


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master

en Informatique

Spécialité : *Ingénierie des systémes d'information et du logiciel*

Théme

Analyse d'opinion politique dans les réseaux sociaux

Encadré par

— Amad Mourad

Réalisé par

— Faci Lina

— Tiabi Imene

2019/2020

Remerciements

” Ils t’interrogeront sur l’esprit ; dis : l’esprit est du domaine exclusif de mon seigneur et vous n’avez reçu de la science que fort peu de choses” Coran- le voyage nocturne 8 4 (El Israa)

Nous tenons à remercier Le Grand Dieu de nous avoir donné la force et le courage pour poursuivre nos études, veuille t’il guider nos pas dans le droit chemin.

Nous tenons à exprimer notre reconnaissance à notre encadreur monsieur AMAD Mourad, pour avoir accepté de nous encadrer dans cette étude.

Nous remercions le président du jury et les examinateurs d’avoir accepter de juger notre travail.

Un merci particulier à nos sources de courage, nos parents, pour leur amour, leurs sacrifices et leurs patiences.

Dédicaces

Je dédie ce travail :

Tout d'abord à mes parents, à mon père qui a tout donné et sacrifié. À ma chère mère qui m'a mis au monde et a veillé à mon bonheur Je leurs épreuves ma profonde gratitude. Je vous aime et que Dieu vous garde pour nous tous. Amène

A ma grande mère à qui je souhaite une longue vie.

- A mes très chers frères que j'admire beaucoup.
- A ma chère sœur Zineb.
- A mes oncles et mes tantes.
- A mon binôme Imane.
- A tous mes amis et à toute ma promotion.

lyna.

Dédicaces

Je dédie ce travail :

- A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études.
- A mes chères sœurs Asma et Soumia pour leurs encouragements permanents, et leur soutien moral et aussi ma nièce Meriem et mon neveu Zakaria que j'adore.
- A mes chères amies pour leur appui et leur encouragement.
- A ma binôme Lina qui a fourni énormément d'efforts.
- A toute ma famille et à toutes les personnes que j'aime pour leur soutien tout au long de mon parcours universitaire.

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fruit de votre soutien infaillible merci d'être toujours là pour moi.

Imane.

Résumé

Les réseaux sociaux sont une excellente source d'information et d'extraction d'opinion ou la majorité des internautes utilisent ces plateformes pour partager leurs sentiments et opinions. L'exploitation de ces opinions ne peut être que fructueuse. Dans notre travail, nous avons exposer le problème de l'analyse des sentiments sur les réseaux sociaux en présentant une nouvelle méthode de classification supervisé des opinions politiques concernant les élections présidentielles républicaines américaines en deux parties : républicain, démocrate faites dans ce contexte sur des Tweets.

Mots-clés : Fouille des opinions, Twitter, Classification supervisé, Elections.

Abstract

Social networks are an excellent source of information and opinion extraction where the majority of Internet users use these platforms to share their feelings and opinions. The exploitation of these opinions can only be fruitful. In our work, we exposed the problem of sentiment analysis on social networks by presenting a new method of supervised classification of political opinions concerning the American Republican presidential elections in two parts : Republican, Democrat made in this context on Tweets.

Keywords : Opinion mining, Twitter, Supervised classification, Elections.

الملخص

تعد الشبكات الاجتماعية مصدرًا ممتازًا للمعلومات واستخلاص الآراء حيث يستخدم غالبية مستخدمي الإنترنت هذه المنصات لمشاركة مشاعرهم وآرائهم. إن استغلال هذه الآراء لا يمكن إلا أن يكون مثمرا. في عملنا ، كشفنا عن مشكلة تحليل المشاعر على الشبكات الاجتماعية من خلال تقديم طريقة جديدة لتصنيف الآراء السياسية المتعلقة بالانتخابات الرئاسية الجمهورية الأمريكية في جزأين: جمهوري ، وديمقراطي صنع في هذا السياق على التغريدات .

الكلمات المفتاحية: التنقيب عن الرأي ، تويتر ، التصنيف المراقب .

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

TALN	Traitement Automatique du Langage Naturel
OM	Opinion Mining
ML	Machine Learning
NB	Naïve Bayes
MNB	MultiNominal Naïve Bayes
NLTK	Natural Language Toolkit
NN	Neural Network
RF	Random Forest
HTML	Hypertext Mark-up Language
URL	Uniform Resource Locator
HTTP	HyperTextTransfer Protocol
WWW	World Wide Web
XML	Extensible Mark-upLanguage

Introduction générale

Au cours des dernières années, les réseaux sociaux en ligne sont devenus de plus en plus populaires. Ces types d'applications ont revendiqué avec succès leur place parmi les services les plus célèbres sur Internet.

A cet effet, le Web nous offre un monde de l'information prodigieux, où tout le monde arrive à exprimer ses opinions et à découvrir les opinions d'autrui.

Concomitant avec cette formidable croissance des plates-formes de réseaux sociaux est née l'émergence de l'analyse des sentiments ou l'exploration de l'opinion qui est l'étude computationnelle des opinions, des sentiments et des émotions exprimés dans le texte, cette dernière est utilisée dans plusieurs domaines tels que le marketing, la médecine, la politique . . . etc.

En effet, l'utilisation des médias sociaux pour l'analyse politique devient une pratique courante, surtout en période électorale. De nombreux chercheurs et médias tentent d'utiliser les réseaux sociaux pour comprendre l'opinion publique et la tendance. Dans ce mémoire, nous utilisons le réseau social Twitter pour en extraire une grande quantité afin de pouvoir prédire l'opinion des utilisateurs en ce qui concerne les élections présidentielles républicaines américaines.

Nous avons analysé des centaines de milliers de tweets de l'année 2016 menant aux élections primaires républicaines. Nous examinons d'abord les méthodes précédentes concernant la prévision des résultats des élections avec les médias sociaux, puis nous intégrons notre compréhension des médias sociaux et proposons un modèle de prédiction pour

prédire les opinions publiques à l'égard des élections présidentielles républicaines. Nos résultats révèlent que c'est possible d'utiliser les médias sociaux pour prédire l'opinion publique.

Notre travail est présenté comme suit :

Chapitre 1 : Des généralités sur les réseaux sociaux, leurs différents types, leur historique ainsi que leur avantages.
Chapitre 2 : Présente les différentes méthodes et approches existantes de classification supervisée et non supervisée, qui pourraient intervenir dans l'analyse des sentiments que nous souhaitons mettre en évidence, ainsi que leurs avantages et limites.
Chapitre 3 : Est consacré à la présentation du dataset utilisé ainsi que le prétraitement des données (*tweets*) ensuite, l'illustration de notre méthode proposée et son implémentation détaillée.
Chapitre 4 : Consacre une présentation des outils de programmation utilisés en premier lieu, ensuite les résultats du déroulement des algorithmes existants sur notre dataset. Et enfin, les résultats du déroulement de notre méthode proposée tout en la comparant avec les méthodes existantes.
Nous clôturons notre travail avec une conclusion générale et des perspectives.

Chapitre 1

Généralités sur les réseaux sociaux

1.1 Introduction

Dans cette section nous allons aborder les différents concepts et définitions des réseaux sociaux, l'analyse des données et les différentes méthodes et travaux qui existent. Ainsi que leurs historiques tout au long des années et notamment les avantages et les dangers qui risquent de survenir.

1.2 Les Réseaux sociaux

Un réseau social est constitué d'un ensemble de personnes liées entre elles et par la force de ces liens. On peut aussi dire qu'un réseau social est un ensemble d'individus liés entre eux par des liens caractérisés par un degré de familiarité variable qui va de simple connaissance aux liens familiaux les plus étroits.
Un réseau social par définition est un regroupement d'individus ou d'organisation reliant les internautes entre eux par des échanges. Cela leurs permet de partager des opinions, des idées ou encore du contenu.
Les fonctionnalités sur tous les types de réseaux sont similaires : après s'être inscrit, on peut effectuer des recherches nominatives. Quand vous trouver un contact, il suffit de cliquer sur un bouton afin de pouvoir se mettre en relation avec cette personne.
Internet a révolutionné le concept des relations humaines en les faisant passer dans le domaine du virtuel.

