- **Systèmes à base de règles** [13] : utilise un ensemble de règles linguistiques faites à la main pour diviser le texte en groupes organisés. Ces règles indiquent au système d'utiliser les éléments sémantiquement liés du texte pour identifier les catégories associées en fonction de son contenu. Chaque règle se compose d'un antécédent ou d'un modèle et d'une catégorie de prédiction.
  Ces systèmes nécessitent une connaissance approfondie du domaine, vu la génération des règles pour un système complexe peut être assez difficile à maintenir et ne sont pas très évolutifs.

- **Systèmes basés sur l'apprentissage automatique** [14] :la première étape de la formation d'un classificateur NLP d'apprentissage automatique est l'extraction de données : une méthode appliquée pour transformer chaque texte en une représentation numérique sous la forme d'un vecteur.
  L'algorithme d'apprentissage automatique est ensuite fourni avec des données d'entraînement, qui se composent d'une paire d'ensembles de fonctionnalités (vecteurs de chaque exemple de texte) et d'étiquettes (par exemple, sports, politique) pour produire des classifications de modèles.

- **Les systèmes hybrides** [15] combinent des classificateurs de base fondés sur l'apprentissage automatique avec des systèmes basés sur des règles pour améliorer les résultats. Ces systèmes hybrides peuvent être facilement perfectionnés en ajoutant des règles spécifiques pour les étiquettes de conflit qui ne sont pas correctement modélisées par le classificateur de base.

2.4.3 Algorithmes

Parmi les algorithmes de classification de texte les plus connus figurent la famille d'algorithmes Naive Bayes, les machines à vecteurs de support (SVM), l'arbre de décision et KNN.

Naive bayes [16] La famille des algorithmes statistiques Naive Bayes est l'un des algorithmes les plus largement utilisés dans la classification et l'analyse de texte en général. L'un des membres de cette famille, Multinomial Naive Bayes (MNB) a un énorme avantage, même si votre ensemble de données n'est pas très grand (environ des milliers d'échantillons étiquetés) et que les ressources informatiques sont rares, vous pouvez obtenir un bon résultat.
Naive Bayes est basé sur le théorème de Bayes, qui peut nous aider à calculer la probabilité conditionnelle de deux événements en fonction de la probabilité de chaque événement. Par conséquent, nous calculons la probabilité de chaque étiquette pour un texte donné, puis nous affichons l'étiquette avec la probabilité la plus élevée.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{2.1}$$

La machine vectorielle de support (SVM) [17] est un autre algorithme d'apprentissage automatique puissant utilisé pour classer le texte, car, comme les Bayes naïfs, les SVM n'ont pas besoin de beaucoup de données à traiter pour commencer à fournir des résultats précis. Le SVM nécessite toutefois plus de ressources informatiques que le Bayes naïf, mais les résultats sont encore plus rapides et plus fiables.

L'arbre de décision : L'une des techniques largement utilisées dans l'exploration de données est Créez un système de classificateurs. Dans l'exploration de données, les algorithmes de classification peuvent gérer de grandes quantités d'informations. Ils peuvent être utilisés pour faire des hypothèses concernant des noms de classe catégoriques, pour classer les connaissances en fonction de l'ensemble d'apprentissage et de l'étiquette de catégorie et pour regrouper les données nouvellement obtenues [18].

KNN est l'un des plus rapides, des plus simples et des plus faciles à conceptualiser parmi tous les algorithmes d'apprentissage automatique. La prédiction de la catégorie

d'échantillon de test est basée sur les k échantillons d'apprentissage les plus proches de l'échantillon de test, où k est un entier positif, généralement très petit [19]. Ensuite, nous attribuons les catégories d'échantillons de test en fonction de la catégorie avec la probabilité de catégorie la plus élevée. La distance euclidienne est la métrique la plus couramment utilisée pour les variables continues. Le meilleur choix de k dépend des données ou peut être sélectionné par une technique heuristique.

2.5 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 [20]-[23], aux émeutes de Londres [24]-[26] et Occupy Wall Street [27]-[29]. Dans la littérature, il y a beaucoup d'analyses prédictives sur Twitter [30], certaines de ces études effectuent des analyses de contenu et de sentiment sur Twitter pendant les événements et les manifestations [27], [30], [31]. D'autres études se penchent sur l'utilisation de Twitter lors de crises et catastrophe naturelle [32]-[35] ou décrivent, modélisent et interprètent les réseaux d'utilisateurs et les relations entre les réseaux sociaux ainsi que les mouvements sociaux [36]-[40]. 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 [41], [42], les interactions sur les médias sociaux, et les cascades d'activités [43], [44], 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 [45] 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.[46] 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. [47] 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 [48] 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 [21] a utilisé les données de plus de 300 000 sites Web différents rassemblés par 'Recorded Future' 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.[20] 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.[49] ont étudié les manifestations de 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.[50] et Doyle et al. [51] 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. [52] 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.Taylor et Veugelers.[53], [54] Deux études plus récentes ont documenté la manière dont les médias sociaux peuvent continuer à jouer certains rôles dans le maintien des mouvement en suspens. De nombreux chercheurs ont proposé leur cadre pour traiter l'agitation civile par des techniques telles que l'apprentissage automatique semi-supervisé, le réseau neuronal, l'arbre de décision, Naive Bayes.Le chercheur Nasser Alsaedi et.al ont proposé une approche consistant à utiliser des caractéristiques temporelles, spatiales et textuelles pour détecter des événements à petite échelle dans un lieu et un temps précis que les algorithmes existants[55].La chercheuse Juhi P Pathak a discuté du rôle des médias sociaux dans la violence ethnique [56].L'objectif était d'analyser de manière critique le rôle des médias sociaux dans la violence ethnique du Nord-Est et de suggérer des moyens de mettre fin à l'incitation à cette violence par le biais des plateformes sociales [56].

La majorité des recherches portent sur la détection d'événements dans des modèles hors ligne. La détection d'événements en temps réel sont très rares. Une autre constatation est l'absence d'un modèle unifié pour détecter tout type d'événement. Le domaine de l'analyse d'événements est toujours en progression et a le potentiel d'obtenir des micro-informations liées aux divers événements qui se produisent, ce qui conduit à la prédiction d'événements futurs [57].

Depuis l'ère de l'apprentissage profond, les réseaux de neurones récurrents (RNN) et leurs variantes sont progressivement devenus la principale méthode de composition des modèles de classification. Le réseau LSTM profond proposé par Shi et al. améliore la méthode d'apprentissage des caractéristiques des phrases et la précision de la classification [58]. Lin R. combine le RNN et CNN pour proposer le réseau RCNN [59] qui améliore l'apprentissage du modèle en résolvant le problème de la lenteur et de la non-convergence, puis le mécanisme d'attention a été proposé. Zhou et al. combinant le réseau LSTM avec le mécanisme d'attention [60] pour résoudre le problème de la classification interlinguistique.

Après l'émergence de BERT, l'enregistrement de la tâche de classification d'étiquettes a été rafraîchi à nouveau.

Chapitre 3

Conception et modélisation

3.1 Introduction

Afin d'atteindre notre objectif de prédire les manifestations liées au « Hirak » à partir des tweets, nous avons suivi les 05 principales étapes présentées dans l'organigramme suivant (Fig 3.1). Dans un premier temps, nous extrairons les données pertinentes de Twitter et préparerons les prochaines étapes. Étant donné que les données extraites sont des données 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 de caractéristiques basées sur des ensembles de données. Une fois les caractéristiques sélectionnées et extraites, nous sélectionnons, formons et évaluons des classificateurs pour faire des prédictions. Dans les sections suivantes, nous décrivons en détail l'utilisation de chaque étape.

Fig. 3.1 : Les étapes principales pour prédire les protestations liées au "Hirak" [61].

3.2 Explication détaillée du processus

3.2.1 Collecte et nettoyage des données

on a utilisé les données des etudiantes de l'année passée en ce basant sur les évènements majeurs suivants :

- **22 Février 2019** : date du début du Hirak.
- **08 Mars 2019** : la journée internationale de la femme.
- **05 Juillet 2019** : jour de l'indépendance algérienne.

- **1er Novembre 2019** : la guerre de la révolution algérienne.

les données extraites des tweets sont d'une durée d'un mois jusqu'à ce que l'évènement survient (y compris le jour de l'évènement).

3.2.2 Extraction des données

On a utilisé la bibliothèque GetOldTweets3 [62] pour extraire des tweets de twitter.com. Initialement, l'API de recherche Twitter a été utilisée. Cependant, l'API Twitter officielle impose certaines restrictions, telles que la limite d'extraction de tweets aux 7 derniers jours. Les données extraites contiennent les informations suivantes : text du tweet, nom de l'utilisateur, nombre de retweets et la date et l'heure.