1.3 Historique des réseaux sociaux

Le premier réseau social à avoir réuni toutes les catégories de base pour faire quelque chose de complet est Sivdegrees.com en 1997.De 1997 à 2001, beaucoup de plates-formes par communautés se sont créées comme par exemple : AsianAvenue [1] (*composé uniquement de la communauté asiatique*), BlackPlanet [2] (*composé uniquement de la communauté noire*), MiGente [3] (*composé uniquement de la communauté latine*).
LinkedIn [4]**,** crée en 2002 en Californie (*Etats-Unis*), qui devient un réseau professionnel très actif : En novembre 2015, le site comptait plus de 400 millions de membres issus de 170 secteurs différents d'activité dans plus de 200 pays et territoires.

Le 22 juillet 1999 ≪ **MSN Messenger** ≫ débarque. Service de messagerie instantanée il connaitra un immense succès sur le web. Il était possible de discuter, de s'envoyer des photos ou d'interférer par webcam. En 2005, MSN devient Windows Live Messenger. C'est en 2004 que la création de Facebook a vu le jour par Marc Zuckerberg alors qu'il étudiait encore à Harvard[**1**].

1. https ://AsianAvenue.com
2. https ://BlackPlanet.com
3. https ://MiGente.com
4. https ://fr.linkedin.com/

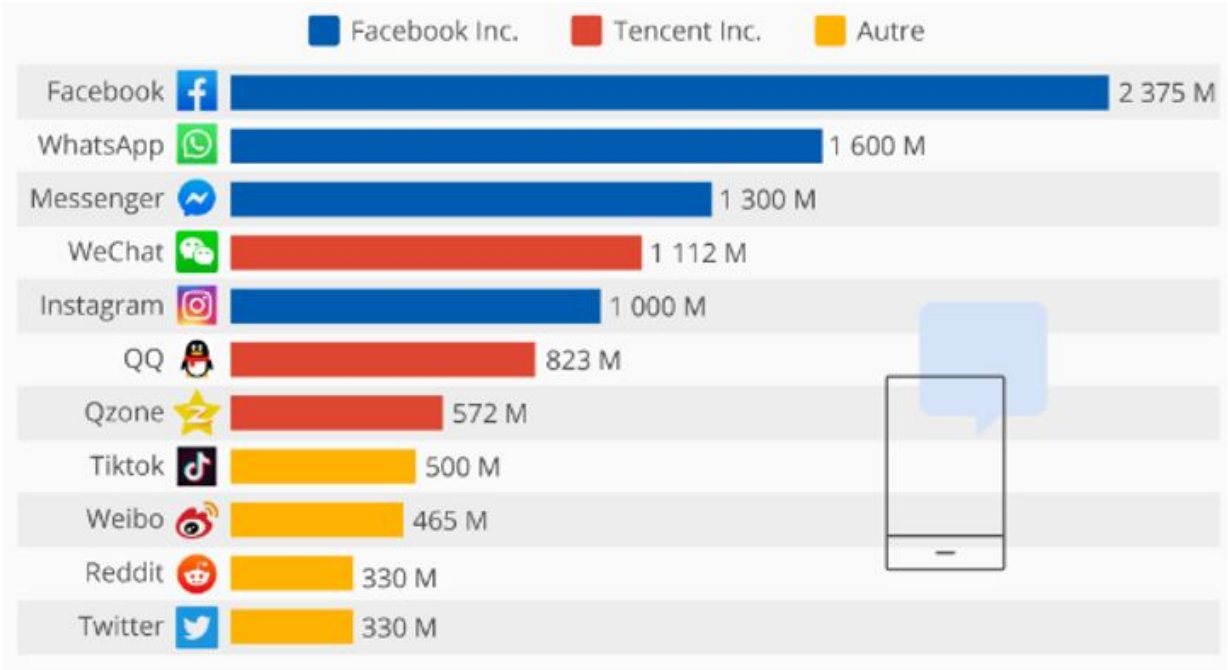

FIGURE 1.1: Enchainement des réseaux sociaux 2019[**2**]

1.4 Différents types de réseaux sociaux

Chaque réseau social a été crée dans un but précis. Ce qui a engendré l'appartenance de ce dernier à un type bien précis issus du but dont on l'avait inventé.

1.4.1 Les réseaux sociaux dits « Généralistes»

Facebook [5] : Chaque internaute a la possibilité de créer son profil limité à un réseau d'amis (*personnes proches ou inconnues*) qu'il a accepté. Il permet de partager : statut, photos, liens et vidéos. Il est aussi utilisé par les entreprises, les artistes pour leur promotion grâce aux pages fans accessibles à tous[**3**].

5. https ://facebook.com

La *figure* 1.2 illustre le logo du réseau social Facebook :

FIGURE 1.2: Logo Facebook

Twitter [6] : outil de microblogging qui permet d'envoyer des messages appelés « tweets » aux internautes qui suivent chaque compte. Ce sont les « followers » soient les abonnés [**3**].

La *figure* 1.3 illustre le logo du réseau social Twitter :

FIGURE 1.3: Logo Twitter

MySpace [7] : espace web personnalisé. Possibilité de présenter des informations personnelles et de faire un blog. Ce réseaux est notamment connut grâce aux nombreux groupes musicaux qui ont pris possession de cet espace, sa popularité a baissé ces dernières années[**3**].

6. https ://twitter.com
7. https ://myspace.com/

La *figure* 1.4 illustre le logo du réseau social MySpace :

FIGURE 1.4: Logo MySpace

1.4.2 Les sites dits « De Partage » pour échanger

YouTube[8] : moins sociaux mais permettent de mettre en ligne et partager des vidéos. On peut y trouver tous types de vidéos politiques, d'humour, de sport, de musique, de cinéma, d'art... il propose également la possibilité de laisser un commentaire au-dessous la vidéo [**3**].

La *figure* 1.5 illustre le logo du réseau social Youtube :

FIGURE 1.5: Logo YouTube

FlickR[9] :site de partage de photos (*amateur ou professionnelles*) gratuit mais le nombre de téléchargement est limité. Il a été fondé en 2004.

8. https ://youtube.com
9. https ://www.flickr.com

1.4.3 Les réseaux sociaux dits « Professionnel » pour les affaires

LinkedIn : c'est un réseau professionnel, Il permet de publier et partager son CV, créer une image professionnelle. En 2020 il compte plus de 575 millions utilisateurs et revendique plus de 660 millions de membres issus de plus 200 pays.

Piwie : c'est un réseau professionnel, Il permet de publier et partager son CV, créer une image professionnelle. En 2020 il compte plus de 575 millions utilisateurs et revendique plus de 660 millions de membres issus de plus 200 pays.

InterFrench [10] : c'est un réseau international de professionnels francophiles qui recense plus de 8000 professionnels à travers le monde. Il permet aux professionnels de développer leur réseau pour trouver de nouveaux clients, fournisseurs ou partenaires.

1.4.4 Les réseaux sociaux dits « Politique »

Coolpol [11] : c'est le réseau social du parti socialiste de « toutes celles et de tous ceux qui veulent débattre et agir à gauche ! » selon le site. C'est un lieu de discussion ou les sympathisants du parti peuvent échanger. On y retrouve les événements, débats, partage d'idées, de liens, de vidéos...

Créateurs de possible : c'est le réseau social de l'UMP lancé en janvier 2010. Propose des fonctionnalités similaires à Coolpol[**3**].

1.4.5 Les réseaux sociaux dits « Géolocalisés »

Foursquare [12] **et Gowalla :** possibilité d'ajouter des amis lorsque l'on se rend quelques part avec possibilité de signaler sa présence... [**3**].

1.4.6 Les réseaux sociaux dits de service

Ma-residence [13] **:** lieu d'échange de bonnes adresses, de services et parler des relations entre voisins.

10. https ://interfrench.com/
11. https ://coolpol.com
12. https ://fr.foursquare.com/
13. https ://www.ma-residence.fr/

Copains d'avant [14] **et Trombi** [15] **:** qui permettent de retrouver des anciens camarades de classes.

Réseaux Lycée et Etnoka [16] **:** réseaux pour lycéens et étudiants où il est possible de discuter, organiser des soirées et le partage de cours[**3**].

1.5 Les réseaux sociaux : quels enjeux ?

Tout d'abord, il faut rappeler que ces informations représentent une véritable manne pour les entreprises de marketing et les publicitaires. En effet, les professionnels de marketing peuvent recueillir ainsi de précieuses informations sur les habitudes de vie, les goûts ou les préférences de millions d'internautes et ainsi mieux cibler leurs opérations de marketing en fonction du profil des internautes. Il s'agit donc de proposer des publicités personnalisées. La popularité des sites communautaires ne cesse d'attirer les annonceurs car elle représente une formidable occasion pour profiter de cette forte fréquentation et de tester l'impact sur les consommateurs. Plus d'1,6 milliard de dollars étaient dépensés en 2008 par les publicitaires sur des réseaux sociaux, contre 920 millions en 2007. MySpace propose ainsi aux annonceurs d'intégrer de la publicité ciblée sur les intérêts des internautes, de même que Facebook depuis novembre 2007 ou encore LinkedIn depuis juillet 2008.

1.6 L'importance de bien gérer son identité

A travers ces espaces personnels, les internautes souvent ne mesurent pas les risques encourus en éparpillant leurs informations personnelles sur ces sites. Ils n'ont pas conscience qu'il s'agit d'informations très personnelles voire sensibles (*opinions politiques ou religieuses, préférences sexuelles...*). Et que toutes ces informations qui font partie de la sphère privée de chaque individu vont être propulsées sur un espace public, visibles de tous (*ou presque*).

De plus, les nombreuses applications offertes sur les sites de socialisation permettent aux

14. https ://copainsdavant.linternaute.com/
15. https ://www.trombi.com/
16. https ://www.etnoka.fr/

internautes de diffuser, d'échanger ou de copier toutes sortes de contenus multimédias (*musique, films, textes, etc.*), ce qui peut constituer un acte de contrefaçon lorsque les œuvres sont protégées par le droit d'auteur. Et quant à la diffusion de photos sur son profil, les internautes ignorent souvent que la simple diffusion de l'image d'un ami sans son consentement peut porter atteinte à l'image de cette personne.

L'usurpation d'identité. Il est très facile de créer une fausse identité à partir d'un nom, d'un e-mail et d'une photo. Rien ne prouve que la personne qui se cache derrière cette identité virtuelle est bien celle qu'elle prétend être. Le risque résidant dans le fait que l'abuseur puisse gagner la confiance du réseau et lui soutirer des informations. Et ensuite, dans le fait qu'il peut sérieusement nuire à votre réputation en se livrant à toutes sortes de méfaits sous votre identité[**4**].

1.7 Identité numérique qui laisse des traces

Tout d'abord parce que cela nécessite en premier lieu de faire la liste à peu près complète de tous les services que l'on souhaite quitter. Après quoi, il ne reste plus qu'à se connecter sur chaque site, l'un après l'autre, à se rendre dans son profil, à trouver le lien « supprimer mon compte » et à cliquer dessus. Mais si certains sites tels que Copains d'Avant, del.icio.us, Flickr, Last.fm, LinkedIn, Myspace, Twitter, Viadeo, Youtube ... proposent cette option. Ce n'est pas le cas de tous. Chez la plupart des autres (*Bloglines, Dailymotion, Ma.gnolia, Netvibes, Technorati, Wikio...*), la suppression se fait en envoyant un email au support technique du site concerné (*en anglais donc*) et attendre leur réponse (*plus ou moins rapide*).

Pour ce qui est de Facebook, le site se réserve le droit de conserver l'intégralité de nos données et se contente de « désactiver » votre compte, sans possibilité de le supprimer réellement. Une fois le compte créé, c'est pour la vie. Un cas unique, puisqu'il est notamment possible de réactiver son compte par la suite. Avant de désactiver son compte, il conviendra donc de supprimer méthodiquement tout ce qui peut ressembler à des données personnelles, en ne laissant que le strict minimum exigé par le site (*nom, prénom, date de naissance, email*)[**5**].

1.8 Utilité des réseaux sociaux

Les réseaux sociaux sont utilisés par deux catégories d'individus dans des buts différents :

Tout d'abord par les particuliers, ils utilisent les réseaux sociaux pour partager des informations, des liens, des centres d'intérêts, rechercher des personnes, publier des photos, des vidéos, etc...

Puis par les professionnels qui eux les utilisent pour générer du business, faire connaître leur entreprise, leurs services, suivre les tendances actuelles, optimiser leurs recrutements, etc...

1.9 Principaux avantages des réseaux sociaux

— L'utilisation des réseaux sociaux est tout a fait gratuite

— Simples d'utilisation,

— Améliorent le référencement,

— Effet de buzz,

— Permettent de se faire connaître (*marketing et publicité*),

— Facilitent l'entrée en contact avec vos clients, fournisseurs, partenaires, amis, famille...

1.10 Dangers des réseaux sociaux

Ces dernières années, le Web est devenu le canal d'informations le plus utilisé par les jeunes. Ceux-ci sont de plus en plus nombreux à posséder un appareil mobile leur permettant l'accès à Internet. La mobilité de ces appareils empêche les parents de pouvoir contrôler ce que leurs enfants consultent sur Internet. De plus, les jeunes peuvent facilement avoir accès à des sites au contenu pouvant heurter leur sensibilité et être nocifs pour eux. Internet est un avantage pour ceux qui souhaitent apprendre, améliorer leur culture, découvrir de nouveaux domaines ou communiquer. Cependant, Internet représente une multitude de dangers envers ses usagers et notamment pour les jeunes qui sont les plus vulnérables car ils n'ont pas le recul ni l'expérience leur permettant de discerner une situation à risque ou un contenu potentiellement nuisible, il peut s'agir de sites au contenu

explicite ou de la rencontre de personnes malintentionnées[**6**].

1.11 Conclusion

Dans ce chapitre, nous avons abordées les notions de base des réseaux sociaux ainsi que leurs effets. Les points de vue des utilisateurs est un moyen pour mesurer l'importance d'un service, vis à vis d'un autre. Pour évaluer l'importance d'un service on utilise les avis des internautes. L'objectif de notre étude est collecter et analyser ces opinions, cette tâche sera détaillée dans le prochain chapitre.

Chapitre 2

Etat de l'art sur l'analyse des opinions

2.1 Introduction

La publication croissante sur internet de textes à teneur politique (*lois, rapports, billets de blogs politiques, etc*) et le constat que la politique ne se fait plus seulement dans les hémicycles mais aussi dans les débats en ligne, a conduit certains chercheurs à utiliser les techniques d'analyse d'opinions pour déterminer l'accord ou le désaccord des commentateurs avec telle ou telle proposition de loi ou des candidats . Les acteurs politiques ont également suivi cette tendance, tel qu'avant de promulguer une nouvelle loi, les politiciens essayent de récolter l'avis des internautes sur cette loi. Il est intéressant de connaître aussi l'avis des internautes sur tel homme politique pour une élection présidentielle par exemple l'analyse des médias sociaux a indiqué que le républicain Donald Trump gagnerait les élections américaines, avant que les sondages confirment que ce soit vrai. Ceci selon les données de médias sociaux analysées par BrandsEye, qui a pointé vers une victoire de Trump avant même que les votes aient été jetés[**7**].

2.2 Généralités et Définition

2.2.1 Opinion

L'opinion est un jugement que l'on porte sur un individu, un être vivant, un phénomène, un fait, un objet ou une chose. Elle peut être considérée comme bonne ou mauvaise, tout dépend de la nature de l'individu en fonction de son caractère, ses émotions, son comportement [**8**].

2.2.2 Types d'opinion

Les opinions pouvant être positives, négatives ou neutres, il existe deux catégories principales d'opinion :

- **Opinion usuelle ou comparative**

  Opinion usuelle : est un simple avis qui peut être visé d'une manière directe ou indirecte sur un sujet principal. Les opinions directes sont les plus exploitées dans les études.

  Opinion comparative : est un sous-domaine de l'exploration d'opinion qui traite de l'identification et de l'extraction d'informations exprimées sous une forme comparative.

- **Opinion explicite ou implicite**

  Opinion explicite : est un extrait sans aucun mot d'opinion c.a.d un avis subjectif.
  Opinion implicite : est une clause sans aucun mot d'opinion ces avis sont généralement objectif c'est la catégorie la moins explorée dans les études.[**9**].

2.3 Disciplines en relation avec l'analyse des opinions

2.3.1 Traitement automatique du langage naturel (TALN)

Le traitement automatique du langage nature ou traitement automatique de la langue naturelle est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle, qui concerne l'application de programmes informatiques à tous les aspects du langage humain[**10**]. Il est utilisé pour la traduction automatique, alimenter les moteurs de recherche, filtrer le spam et obtenir des analyses de manière rapide et évolutive.

2.3.2 Machine Learning (ML)

L'apprentissage automatique vise le développement, l'analyse et l'implémentation des méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage[**11**], et ainsi de remplir des tâches qu'il est difficile ou même impossible de les faire

par des moyens algorithmiques classiques.