3.2.3 Prétraitement

Le prétraitement est une étape primordiale pour toute tâche NLP, il vise à transformer les données brutes en un format approprié. Il comprend de nombreuses étapes. Dans ce qui suit, nous allons expliquer chaque étape séparément.

1. **Nettoyage des données** Le nettoyage des données est la première étape pour la transformation des données.

2. **Suppression des tweets répétés** Puisque nous utilisons des hashtags et des mots-clés séparés pour extraire les données, Certains tweets peuvent apparaître plus d'une fois dans notre ensemble de données combiné. Par conséquent, les tweets en double doivent être supprimés.

3. **Suppression des émoticônes et des mentions** Comme nous n'avons pas besoin d'emojis et de mentions dans notre champ d'exploration, nous les avons supprimés de notre texte de tweet en remplaçant tous les symboles par replace().

4. **Suppression de la ponctuation et les nombres** Cette dernière consiste à supprimer la ponctuation et les chiffres, car ces derniers n'ajouteront pas d'informations supplémentaires lors du traitement des données de texte. Par conséquent, les supprimer nous aidera à réduire la taille des données d'apprentissage.

5. **Suppression des mots d'arrêt** Les mots d'arrêt doivent être supprimés des données textuelles. La suppression des mots d'arrêt supprimera les mots courants et fréquents qui n'ont pas d'impact significatif sur la phrase.

3.2.4 Extraction des Caractéristiques

La tokenisation fait référence à la division du texte en unités minimales significatives. Il y a un tokenizer de phrase et un tokenizer de mot. Nous avons utilisé un tokenizer 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 tokenisation comme NLTK, SpaCy et TextBlob. on a basé sur deux méthodes :

1. **Sac de mots** : La tokenisation consiste à décomposer le texte en une série de mots ou de phrases. Dans notre exemple, nous utilisons la fonction split pour convertir 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 du sac de mots a un inconvénient. En supposant qu'un mot spécifique apparaisse dans tous les documents du corpus, il deviendra important dans notre méthode précédente. Ce n'est pas bon pour notre analyse.

2. **TF-IDF** : est l'une des méthodes les plus efficaces pour calculer les pondérations des termes. L'idée d'avoir TF-IDF est de considérer l'importance d'un mot pour les documents d'une collection, de manière à standardiser les mots apparaissant fréquemment dans tous les documents. La fréquence du terme, notée TF, est calculée dans le modèle sac de mots dans la section précédente. La fréquence d'un terme dans n'importe quel vecteur de document est représentée par la valeur de fréquence d'origine du terme dans un document particulier.
   La fréquence de document inverse donnée par IDF est l'inverse de la fréquence de document de chaque mot. La méthode de calcul consiste à diviser le nombre total de documents dans notre corpus par la fréquence de document de chaque mot, puis à appliquer une échelle logarithmique mise à jour au résultat . Nous avons utilisé la méthode TfidfVectorizer.

3.3 Classification

3.3.1 Définition des réseaux de neurones artificielles

Les réseaux de neurones artificiels [63] sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Toute structure hiérarchique de réseaux est évidemment un réseau.
Les réseaux de neurones artificiels, ou réseaux neuromimétiques, sont des modèles inspirés du fonctionnement du cerveau animal, et dont le but est de voir surgir des propriétés analogues au système biologique.

Ils en reprennent quelques grands principes :

- **Le parallélisme** : les neurones sont des entités réalisant une fonction très simple, mais ils sont très fortement interconnectés entre eux, ce qui rend le traitement du signal massivement parallèle.
- **Les poids synaptiques** : les connexions entre les neurones ont des poids variables, qui déterminent la force de l'interaction entre chaque paire de neurones.
- **L'apprentissage** : ces coefficients synaptiques sont modifiables lors de l'apprentissage, dans le but de faire réaliser au réseau la fonction désirée.

3.3.2 Structure du réseau de neurones

Les réseaux de neurones artificiels sont composés d'unités de calcul élémentaires appelées neurones combinées selon différentes architectures. Par exemple, ils peuvent être disposés en couches (réseau multicouche), ou avoir une topologie de connexion. Les réseaux en couches sont constitués de trois couches comme le montre (Fig :3.2)

- **Couche d'entrée** : composée de n neurones (un pour chaque entrée du réseau).
- **Couche cachée** : composée d'une ou plusieurs couches cachées (ou intermédiaires) constituées de m neurones.
- **Couche de sortie**, constituée de p neurones (un pour chaque sortie du réseau).

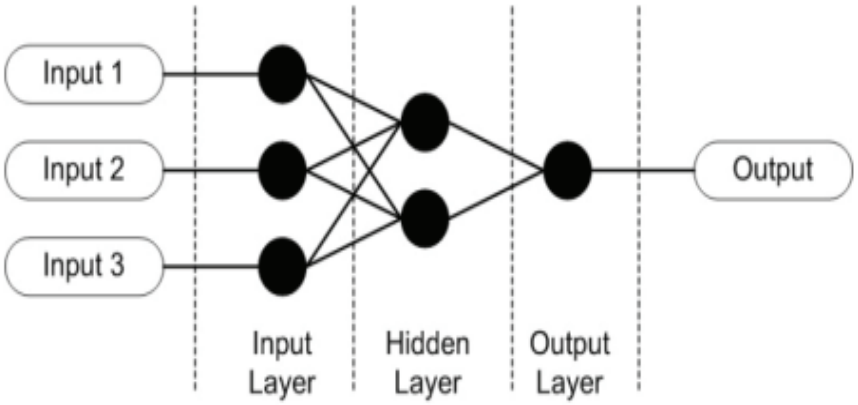

Fig. 3.2 : Réseau de neurones artificiels [64].

3.3.3 Types des réseaux de neurones

Les connexions entre les neurones qui composent le réseau décrivent la topologie du modèle. Elle peut être quelconque, mais le plus souvent il est possible de distinguer une certaine régularité (réseau à connexion complète) [65]. Les différent type de RNA sont distinguées par le type d'interconnexion (topologie du réseau), le choix des fonctions de transferts (types de neurone) et au Mode (règle) d'apprentissage associé aux réseaux comment estimer les poids.

Nous distinguons 4 modèles des RNA :

Recurrent Neural Network (RNN)	Convolutional Neural Network (CNN)	Long Short Term Memory (LSTM)	Bidirectional Long Short Term Memory(BiLSTM)

Tab. 3.1 : Modèles de réseaux de neurones artificiels

1. **Le réseau de neurone récurrent (RNN)** est l'algorithme le plus avancé pour les données de séquence. En raison de la mémoire interne, c'est le premier algorithme à se souvenir de son entrée, ce qui le rend très approprié pour les problèmes d'apprentissage impliquant automatiquement des données de séquence. C'est l'un des algorithmes à l'origine des réalisations étonnantes observées dans le domaine du deep learning ces dernières années.

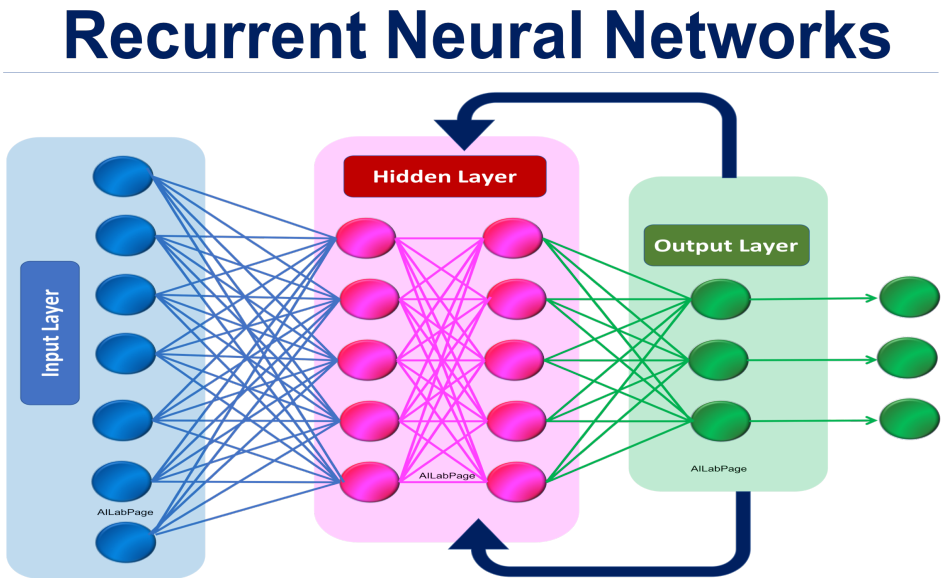

FIG. 3.3 : réseaux de neurones récurrents [66].