2.3.3 Deep learning

Le Deep Learning ou apprentissage en profondeur est un sous-domaine du machine learning qui s'intéresse à des algorithmes, inspirés par la structure et la fonction du cerveau humain pour faire des réseaux de neurones artificiels.

2.4 Techniques de classification des opinions

La *figure* 2.1 illustre les différentes approches permettant de classifier des opinions :

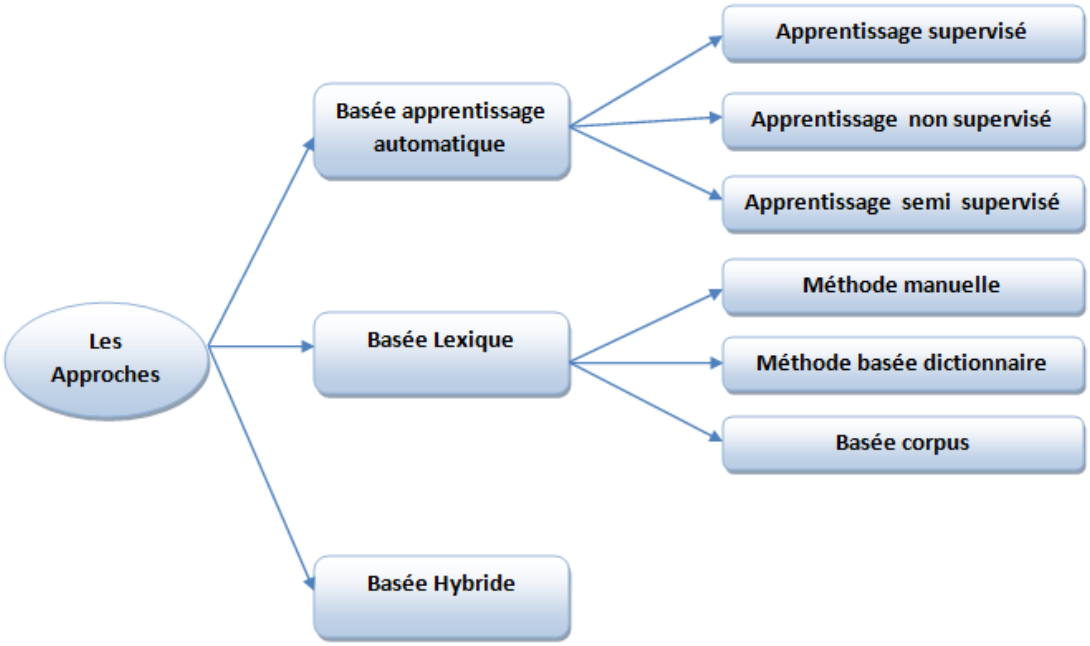

FIGURE 2.1: Méthodes de classifications

2.4.1 Apprentissage automatique

Dans l'apprentissage automatique, les tâches sont généralement classées en grandes catégories. Ces catégories sont basées sur la façon dont l'apprentissage est reçu ou comment le feedback sur l'apprentissage est donné au système développé.

Les méthodes d'apprentissage automatique les plus adoptées sont l'apprentissage supervisé qui forme des algorithmes basés sur des données d'entrée et de sortie étiquetées par le développeur et l'apprentissage non supervisé qui ne fournit pas à l'algorithme des données étiquetées permettre de trouver une structure et de découvrir des données entrées.

Apprentissage supervisé

On appelle apprentissage supervisé la branche de machine learning qui s'intéresse aux problèmes basés sur des labels (*catégories ou classes*) afin de classer de nouvelles données[**11**]. Il existe plusieurs algorithmes et techniques utilisés pour la classification supervisée telles que :

- **Classification naïve Bayésienne**

  C'est une méthode de classification statistique qui base sur le théorème de Bayes [**11**]. Elle est utilisée dans plusieurs applications telles que les applications de détection des Sparns pour séparer les bons courriels des mauvais aussi dans la détection des Risques de crédit.

  Le principe d'un classificateur naïve bayésien consiste à maximiser la probabilité $P(y|d)$, soit la probabilité d'occurrence de la classe de prédiction y connaissant la représentation de la nouvelle donnée x (*on suppose donc ici d = d(x) = (d1, d2, ..., dn)*), et ce pour toutes les classes y ∈ Y et toutes les composantes qui interviennent dans la définition de l'espace de représentation D.

  Pour cela, on fait appel à la règle de Bayes. Soient A et B deux évènements. La règle de Bayes dit alors que la probabilité de l'évènement A sachant l'évènement $B(P(A|B))$ peut se calculer à l'aide des probabilités des évènements A et B (P(A) et P(B)) et connaissant la probabilité de l'évènement B sachant l'évènement $A(P(B|A))$ par la formule 2.1 :

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

  En appliquant la règle de Bayes à la problématique de la classification, on obtient l'équation 2.2 :

$$P(y|d) = P(d|y) * P(y)/P(d) \tag{2.2}$$

Les probabilités de l'expression de droite doivent être estimées, à l'aide du corpus d'apprentissage S, afin de calculer la quantité qui nous intéresse, soit $P(y|d)$:

— $P(y)$ est la probabilité d'observer la classe y.

— $P(d)$ est la probabilité d'observer la représentation d.

— $P(d|y)$ la vraisemblance de l'évènement « observer la représentation d » si s∈ S est de classe y. Ce terme est plus difficile à estimer que le précédent.

En pratique, on ne s'intéresse qu'au numérateur, le numérateur ne dépendant pas de y. Concernant $P(d|y)$, l'hypothèse habituellement faite dans ce type de classifieurs est que toutes les composantes *di* sont indépendantes, ce qui permet de calculer facilement la probabilité globale d'une classe connaissant une donnée. Cette non-dépendance des composantes correspond à « l'hypothèse de Bayes naïve ».

On considère donc que :

$$P(d|y) = \prod iP(di|y) \tag{2.3}$$

Maximiser $P(y|d)$ revient donc à maximiser $(\prod iP(di|y))P(y)$.

Les $P(di|y)$ sont évalués par les fréquences observées dans les exemples de l'ensemble S.

L'inconvenient majeure de cette méthode est :

L'algorithme Naive Bayes Classifier suppose l'indépendance des variables ce qui implique que chaque fonctionnalité soit indépendante.

- **Arbre de décision**

L'apprentissage se fait par partitionnement récursif selon des règles sur les variables explicatives. Suivant les critères de partitionnement et les données, on dispose de différentes méthodes, dont CART, CHAID ... Ces méthodes peuvent s'appliquer à

une variable à expliquer qualitative ou quantitative [**11**].

La *figure* 2.2 illustre un arbre de décision ayant trois Features temps, humidité, vent :

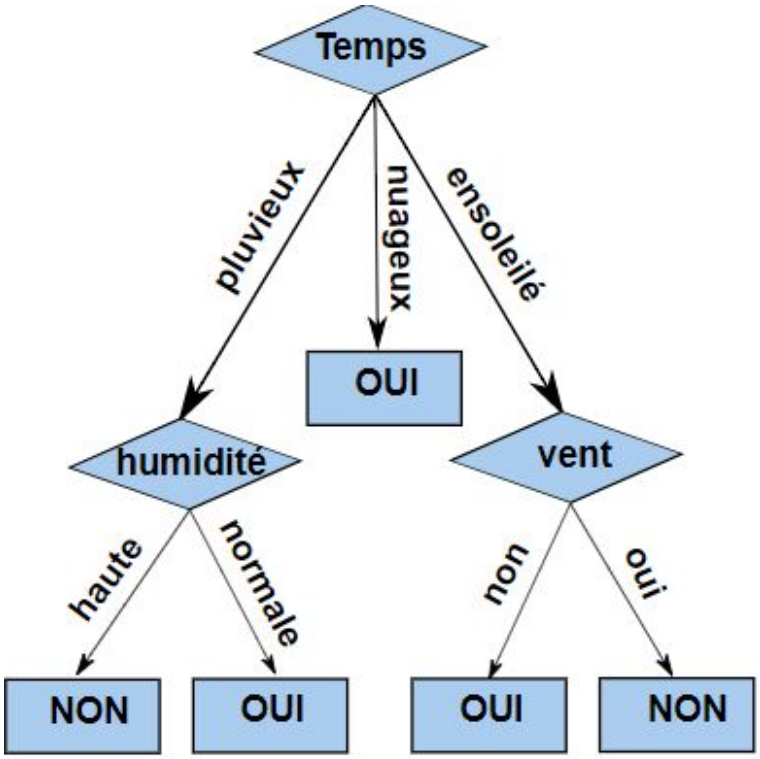

FIGURE 2.2: Exemple d'un arbre de décision[**12**]

L'incovénient majeure de cette approche c'est :

Les arbres de décision sont instables et ont tendance à overfiter.