2. **Le réseau de neurones convolutifs (ConvNet / CNN)** est un algorithme d'apprentissage profond qui peut prendre des images en entrée, attribuer une importance (poids et biais apprenables) à divers aspects/objets de l'image et être capable de les distinguer les uns des autres. Le prétraitement requis dans ConvNet est bien inférieur à celui des autres algorithmes de classification.

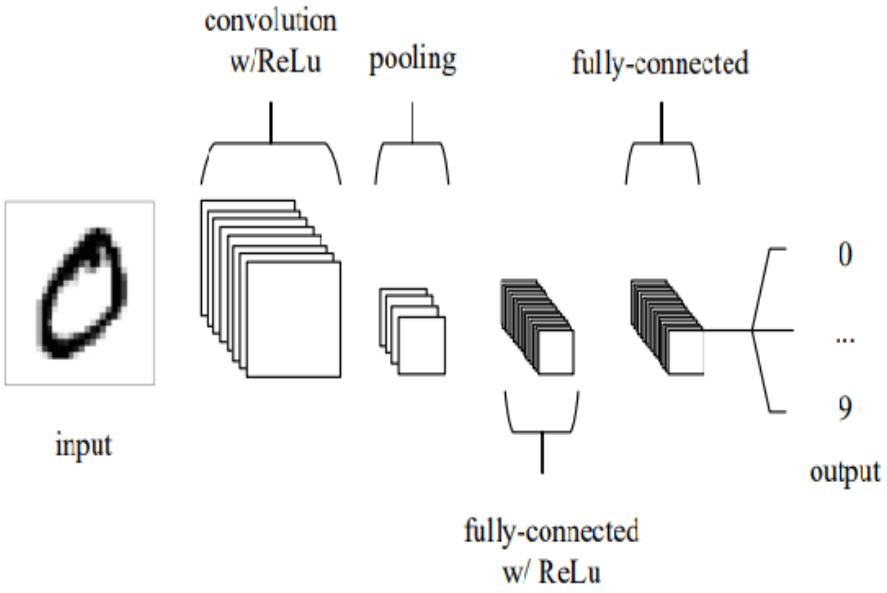

FIG. 3.4 : réseaux de neurones convolutif [67].

3. **Les réseaux à mémoire à long terme (LSTM)** sont un type particulier de RNN, capable d'apprendre des dépendances à long terme, ils sont explicitement conçus pour éviter le problème de dépendance à long terme. Se souvenir d'informations pendant de longues périodes est pratiquement leur comportement par défaut. Tous les réseaux neuronaux récurrents ont la forme d'une chaîne de modules répétitifs de réseau neuronal.

4. **Le réseau de neurones récurrents bidirectionnels (BiLSTM)** consiste en fait à assembler deux RNN indépendants. Cette structure permet au réseau d'avoir des informations en amont et en aval de la séquence à chaque pas de temps. Cette méthode est différente de la méthode à sens unique car dans le LSTM fonctionnant en sens inverse, vous conservez des informations du futur et en combinant deux états cachés, vous pouvez conserver des informations du passé et du futur à tout moment.

3.3.4 Modèles de base

1. **Naive bayes** : Le Classificateur naïf de Bayes est très simple et efficace. Par conséquent, la recherche sur les paramètres d'évaluation de la fonction est très néces-

saire. exister Étant donné que de nombreux ensembles de données de texte sont multi-catégories, il est naturel d'ajuster le ratio De la côte à de nombreux types de problèmes, bien que le modèle soit simple et les hypothèses limitées Il est formulé pour l'indépendance [68].

2. **Arborescence des décisions (DT)** : Le classificateur d'arbre de décision est un algorithme de classification simple et couramment pour classifier Les données. L'arbre de décision représente une structure arborescente et les nœuds internes représentent Les conditions de test et les nœuds feuilles sont utilisés comme étiquettes de classe [69]. Cette méthode de classement peut être appliquée à tous les types de données, tels que nominaux, ordonnés et numériques. Les données de test sont classifiées très rapide à l'aide de l'algorithme d'arbre de décision.

3. **Machine Vectorielle de Support (SVM)** :Cet algorithme est adapté à une stratégie de séparation d'hyperplans simple. considérant Pour les données d'entraînement, l'algorithme classe les données de test dans un hyperplan optimal. Les points de données sont tracées dans un espace vectoriel à N dimensions (N dépend des caractéristiques du point Les données) [70]. L'algorithme SVM est utilisé pour les tâches de classification binaire et de régression.

4. **Classificateur de régression logistique (RL)** :L'algorithme est nommé d'après la fonction centrale qui y est utilisée, c'est-à-dire la fonction logistique. Les fonctions logiques sont également appelées fonctions sigmoïdes. C'est une courbe en forme de S, prenez la valeur réelle comme entrée et convertissez-la dans la plage comprise entre 0 et 1 [71].

3.4 Évaluation

pour evaluer les resultats nous utiliserons plusieurs paramètres tels que (Accuracy, Precision, Recall) qui seront décrits ci-dessous. décrits ci-dessous

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

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

- **Precision (P)** : c'est le pourcentage de l'exactitude des situations positives qui ont été déterminées avant.

$$P = \frac{TP}{(TP + FP)} \tag{3.2}$$

- **Recall (R)** : appelé aussi Sensibilité, c'est le pourcentage des situations réelles qui ont été déterminées avec précision.

$$R = \frac{TP}{(TP + TN)} \tag{3.3}$$

Chapitre 4

Implémentation et résultats

4.1 Introduction

Les environnements d'apprentissage profond et d'apprentissage automatique sont très puissants et facilitent la révolution de l'IA car ils fournissent des outils pré-implémentés et des modèles pour le prétraitement, la classification et l'évaluation. Sans ces outils, il serait très difficile pour les scientifiques de travailler sur des tâches d'apprentissage automatique. Cette sous-section passe en revue les différents cadres de développement et services de Cloud utilisés dans cette thèse.

4.2 L'environnement de travail et les outils utilisés

4.2.1 Matériel

Le matériel utilisé est représenté dans le tableau suivant :

	POSTE DE TRAVAIL
Pc	ASUS
Système d'exploitation	Windows 10 Professionnel
Processeur	Processeur Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz
RAM	8,00 Go
Type de système	SE 64 bits

4.2.2 Python

Python est un langage de programmation de haut niveau et sa philosophie de conception est axée sur la lisibilité du code et une syntaxe qui permet aux programmeurs d'exprimer des concepts en quelques lignes de code. Python est développé sous une licence open source approuvée par l'OSI, ce qui le rend librement utilisable et distribuable, même à des fins commerciales. Il est utilisé avec succès dans des milliers d'applications professionnelles réelles dans le monde entier, y compris dans de nombreux systèmes importants et critiques. La version de python utilisée dans ce travail est la 3.9 [72].

FIG. 4.1 : python Logo.

4.2.3 Editeur de code

Nous avons utilisé jupyter notebook pour éditer le code de notre systéme , qui est une application Web open source qui vous permet de créer et de partager des documents contenant du code en temps réel, des équations, des visualisations et du texte narratif. Les utilisations incluent : le nettoyage et la conversion des données, la simulation numérique, la modélisation statistique, la visualisation des données et l'apprentissage automatique [73].

4.2.4 Librairies et bibliothèques Python

FIG. 4.2 : : Editeur de code et bibliothèques Python utilisés.

- **Pandas** : C'est une bibliothèque ecrite pour le langage python permet la manipulation et l'analyse de données.
- **NLTK** : est la principale plate-forme Python pour le traitement des données de langage la nature. Nltk fournit également une bibliothèque de prétraitement des données, Classification, segmentation,ect.
- **Scikit-learn** :est une bibliothèque Python gratuite pour l'apprentissage automatique. Elle comprend de nombreuses classifications, régressions et Regroupement.
- **Matplotlib** [74] : est une bibliotheque destinée à tracer et visualiser des données sous formes de graphiques. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy.
- **Keras** : est une API de réseau de neurones écrite en langage Python capable de s'exécuter sur TensorFlow [75].

4.3 Analyse exploratoire des données

4.3.1 Popularité des hashtags

Dans cette section nous avons utilité les tweets qui contient une liste de hashtags spécifiques pour decouvrir quels sont les hashtags les plus populaires utilisés durant une durée précise.

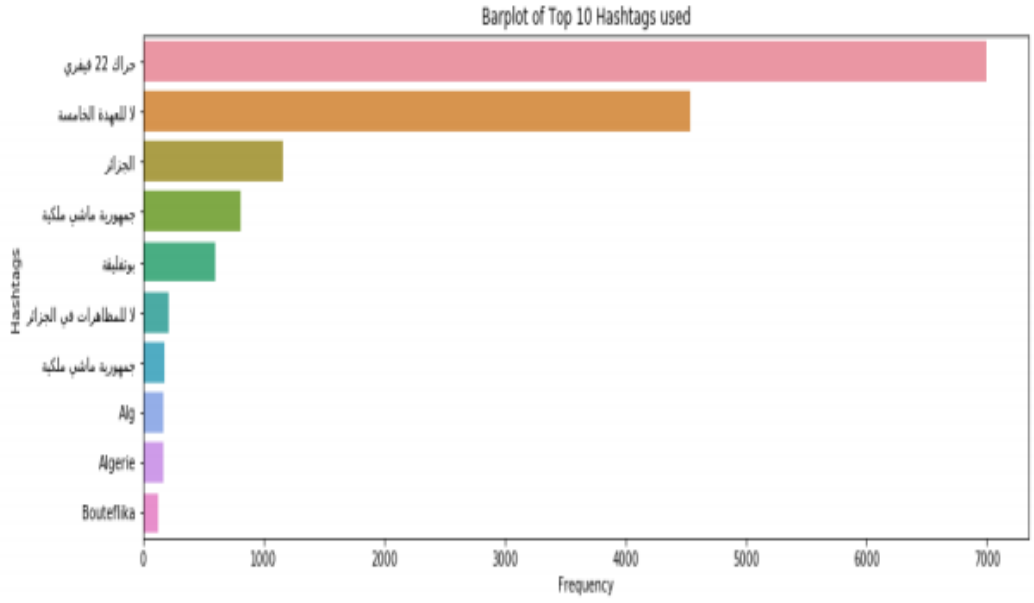

Fig. 4.3 : : Top 10 Hashtags avant et durant le 22-02-2019 Bigrams.

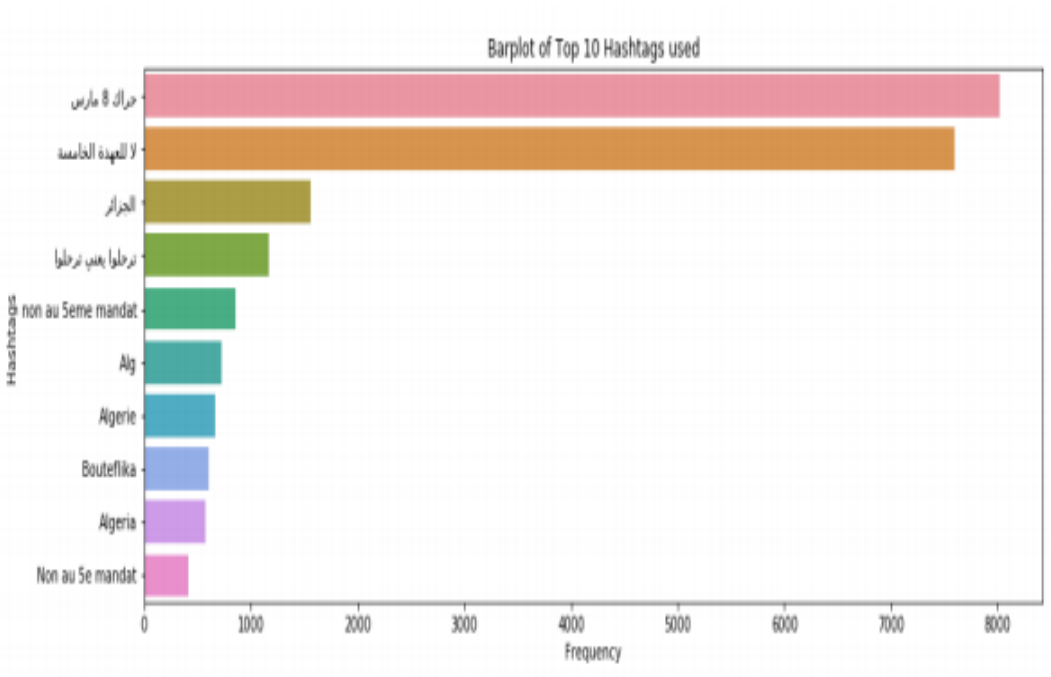

Fig. 4.4 : : Top 10 Hashtags avant et durant le 08-03-2019 Bigrams.

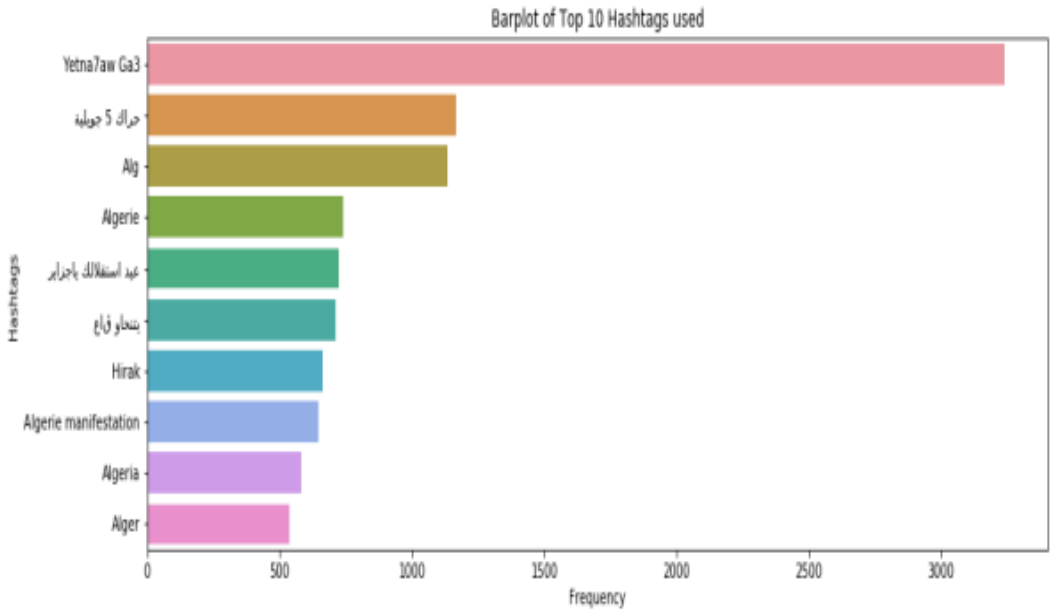

Fig. 4.5 : : Top 10 Hashtags avant et durant le 05-07-2019 Bigrams.

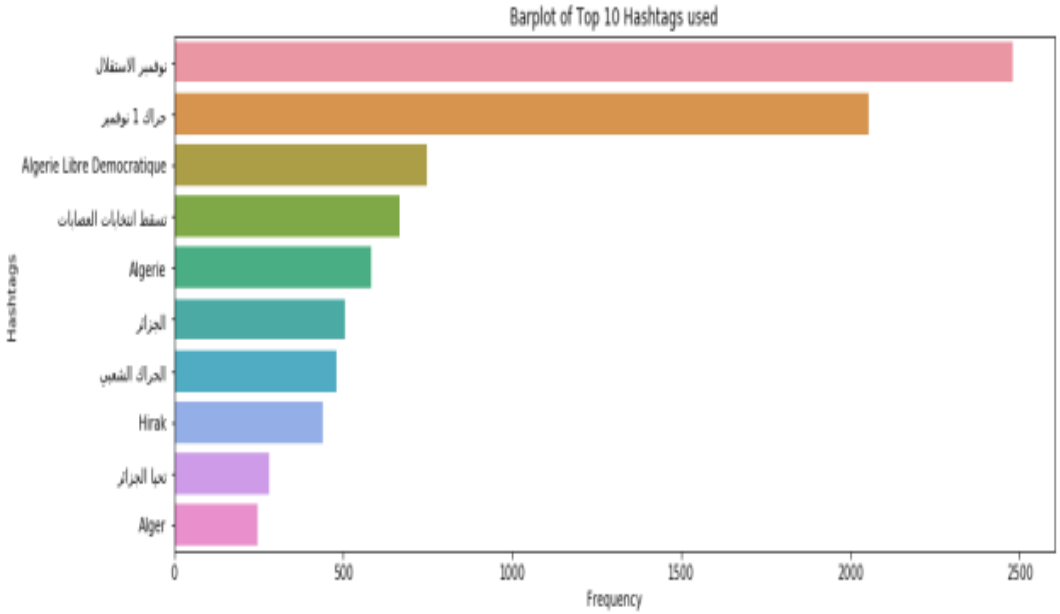

Fig. 4.6 : : Top 10 Hashtags avant et durant le 01-11-2019 Bigrams.