Réseaux de neurone

En anglais neural network est un ensemble de neurones formels interconnectés permettant la résolution de problèmes.

Principe de fonctionnement Le neurone reçoit les entrées $x1, \ldots, xn$. Le potentiel d'activation du neurone N est défini comme la somme pondérée (*les poids sont les coefficients synaptiques*)wi des entrées. La sortie S est calculée en fonction du seuil θ.

$$N = x.w = x1.w1 + \ldots + xi.wi + \ldots + xn.wn \tag{2.4}$$

Alors : $S = 1 \quad si \quad N > \theta$

$S = 0 \quad si \quad N \leq \theta$

Applications

- Statistiques : analyse de données / prévision / classification .
- Robotique : contrôle et guidage de robots ou de véhicules autonomes.
- Imagerie / reconnaissance de formes.

La *figure* présente un réseau de neurones dotés de 1 jusqu'à n entrées et une fonction d'agrégation qui fait la somme pondérées des entrées ensuite la faire passer par une fonction d'activation qui détermine la sortie selon un seuil bien précis :

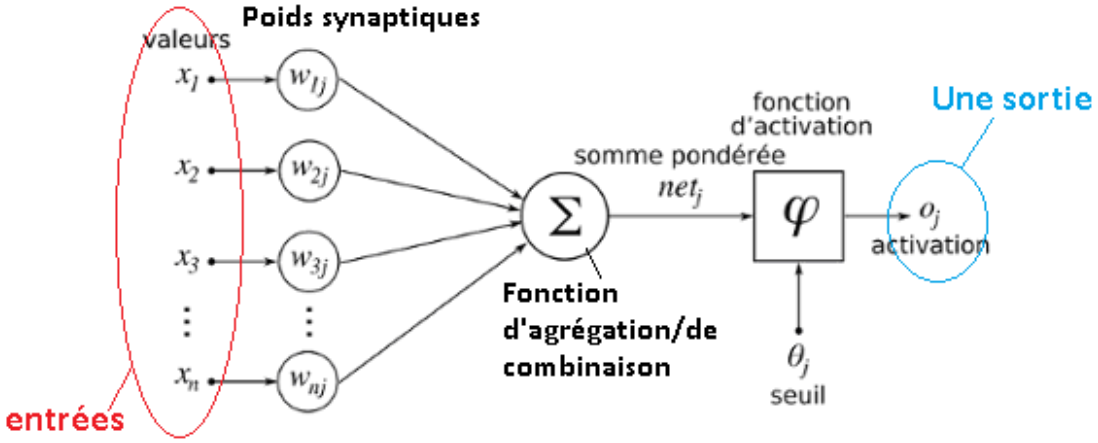

FIGURE 2.3: Exemple d'un neurone[**13**]

L'inconvénient majeure de cette approche c'est :

Le choix des valeurs initiales des poids du réseau est difficile.

Boosting

Il s'agit d'une méthode de classification émettant des hypothèses qui sont au départ de moindre importance. Plus une hypothèse est vérifiée, plus son indice de confiance augmente. Ce qui prend de l'importance dans la classification [**14**].

Apprentissage non supervisé

L'apprentissage non-supervisé est un problème d'apprentissage automatique. Il s'agit, pour un logiciel, de trouver des structures sous-jacentes à partir de données non étiquetées. Puisque les données ne sont pas étiquetées, il n'est pas possible d'affecter au résultat de l'algorithme utilisé un score d'adéquation. Cette absence d'étiquetage (*ou d'annotation*) est ce qui distingue les tâches d'apprentissage non-supervisé des tâches d'apprentissage supervisé. L'introduction dans un système d'une approche d'apprentissage non supervisé est un moyen d'expérimenter l'intelligence artificielle, les algorithmes d'apprentissage non supervisé peuvent exécuter des tâches de traitement plus complexes que les systèmes d'apprentissage supervisé, mais ils peuvent aussi être plus imprévisibles.

Le processus **Clustering** est la forme la plus répandue d'apprentissage non supervisé :

— Regroupement d'instances ayant des traits communs.
— Utile pour identifier des tendances dans les données.
— Et pour dégager des thèmes communs dans des documents.

- **Algorithme k-Means**

  Dans l'algorithme k-Means[**16**], nous avons des instances à regrouper :

  — k est le nombre de groupe que l'on veut créer (*clusters*).
  — On adopte une structure linéaire Æ d'arbre.
  — Idée générale : on définit un groupe par son centre de masse.

  Approche

  - Assigner arbitrairement chaque instance à un des k groupes.
    — Peut-être fait aléatoirement.
  - Calculer la moyenne de chaque groupe (*centroïde*).
    — Pourrait être aussi la médiane :

$$\mu = \frac{1}{M} \sum_{x \in C} x \quad où \quad M = |C| \tag{2.5}$$

  - Déplacer chaque instance dans le groupe dont la moyenne est la plus près :

$$c_j = \{x_i | \forall \mu_l \ , dist(x_i, \mu_j) \leq dist(x_i, \mu_l)\} \tag{2.6}$$

- Recommencer les étapes 2 et 3 quelques fois.
  - On arrête lorsque les déplacements d'instances se stabilisent.

Exemple de k-Means

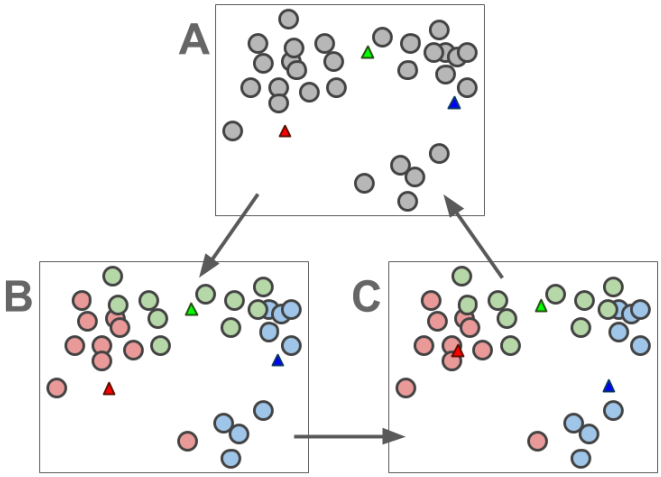

FIGURE 2.4: Exemple de k-Means[**17**]

A : positionnement des centres (*aléatoirement la première fois*).
B : affectation de chaque élément au centre le plus proche, selon la mesure de similarité choisie.
C : repositionnement des centres au barycentre des groupes ainsi constitués.
Puis on recommence jusqu'à la stabilisation de la position des centres.

Les **inconvénients** de cette approche sont :

- Le Manque de cohérence,
- Ensemble non optimal de clusters,
- Limitation des calculs,
- Traiter que les données numériques.

- **Algorithme des c-moyennes flous *Fuzzy K-Means***

  La variante floue de la méthode des k-moyennes, proposant qu'un objet ne soit pas associé qu' à un seul groupe[**18**].

  Les étapes de cet algorithme sont comme suit :

  Entrées : k est le nombre de clusters désiré, d'une mesure de dissimilarité sur l'ensemble X des objets à traiter, T est le nombre maximum d'itérations, un poids $m > 1$ et un seuil $\epsilon > 0$.

  Sortie : Une partition floue $C = C_1, ..., C_k$ définie par les fonctions. d'appartenance $(u_h)_h = 1...k$

  1. A l'instant (t=0) Choisir ou tirer aléatoirement une partition initiale $u_{(h,t)_h} = 1...k$.

  2. Calculer les centres de gravité$x^*_{1,t}, ..., x^*_{k,t}$ de chacune des k classes (*à l'instant t*)

$$\frac{1}{\sum_{x_i \in X}[u_{h,t}(x_i)]^m} \cdot \sum_{x_i \in X}[u_{h,t}(x_i)]^m . v_j(x_i) \tag{2.7}$$

  3. Calculer les nouvelles valeurs d'appartenance$(u_{h,t+1}(x_i))_h = 1...k$ de chaque objet x_i à chaque centre de classe $x^*_{h,t}$

$$u_{h,t+1}(x_i) = \frac{[d(x^i, x^*_{h,t})]^{\frac{2}{1-m}}}{\sum_{h=1}^{k}[d(x_i, x^*_{h,t})]^{\frac{2}{1-m}}} \tag{2.8}$$

  4. Calculer les centres de gravité de chaque classe $x^1_{1,t+1}, ..., x^*_{k,t+1}$ de chacune des k classes à l'instant t + 1.