4.3.2 Distribution de données

Les figures suivantes illustre la distribution de données selon les attributs «Text» et «Target».

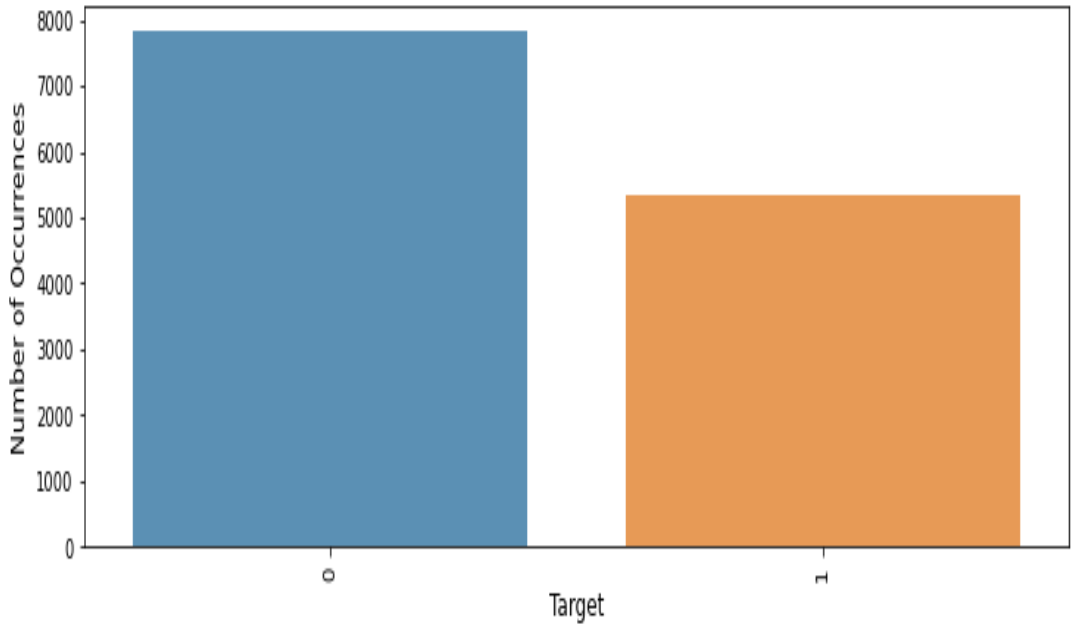

Fig. 4.7 : : Distribution des données du 22-02-2019.

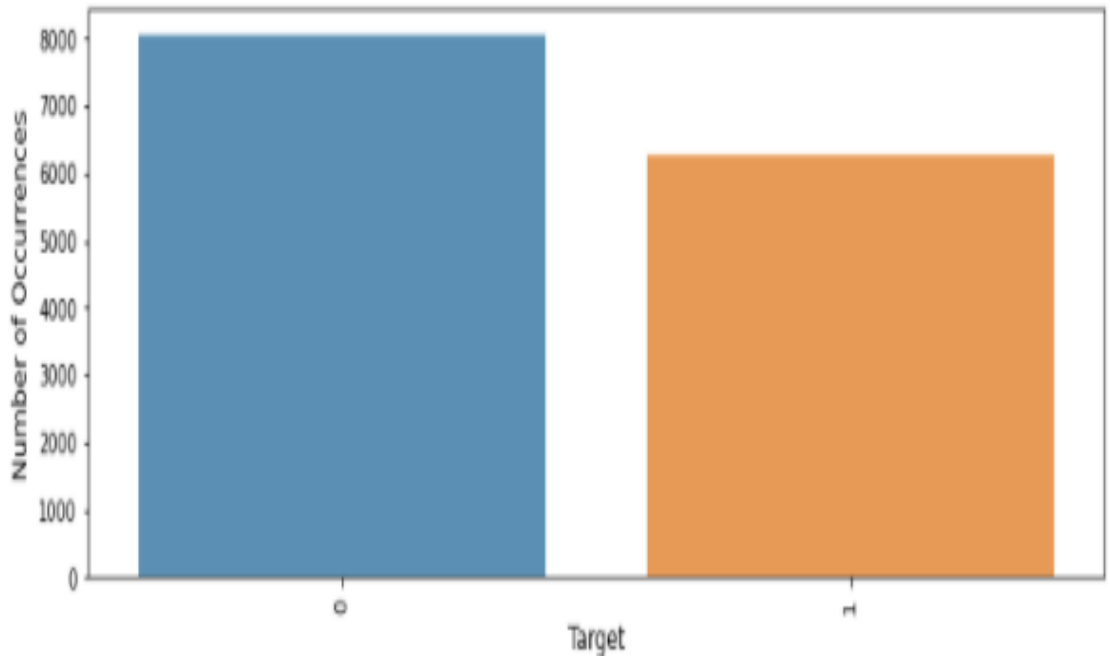

Fig. 4.8 : : Distribution des données du 08-03-2019.

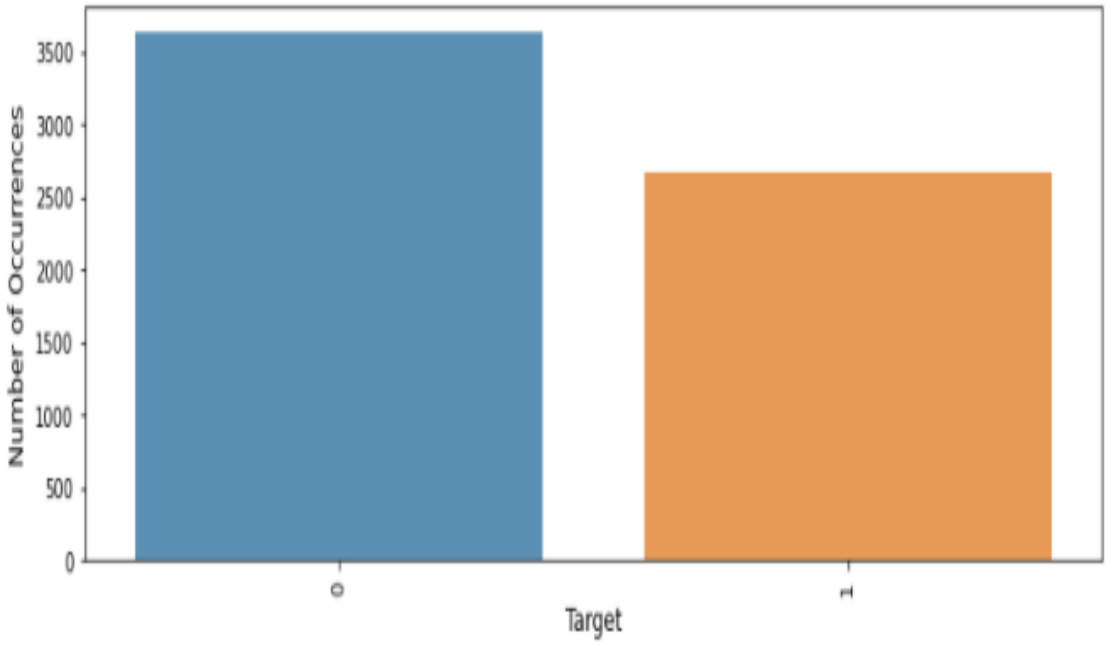

FIG. 4.9 : : Distribution des données du 05-07-2019.

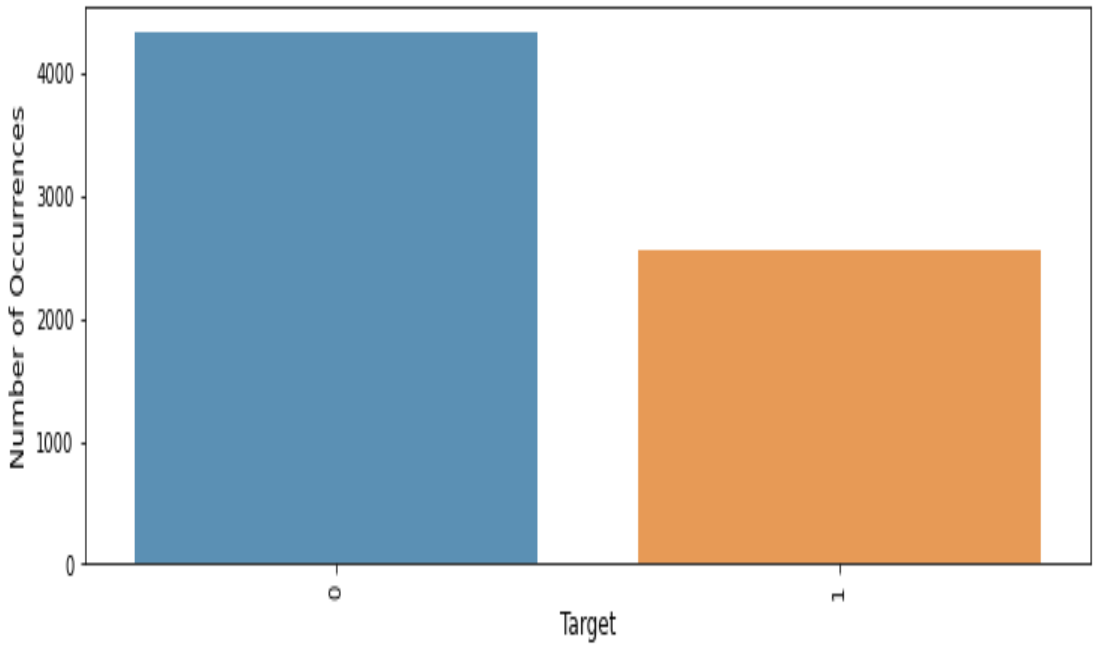

FIG. 4.10 : : Distribution des données du 01-11-2019.