  5. Calculer le déplacement global

$$E_t = \sum_{h=1}^{k} d(x^*_{h,t+1}, u_{h,t}) \tag{2.9}$$

  6. Si $E_t \preceq \epsilon$ alors retourner la partition floue définie par$((x^*_{h,t+1}, u_{h,t+1}))_h = 1, ..., k))$, sinon (*(t=t+1) retourner en 3*).

  Les **inconvénients** de cette approche :

  — L'algorithme des c-moyennes flous n'est pas robuste face aux bruits introduits par l'imprécision des attributs,

  — Son efficacité dépend fortement de l'étape d'initialisation des classes,

  — Utilisable que pour détecter des classes de forme sphérique.

- **Méthodes hiérarchiques**

  Les méthodes hiérarchiques[**19**] consistent à effectuer une suite de regroupements en Clusters de moins en moins fines en agrégeant à chaque étape les objets (*simple élément*) ou les groupes d'objets (*un Cluster-partition*) les plus proches. Ce qui nous donne une arborescence de clusters. Cette approche utilise la mesure de similarité pour refléter l'homogénéité ou l'hétérogénéité des classes.

  Son principe est simple, initialement chaque individu forme une classe, soit n classes. Donc on cherche à réduire ce nombre de classe newnbrclss < n itérativement de sorte que dans chaque étape on fusionne deux classes ensemble (*Les deux classes choisies pour être fusionnées sont celles qui sont les plus "proches" en fonction de leur dissimilarité*) ou ajouter un nouveau élément à une classe (*un élément appartient à une classe s'il est plus proche de cette classe que de toutes les autres*) La valeur de dissimilarité est appelée indice d'agrégation. Il commence dans la première itération faible, et croîtra d'itération en itération.

- **Espérance Maximisation *EM***

  L'Espérance Maximisation EM[**20**] est un algorithme itératif qui permet de trouver les paramètres du maximum de vraisemblance [1] d'un modèle probabiliste lorsque ce dernier dépend de variables latentes non observables. De nombreuses variantes ont par la suite été proposées, formant une classe entière d'algorithmes.

  On utilise souvent l'algorithme EM pour la classification de données, l'apprentissage automatique, ou la vision artificielle. On peut également citer son utilisation en imagerie médicale dans le cadre de la reconstruction tomographique.

  L'algorithme d'espérance-maximisation comporte :

  — Une étape d'évaluation de l'espérance E, où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées.

  — Une étape de maximisation M, où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

  On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi.

1. Caractère de ce qui semble vrai, juste, aux yeux du sens commun.

Les **inconvénients** de cette approche :

— A moins de bénéficier de conditions spéciales, il faut recommencer la simulation a chaque itération,

— Besoin de grandes valeurs de m pour obtenir la stabilité.

Les applications de l'algorithme EM

- **Traitement du signal**
  Le traitement du signal c'est la réalisation d'opérations sur le signal.
- **Applications du traitement du signal**
  — Elaboration de signaux : Synthèse de parole ou de musique, modulation et codage,
  — Interprétation des signaux : filtrage, extraction, identification, analyse (*spectrale ou temporelle*) ou mesure,
  — Mixage : utilisation de plusieurs signaux (*audio la plupart du temps*) pour la diffusion d'un ou deux signaux résultats[**21**].
- **Traitement d'image notamment en imagerie médicale**
  Le traitement d'images est une discipline de l'informatique et des mathématiques appliquées qui étudie les images numériques et leurs transformations, dans le but d'améliorer leur qualité ou d'en extraire de l'information[**22**].
- **Applications du traitement d'image**
  — Reconstruction d'un objet d'après plusieurs images,
  — Calcul de mouvements entre des séries d'images,
  — Détecter la présence d'un objet ou son absence.

Apprentissage semi_supervisé

L'apprentissage semi-supervisé est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non étiquetées [**14**]. Il est entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées se type permet d'améliorer significativement la qualité de l'apprentissage.

2.4.2 Approche basée lexique

L'approche basée lexique utilise un lexique des sentiments pour décrire la polarité (*positive, négative et neutre*) du contenu d'un texte. Cette approche est plus compréhensible et peut-être facilement mise en œuvre contrairement aux algorithmes de base de l'apprentissage automatique. Mais l'inconvénient est que ça nécessite l'implication des êtres humains dans le processus de l'analyse du texte. Plus le volume d'information est important, plus l'effort humain fournit sera remarquable notamment pour le nettoyage des données, l'identification du sentiment et la distinction des données utiles provenant de diverses sources de contenu.

L'approche basée sur le lexique inclut trois catégories :

- **Approche basée sur le dictionnaire**
  Un petit ensemble de mots d'opinion est collecté manuellement avec des orientations connues [**23**]. Cet ensemble est développé en recherchant dans les corpus bien connus WordNet [2] ou dans le dictionnaire des synonymes et antonymes. Les mots récemment trouvés sont ajoutés à la liste de semences, puis la prochaine itération commence. Le processus itératifs arrête lorsqu'aucun nouveau mot n'est trouvé. Une fois le processus terminé, une inspection manuelle peut être exécutée pour supprimer ou corriger les erreurs.
- **Approche basée sur le corpus**
  Un corpus est un grand corps de texte en langage naturel utilisé pour accumuler des statistiques sur le texte en langage naturel. Les corpus incluent souvent des informations supplémentaires comme une étiquette pour chaque mot indiquant sa partie de discours, et peut-être l'arbre d'analyse pour chaque phrase. Un lexique est une collection d'informations sur les mots d'une langue à propos des catégories lexicales auxquelles ils appartiennent. Un lexique est généralement structuré comme une collection d'entrées lexicales. Une entrée lexicale inclura d'autres informations sur les rôles joués par le mot, tels que les informations sur les caractéristiques (*par exemple, si un verbe est transitif, intransitif etc, quelle forme prend le verbe : participe, présent, passé, etc...*)

2. http ://wordnetweb.princeton.edu/perl/webwn ?fbclid=IwAR1qPFWxaz9bFxRgQg9Xf d3TWoIopAtoRxLtp4tnkI2avsZcM7Eh4tbQLxg.

- **Méthode manuelle**

  Utilisée en combinaison avec des approches automatisées [**24**] telles que l'approche basée sur un dictionnaire et sur le corpus, l'annotation de sentiment manuel prend beaucoup de temps.

2.4.3 Approche hybride

Cette approche est une combinaison entre l'approche basée lexique et l'approche basée apprentissage automatique [**14**]. L'utilisation de l'approche hybride permet d'annoter automatiquement le corpus[**23**]d'apprentissage avec la méthode basée lexique, et ensuite entrainer le classificateur sur ce corpus avec une méthode issue des méthodes de l'apprentissage automatique.

2.5 Avantages et limites des approches

Le tableau 2.1 illustre les avantages et limites des deux approches Apprentissage automatique et l'approche basée sur le lexique :

Approche	**Avantages**	**Limites**
L'apprentissage Automatique	-Dictionnaire n'est pas nécessaire -Meilleurs résultats, haute précision de classification.	-Les classificateurs s'entrainent sur un domaine. -Spécifique.
lexique	-Aucune donnée d'entrainement. -Moins d'opérations de calcul.	-Exige des ressources linguistiques puissantes.

TABLE 2.1: Avantages et limites des approches

2.6 Conclusion

Nous avons eu un aperçu sur les principes de des grandes approches/outils d'analyse des opinions, les méthodes de classification, de clustering, ainsi que toutes les autres Méthodes. Elles ont leurs avantages, faiblesses et limites. Dans le chapitre suivant, nous

allons proposer une nouvelle méthode de classification permettant de prédire des connaissances et bien évidement la comparer aux méthodes existantes.

Chapitre 3

Nouvelle solution : Conception, mise en œuvre

3.1 Introduction

Twitter [1] donne au grand public un accès direct non filtré aux idées et aux opinions politiques. Cela signifie que la compréhension et l'analyse du contenu des tweets [2] se trouvant dans ce réseau social, peut nous aider à comprendre les différents avis à propos des politiciens. Cet ensemble de données est destiné à faciliter cette exploitation. Dans ce chapitre, nous allons proposer une amélioration d'une méthode existante, dans le but d'augmenter la précision de notre model de classification.