4.4 Résultats de classification

4.4.1 par les modèles de base

La figure (Fig 4.11) montre les résultats de classification par les modéles de base pour chaque évènement.

Méthodes de classification / Evenements		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	0.706	**0.728**	**0.730**	0.716	**0.704**	0.689	0.730	**0.738**
	Precision	0.635	**0.676**	0.675	**0.746**	**0.642**	0.612	0.666	**0.705**
	Recall	0.653	**0.663**	**0.669**	0.577	0.637	**0.638**	**0.677**	0.663
Le 08-03-2019	Accuracy	**0.703**	0.702	**0.719**	0.706	**0.691**	0.674	**0.733**	0.726
	Precision	0.655	**0.673**	0.686	**0.748**	**0.650**	0.627	0.700	**0.750**
	Recall	**0.673**	0.655	**0.680**	0.604	**0.652**	0.637	**0.697**	0.648
Le 05-07-2019	Accuracy	0.695	**0.713**	0.691	**0.693**	**0.690**	0.672	**0.734**	0.726
	Precision	0.663	**0.711**	0.648	**0.726**	**0.655**	0.629	0.736	**0.772**
	Recall	**0.638**	0.637	**0.644**	0.576	**0.634**	0.617	**0.666**	0.628
Le 01-11-2019	Accuracy	0.696	**0.705**	**0.685**	0.682	**0.686**	0.658	**0.720**	0.703
	Precision	0.607	**0.644**	0.610	**0.742**	**0.586**	0.546	0.668	**0.703**
	Recall	**0.567**	0.543	**0.510**	0.354	**0.564**	0.522	**0.571**	0.475

FIG. 4.11 : Résultats de classification par les modéles de base pour chaque évènement.

4.4.2 Par les réseaux de neurones

Dans cette étape nous allons afficher les courbes montrant les étapes de développement de ”Accuracy” et ”epochs” dans chaque évènement.

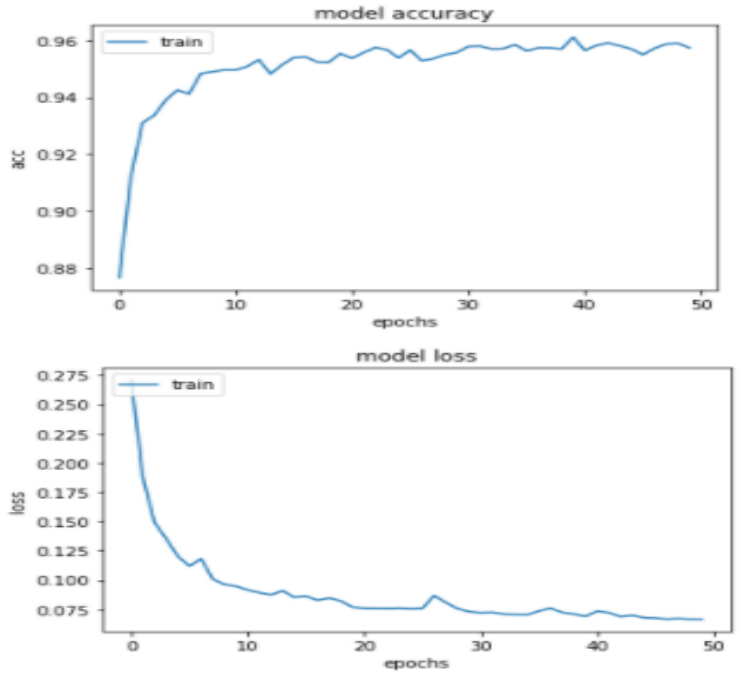

Fig. 4.12 : précision de l'entraînement et du test par rapport aux époques du 22-02-2019.

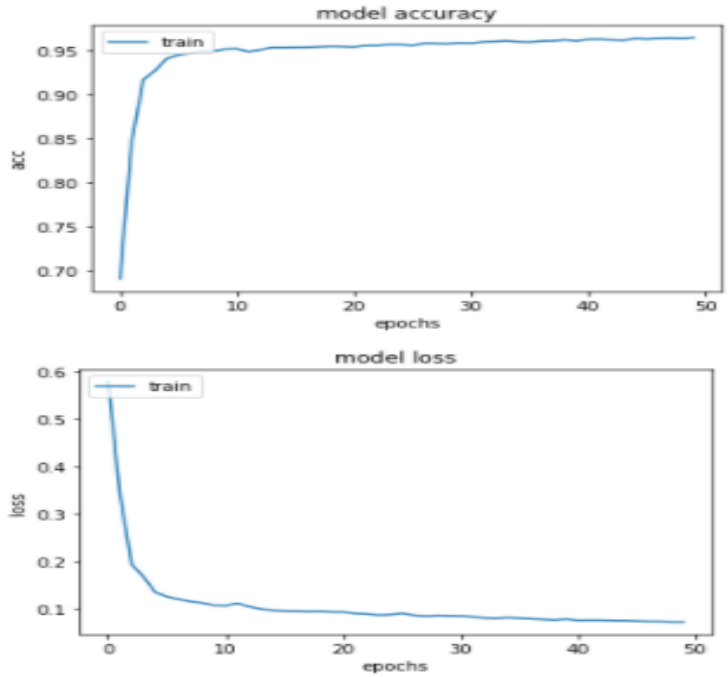

Fig. 4.13 : précision de l'entraînement et du test par rapport aux époques du 08-03-2019.

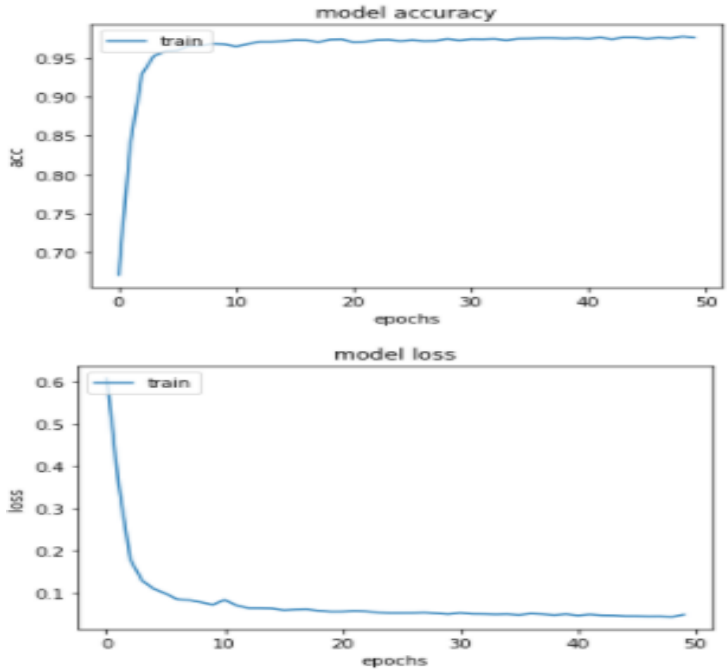

FIG. 4.14 : précision de l'entraînement et du test par rapport aux époques du 05-07-2019.