3.2 Généralités sur twitter

En mars 2006 Jack Dorsey, Noah Glass,Biz Stone et Evan Williams [**25**] ont développé Twitter, ce réseau social en ligne est devenu rapidement populaire dans le monde, les messages sont appelés tweets, qui sont en toutes langues et limités à 280 caractères.

3.3 Caractéristiques du Tweet

Le tweet est composé de :

1. https ://twitter.com/
2. Court message informatif posté sur le réseau social Twitter par l'intermédiaire d'un service qui le transmet à des abonnés.

— Id tweet,

— Nom de l'utilisateur : indiqué par ' @ ',

— Le texte du tweet : ensemble d'information / opinions plus des émoticônes *emojis*,

— La photo de profile (*une photo personnelle ou une image*),

— L'emplacement de l'utilisation (*localisation*),

— Date du tweet,

— Hashtag : c'est un mot ou un ensemble de mots précédé par le symbole '#' Comme on a la possibilité de aimer, Mentionner une autre personne ou page, Retweeter, répondre à un Tweet.

3.4 Dataset

Le dataset "Democrat Vs. Republican Tweets " : est un dataset de Kaggle[3] qui contient 86461 tweets, qui sont collectés à partir d'un API Twitter où il inclut 3 champs : Party (*Républicain /Démocrate*), Handle (*info*), Tweet (*le texte/message*).

Ces données sont répartis en deux classes (*Républicain / Démocrate*) dont 44393 tweets pour Républicain(*51%*) et 42068 tweets pour Démocrate(*49%*).
Dans chaque partie (*Républicain / Démocrate*) on a un ensemble de mot qui construit les tweets, pour cela on a fait un petit traitement :

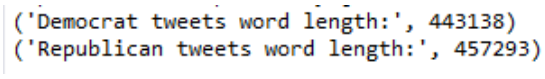

FIGURE 3.1: Nombre de mots dans chaque partie

3. https ://www.kaggle.com/

```
from nltk.probability import FreqDist
fdist_democrat = FreqDist(democrat_tweets)
fdist_republican=FreqDist(republican_tweets)
fdist_republican
import matplotlib.pyplot as plt
plt.subplots(figsize=(10,5))
fdist_democrat.plot(30,title="Democrat Tweets")
plt.subplots(figsize=(10,5))
fdist_republican.plot(30,title="Republican Tweets")
de=pd.DataFrame(list(fdist_democrat.items()), columns = ["Word","FrequencyDemocrat"])
re=pd.DataFrame(list(fdist_republican.items()), columns = ["Word","FrequencyRepublican"])
```

FIGURE 3.2: Code pour visualiser les mots les plus fréquents dans la partie démocrate

La partie démocrate : La figure 3.3 montre les mots les plus fréquents de la partie démocrate

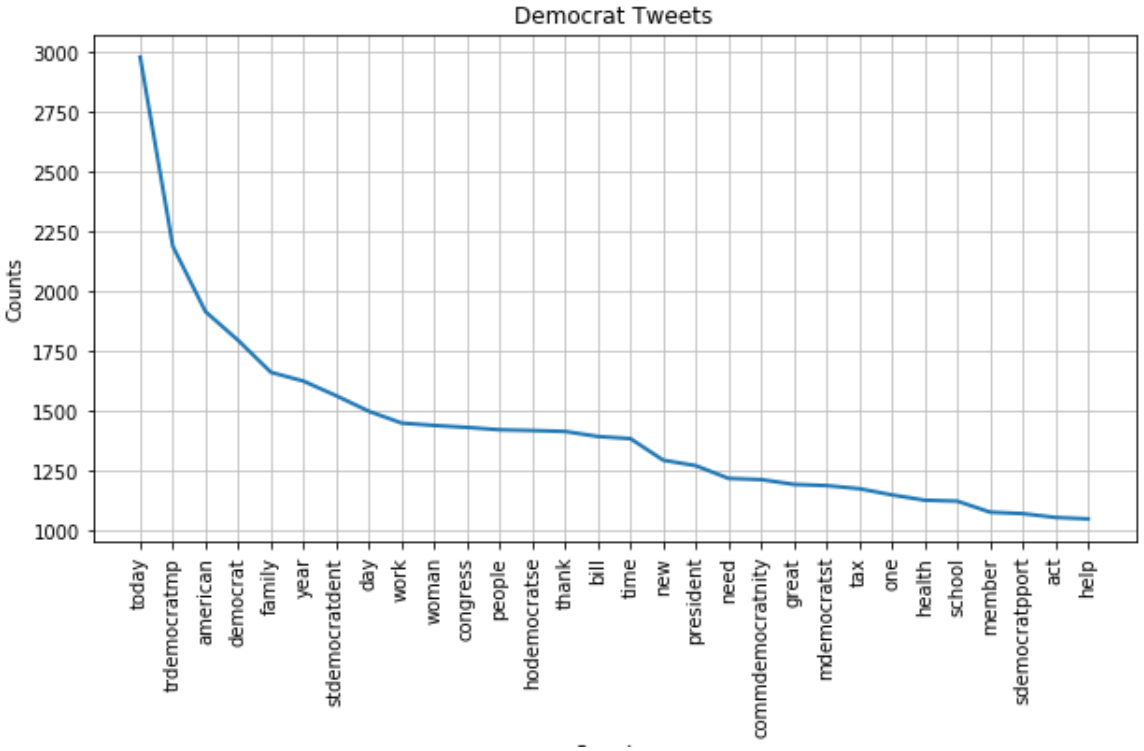

FIGURE 3.3: Courbe qui présente les mots les plus fréquents de la partie démocrate

La partie république : La figure 3.4 montre les mots les plus fréquents de la partie républicaine :

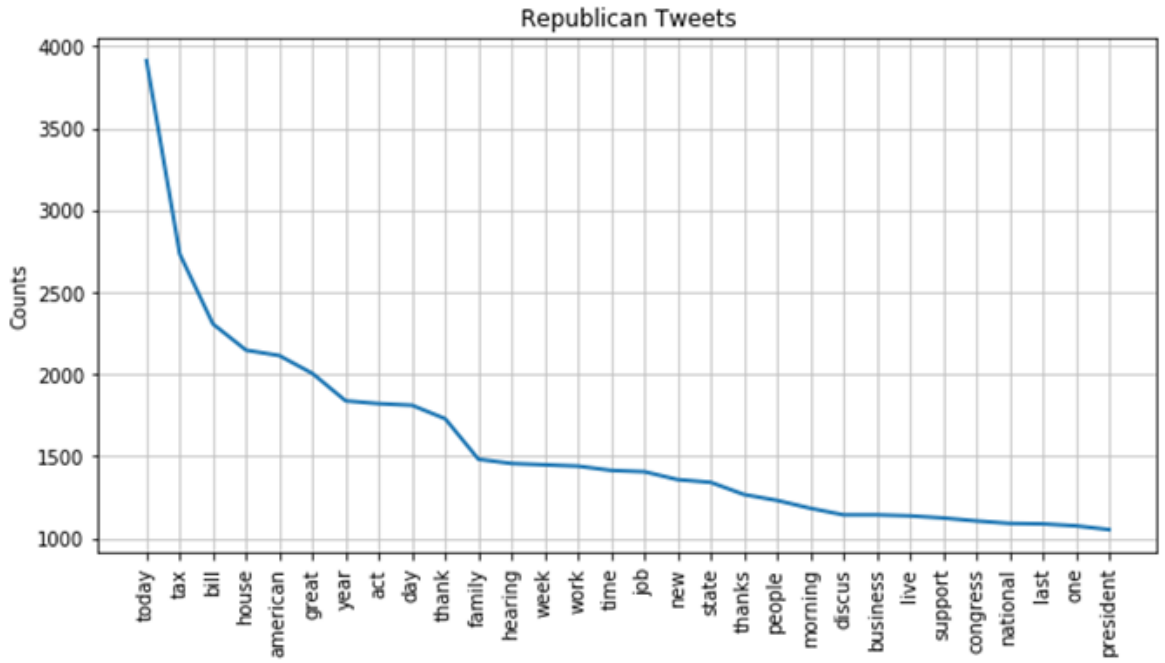

FIGURE 3.4: Courbe qui présente les mots les plus fréquents dans la partie républicaine

3.5 Prétraitement des données

Le prétraitement et le nettoyage des données sont des tâches importantes qui doivent intervenir avant d'utiliser un jeu de données pour la formation de modèles. Les données brutes sont souvent bruyantes, peu fiables et incomplètes, pour cela nous avons nettoyés nos données dans cet ordre.