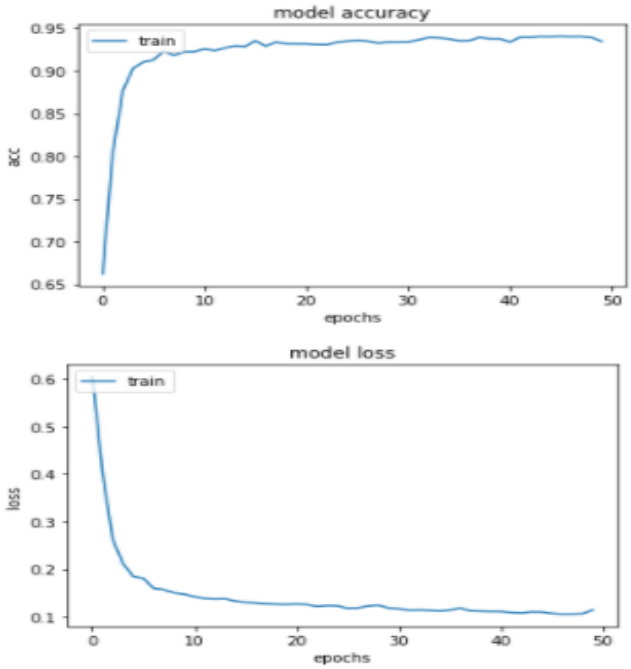

FIG. 4.15 : précision de l'entraînement et du test par rapport aux époques du 01-11-2019.

Le Tableau suivant montre les résultats de classification avec les réseaux de neurones (LSTM) pour chaque évènement.

	22-02-2019	08-03-2019	05-07-2019	01-11-2019
Accuracy	0.63	0.64	0.67	0.64

Tab. 4.1 : Les résultats obtenus par les réseaux de neurones (LSTM)

4.5 Comparaison et discussion des résultats

D'après les résultats que nous avons obtenus, nous avons remarqué que le modèle Classification de base Naive Bayes, arbre de décision, régression logistique et Vecteur de support SVM, donnent de bonnes performances performances par rapport au modèle de l'apprentissage profond.D'autre part, nous constatons que le modèle de régression logistique dépasse les autres classificateurs de base et fonctionnent mieux en termes de toutes les mesures. Ceci n'est pas surprenant parce que LR est connu de sa capacité à supporter les données textuelles par rapport aux autres classificateurs classiques.
Les résultats du modèle d'apprentissage profond LSTM ont été obtenus par la quantité des données liées au évenements du <hirak>. En effet, la taille de l'ensemble de données était bénéfique pour les classificateurs de base, a été de l'autre côté négative pour les classificateurs profonds. Les modèles d'apprentissage profond exigent des très grandes quantités de données pour aboutir à des meilleurs scores de classification.

4.6 Conclusion

Ce chapitre a englobé les principaux résultats de la thèse, les expériences ont été détaillées et expliquées, les résultats ont été illustrés sous forme de figures et de tableaux afin de bien comprendre les limitations et les performances des modèles de classification appliqués. Les résultats peuvent être améliorés en se concentrant sur les étapes de prétraitement.

Conclusion et perspectives

Conclusion générale

La connaissance du traitement du langage associée aux concepts de l'intelligence artificielle, l'apprentissage automatique et l'apprentisage profond 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 ces apprentissages 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 ce contexte, notre travail vise plusieurs objectifs. En premier lieu, il nous a permis d'explorer le domaine d'échange informationnel humain sur le web et les plateformes sociales et comprendre toutes ses complexités en termes de traitement, de contrôle et d'orientation vers la bonne voie. En outre, nous avons eu l'opportunité de mettre en pratique toutes nos connaissances dans le domaine d'apprentissage automatique et d'apprentissage profond sur un thème aussi important. Nous avons eu également la chance de travailler sur le domaine de traitement de langage naturel, qui est un domaine très actif et très prometteur et dont la nécessité pour la société moderne s'accroit jour après jour.

Les méthodes décrites dans cette thèse ont données des résultats raisonnables, comme le montre le dernier chapitre. Cependant, Par rapport à l'apprentissage profond nous pouvons penser à améliorer notre système en entraînant le modèle sur un plus grand jeu de données, et en effectuant plus d'expériences pour évaluer notre extracteur d'événements. Concernant l'apprentissage automatique nous proposons de pratiquer différents algorithmes d'apprentissage automatique pour améliorer la précision de la prédiction.

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

- 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 rédaction avec LATEX en ligne (Overleaf) à cause des coupures d'internet et le mauvais flux d'Internet pendant cette période avec le stress que nous avons vécu de l'émergence du virus Covid-19.

Bibliographie

[1] S. P. Borgatti, M. G. Everett et J. C. Johnson, *Analyzing social networks*. Sage, 2018.

[2] C. T. Carr et R. A. Hayes, "Social media : Defining, developing, and divining," *Atlantic journal of communication*, t. 23, n° 1, p. 46-65, 2015.

[3] A. Drif et K. Hadjoudj, "An Opinion Spread Prediction Model With Twitter Emotion Analysis During Algeria's Hirak," *The Computer Journal*, t. 64, n° 3, p. 358-368, 2021.

[4] F. Barboza, H. Kimura et E. Altman, "Machine learning models and bankruptcy prediction," *Expert Systems with Applications*, t. 83, p. 405-417, 2017.

[5] R. Vargas, A. Mosavi et R. Ruiz, "Deep learning : a review," *Advances in intelligent systems and computing*, 2017.

[6] M. Al-Ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh et B. Gupta, "Deep learning for Arabic NLP : A survey," *Journal of computational science*, t. 26, p. 522-531, 2018.

[7] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, t. 37, n° 1, p. 51-89, 2003.

[8] F. Millstein, *Natural language processing with python : natural language processing using NLTK*. Frank Millstein, 2020.

[9] E. D. Liddy, "Natural language processing," 2001.

[10] S. Maurel, P. Curtoni et L. Dini, "L'analyse des sentiments dans les forums," *Atelier Fouille des Données d'Opinion*, 2008.

[11] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter et H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data*, t. 2, n° 1, p. 1-24, 2015.

[12] I. H. Sarker, "Machine learning : Algorithms, real-world applications and research directions," *SN Computer Science*, t. 2, n° 3, p. 1-21, 2021.

[13] L. Yao, C. Mao et Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," *BMC medical informatics and decision making*, t. 19, n° 3, p. 71, 2019.

[14] M. A. Hanif, F. Khalid, R. V. W. Putra, M. T. Teimoori, F. Kriebel, J. J. Zhang, K. Liu, S. Rehman, T. Theocharides, A. Artusi et al., "Robust Computing for Machine Learning-Based Systems," in *Dependable Embedded Systems*, Springer, Cham, 2021, p. 479-503.

[15] S. Jain, A. K. Jain et S. P. Singh, "Building a Machine Learning Model for Unstructured Text Classification : Towards Hybrid Approach," in *Rising Threats in Expert Applications and Solutions*, Springer, 2021, p. 447-454.

[16] K. P. Murphy et al., "Naive bayes classifiers," *University of British Columbia*, t. 18, n° 60, 2006.

[17] I. Kerenidis, A. Prakash et D. Szilágyi, "Quantum algorithms for second-order cone programming and support vector machines," *Quantum*, t. 5, p. 427, 2021.

[18] B. Charbuty et A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, t. 2, n° 01, p. 20-28, 2021.

[19] Z. Yong, L. Youwen et X. Shixiong, "An improved KNN text classification algorithm based on clustering," *Journal of computers*, t. 4, n° 3, p. 230-237, 2009.

[20] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani et J. Fowler, "Online social networks and offline protest," *EPJ Data Science*, t. 4, n° 1, p. 1-9, 2015.

[21] N. Kallus, "Predicting crowd behavior with big public data," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, p. 625-630.

[22] K. Clarke et K. Kocak, "Launching revolution : Social media and the Egyptian uprising's first movers," *British Journal of Political Science*, t. 50, n° 3, p. 1025-1045, 2020.

[23] A. Bruns, T. Highfield et J. Burgess, "The Arab Spring and social media audiences : English and Arabic Twitter users and their networks," *American behavioral scientist*, t. 57, n° 7, p. 871-898, 2013.

[24] P. Panagiotopoulos, A. Z. Bigdeli et S. Sams, "" 5 Days in August"–How London local authorities used Twitter during the 2011 riots," in *International Conference on Electronic Government*, Springer, 2012, p. 102-113.

[25] A. Gupta, A. Joshi et P. Kumaraguru, "Identifying and characterizing user communities on twitter during crisis events," in *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media*, 2012, p. 23-26.

[26] M. Cheong, S. Ray et D. Green, "Interpreting the 2011 London riots from twitter metadata," in *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, IEEE, 2012, p. 915-920.

[27] M. Tremayne, "Anatomy of protest in the digital era : A network analysis of Twitter and Occupy Wall Street," *Social Movement Studies*, t. 13, n° 1, p. 110-126, 2014.

[28] Y. Theocharis, W. Lowe, J. W. Van Deth et G. García-Albacete, "Using Twitter to mobilize protest action : online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements," *Information, Communication & Society*, t. 18, n° 2, p. 202-220, 2015.

[29] M. D. Conover, E. Ferrara, F. Menczer et A. Flammini, "The digital evolution of occupy wall street," *PloS one*, t. 8, n° 5, e64679, 2013.

[30] K. Bajpai et A. R. Jaiswal, "A framework for analyzing collective action events on Twitter.," in *ISCRAM*, 2011.

[31] Y. Hu, F. Wang et S. Kambhampati, "Listening to the crowd : Automated analysis of events via aggregated twitter sentiment.," in *IJCAI*, Citeseer, 2013, p. 2640-2646.

[32] T. Sakaki, M. Okazaki et Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Transactions on Knowledge and Data Engineering*, t. 25, n° 4, p. 919-931, 2012.

[33] C. W. Woo, M. P. Brigham et M. Gulotta, "Twitter Talk and Twitter Sharing in Times of Crisis : Exploring Rhetorical Motive and Agenda-Setting in the Ray Rice Scandal," *Communication Studies*, t. 71, n° 1, p. 40-58, 2020.

[34] S. Brown, "Twitter Usage in Times of Crisis," *Open Access Journals for School Teachers in Indonesia*, t. 29, 2011.

[35] T. Sakaki, M. Okazaki et Y. Matsuo, "Earthquake shakes twitter users : real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, 2010, p. 851-860.

[36] K. Starbird, G. Muzny et L. Palen, "Learning from the crowd : Collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions.," in *ISCRAM*, Citeseer, 2012.

[37] M. Li, N. Turki, C. R. Izaguirre, C. DeMahy, B. L. Thibodeaux et T. Gage, "Twitter as a tool for social movement : An analysis of feminist activism on social media communities," *Journal of community psychology*, t. 49, n° 3, p. 854-868, 2021.

[38] D. Ray et M. Tarafdar, "How Does Twitter Influence a Social Movement ?," 2017.

[39] K. Hunt, "Twitter, social movements, and claiming allies in abortion debates," *Journal of Information Technology & Politics*, t. 16, n° 4, p. 394-410, 2019.

[40] D. Isa et I. Himelboim, "A social networks approach to online social movement : Social mediators and mediated content in# FreeAJStaff Twitter network," *Social Media+ Society*, t. 4, n° 1, p. 2 056 305 118 760 807, 2018.

[41] J. M. Larson, J. Nagler, J. Ronen et J. A. Tucker, "Social networks and protest participation : Evidence from 130 million Twitter users," *American Journal of Political Science*, t. 63, n° 3, p. 690-705, 2019.

[42] C. A. D. L. Salge et E. Karahanna, "Protesting corruption on Twitter : Is it a bot or is it a person ?" *Academy of Management Discoveries*, t. 4, n° 1, p. 32-49, 2018.

[43] J. Cadena, G. Korkmaz, C. J. Kuhlman, A. Marathe, N. Ramakrishnan et A. Vullikanti, "Forecasting social unrest using activity cascades," *PloS one*, t. 10, n° 6, e0128879, 2015.

[44] S. González-Bailón, J. Borge-Holthoefer, A. Rivero et Y. Moreno, "The dynamics of protest recruitment through an online network," *Scientific reports*, t. 1, n° 1, p. 1-7, 2011.

[45] A. Hermida et V. Hernández-Santaolalla, "Twitter and video activism as tools for counter-surveillance : the case of social protests in Spain," *Information, Communication & Society*, t. 21, n° 3, p. 416-433, 2018.

[46] R. Compton, C. Lee, T.-C. Lu, L. De Silva et M. Macy, "Detecting future social unrest in unprocessed twitter data :"emerging phenomena and big data"," in *2013 IEEE International Conference on Intelligence and Security Informatics*, IEEE, 2013, p. 56-60.

[47] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz et N. Ramakrishnan, "Planned protest modeling in news and social media," in *Twenty-Seventh IAAI Conference*, Citeseer, 2015.

[48] K. Radinsky et E. Horvitz, "Mining the web to predict future events," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, p. 255-264.

[49] R. Korolov, D. Lu, J. Wang, G. Zhou, C. Bonial, C. Voss, L. Kaplan, W. Wallace, J. Han et H. Ji, "On predicting social unrest using social media," in *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, IEEE, 2016, p. 89-95.

[50] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz et al., "'Beating the news' with EMBERS : Forecasting civil unrest using open source indicators," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, p. 1799-1808.

[51] A. Doyle, G. Katz, K. Summers, C. Ackermann, I. Zavorin, Z. Lim, S. Muthiah, P. Butler, N. Self, L. Zhao et al., "Forecasting significant socie-

tal events using the embers streaming predictive analytics system," *Big data*, t. 2, n° 4, p. 185-195, 2014.

[52] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti et 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 2015*, 2015, p. 258-265.

[53] V. Taylor, "Social movement continuity : The women's movement in abeyance," *American Sociological Review*, p. 761-775, 1989.

[54] J. W. Veugelers, "Dissenting families and social movement abeyance : the transmission of neo-fascist frames in postwar Italy 1," *The British journal of sociology*, t. 62, n° 2, p. 241-261, 2011.

[55] N. Alsaedi, P. Burnap et O. Rana, "Can we predict a riot ? Disruptive event detection using Twitter," *ACM Transactions on Internet Technology (TOIT)*, t. 17, n° 2, p. 1-26, 2017.

[56] P. Pathak, "„Role of Social media in reference to North East Ethnic violence ," *IOSR Journal of Humanities And Social Science (IOSR-JHSS)*, t. 19, n° 4, p. 59-66, 2014.

[57] P. Tijare et J. R. Prathuri, "A Survey on Event Detection and Prediction Online and Offline Models using Social Media Platforms," *Materials Today : Proceedings*, 2021.

[58] Y. Shi, K. Yao, L. Tian et D. Jiang, "Deep LSTM based feature mapping for query classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics : Human language technologies*, 2016, p. 1501-1511.

[59] Y. Zhang, L. Dong et S. Li, "Extraction of Elements of Protest Based on BERT Model and TextTeaser improved algorithm," in *Journal of Physics : Conference Series*, IOP Publishing, t. 1955, 2021, p. 012 107.

[60] X. Zhou, X. Wan et J. Xiao, "Attention-based LSTM network for cross-lingual sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, p. 247-256.

[61] S. Belguidoum, "Hirak et crise du système néo-patrimonial en Algérie : rupture générationnelle et nouvelle temporalité historique," *Insaniyat/ . Revue algérienne d'anthropologie et de sciences sociales*, nº 88, p. 31-50, 2020.

[63] E. Guresen et G. Kayakutlu, "Definition of artificial neural networks with comparison to other networks," *Procedia Computer Science*, t. 3, p. 426-433, 2011.

[64] S. Shanmuganathan, "Artificial neural network modelling : An introduction," in *Artificial neural network modelling*, Springer, 2016, p. 1-14.

[65] C. Touzet, *les réseaux de neurones artificiels, introduction au connexionnisme*. Ec2, 1992.

[66] M. N. Fekri, H. Patel, K. Grolinger et V. Sharma, "Deep learning for load forecasting with smart meter data : Online adaptive recurrent neural network," *Applied Energy*, t. 282, p. 116 177, 2021.

[67] K. O'Shea et R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv :1511.08458*, 2015.

[68] P. Abhilash et D. Chakradhar, "Sustainability improvement of WEDM process by analysing and classifying wire rupture using kernel-based naive Bayes classifier," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, t. 43, nº 2, p. 1-9, 2021.

[69] S. Koga, X. Zhou et D. W. Dickson, "Machine learning-based decision tree classifier for the diagnosis of progressive supranuclear palsy and corticobasal degeneration," *Neuropathology and Applied Neurobiology*, 2021.

[70] R. Kashef, "A boosted SVM classifier trained by incremental learning and decremental unlearning approach," *Expert Systems with Applications*, t. 167, p. 114 154, 2021.

[71] J. S. Humphries, "Surviving the Veracity of a Data Onslaught,"

[74] J. D. Hunter, "Matplotlib : A 2D graphics environment," *Computing in science & engineering*, t. 9, nº 03, p. 90-95, 2007.

[75] N. Ketkar, "Introduction to keras," in *Deep learning with Python*, Springer, 2017, p. 97-111.

Webographie

[62] Web, *GetOldTweets3*, 2019. adresse : https://pypi.org/project/GetOldTweets3/.

[72] ——, *python*. adresse : https://www.python.org/.

[73] *jupyter notebook*, 2020. adresse : https://jupyter.org/.