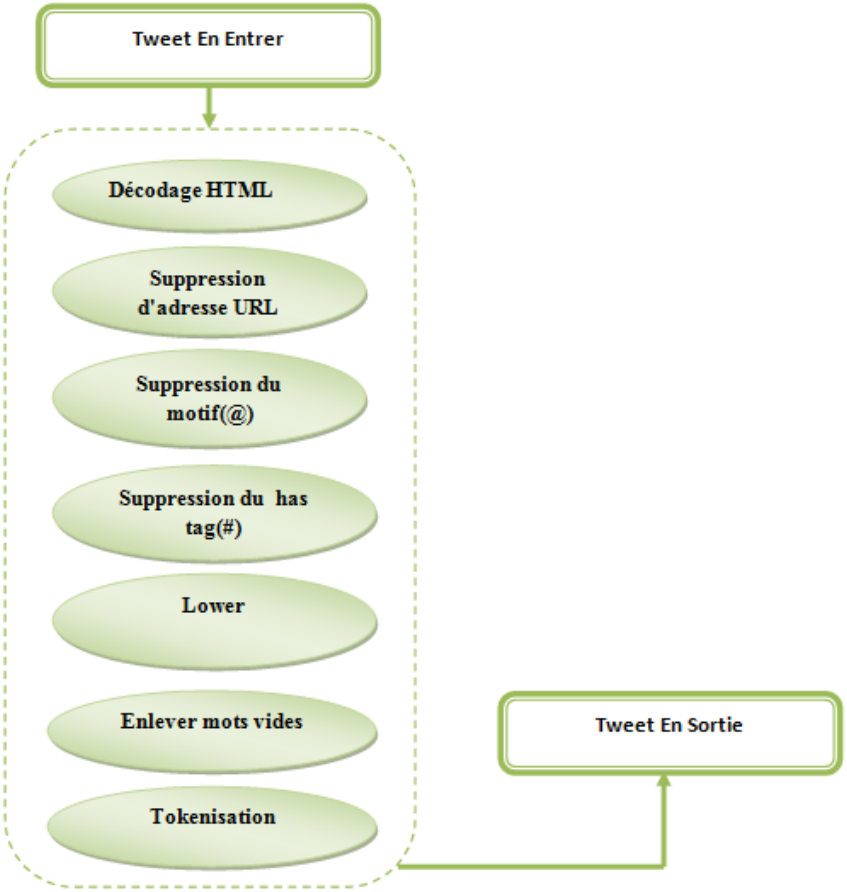

FIGURE 3.5: Prétraitement des Tweets

Décodage HTML : Les balises HTML qui ne se convertissent pas en texte et ils apparaissent sous la forme " &", "&quot".

```
In [42]: runfile('C:/Users/acer/Desktop/Memoire_M2/ytgfyt.py', wdir='C:/Users/acer/Desktop/
Memoire_M2')
Recently met @floridastate students Saphicher Gonzalez &amp; Elena Princi in #Tallahassee. As a
@FSUSGA senator, Saphic… https://t.co/fg3NGed9xH
tweet after cleaning:
recently met student saphicher gonzalez elena princi tallahassee senator saphic http co fg nged
```

FIGURE 3.6: Balise HTML

Suppression d'adresse URL : Motif(http...)

```
In [43]: runfile('C:/Users/acer/Desktop/Memoire_M2/ytgfyt.py', wdir='C:/Users/acer/Desktop/
Memoire_M2')
Recently met @floridastate students Saphicher Gonzalez &amp;amp; Elena Princi in #Tallahassee. As a
@FSUSGA senator, Saphic… https://t.co/fg3NGed9xH
tweet after cleaning:
recently met student saphicher gonzalez elena princi tallahassee senator saphic
```

FIGURE 3.7: Enlever l'URL

Suppression du motif(@) : Le code ci-dessous permet d'enlever le symbole @ des tweets .

```
In [38]: runfile('C:/Users/acer/Desktop/Memoire_M2/ytgfyt.py', wdir='C:/Users/acer/Desktop/
Memoire_M2')
RT @garywhite13: @SenBillNelson, @RepDarrenSoto discuss restoration of voting rights for
felons, census questions during town hall in Haine…
tweet after cleaning:
discus restoration voting right felon census question town hall haine
```

FIGURE 3.8: Enlever '@'

Suppression du hashtag(#) : Le code ci-dessous permet de supprimer le symbole # des tweets.

```
In [39]: runfile('C:/Users/acer/Desktop/Memoire_M2/ytgfyt.py', wdir='C:/Users/acer/Desktop/
Memoire_M2')
It was great to host a #NationalDayOfPrayer breakfast in #WinterSprings to discuss the need
for unity and the impor… https://t.co/mIJRRFrHHn
tweet after cleaning:
great host nationaldayofprayer breakfast wintersprings discus need unity impor http co
```

FIGURE 3.9: Enlever '#'

Lower : Utilisé pour transformer le texte en minuscule afin de faciliter à l'algorithme l'appariement lorsqu'il compare entre le tweet du test et ceux du trainset.

```
In [5]: runfile('C:/Users/acer/Desktop/Memoire_M2/ytgfyt.py', wdir='C:/Users/acer/Desktop/
Memoire_M2')
Today at 5:30PM PT, the @LosAngelesVA will be holding a public hearing on the VA Greater Los
Angeles Healthcare Sys… https://t.co/Pszdvl9koM
tweet after cleaning:
today pm pt holding public hearing va greater los angeles healthcare sys http co pszdvl kom
```

FIGURE 3.10: Transformer en minuscule.

Tokenisation : Enlever mots vides.

```
In [39]: runfile('C:/Users/acer/Desktop/Memoire_M2/ytgfyt.py', wdir='C:/Users/acer/Desktop/
Memoire_M2')
It was great to host a #NationalDayOfPrayer breakfast in #WinterSprings to discuss the need
for unity and the impor… https://t.co/mIJRRFrHHn
tweet after cleaning:
great host nationaldayofprayer breakfast wintersprings discus need unity impor http co
```

FIGURE 3.11: Enlever les mots vides

```
In [41]: runfile('C:/Users/acer/Desktop/Memoire_M2/ytgfyt.py', wdir='C:/Users/acer/Desktop/
Memoire_M2')
It was great to host a #NationalDayOfPrayer breakfast in #WinterSprings to discuss the need
for unity and the impor… https://t.co/mIJRRFrHHn
tweet after cleaning:
['great', 'host', 'nationaldayofprayer', 'breakfast', 'wintersprings', 'discus', 'need',
'unity', 'impor']
```

FIGURE 3.12: Enlever les mots vides

3.6 Méthode proposée

La méthode proposée est inspirée de la méthode Naive Bayes[**11**] qui se base sur le théorème de Bayes. Ce dernier est un classique de la théorie des probabilités. Ce théorème est fondé sur les probabilités conditionnelles(*voir la section 2.5.1*) :

$P(A|B)$: la probabilité que l'événement A se réalise sachant que l'événement B s'est déjà réalisé.

Dans notre étude, A c'est le mot appartenant à un tweet et B c'est la classe (*Démocrate ou Républicain*).

La matrice TF*IDF qui est une méthode de pondération utilisée en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document qui est un Tweet dans notre cas, relativement à une collection qui est le dataset dans notre étude. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le dataset.

La formule pour déterminer le Term Frequency est la suivante :

$$TF(i) = \frac{\log(Freq(i,j)) + 1}{\log(l)} \quad \textbf{[26]} \tag{3.1}$$

i : Terme dont le Term Frequency dans le document doit être déterminé.

j : Document analysé.

L : Nombre total de mots dans le document « j ».

La formule pour déterminer l'Inverse Document Frequency est la suivante :

$$IDF(i) = \log(\frac{N_D}{f_i} + 1) \quad \textbf{[26]} \tag{3.2}$$

N_D : Nombre de tous les documents dans le corpus des documents (*qui contiennent les termes pertinents*).

f_i : Nombre de tous les documents dans lesquels le terme « i » apparaît. La matrice TF*IDF : est calculée pour chaque mot i dans le document j comme suit :

$$TF(i,j) = TF_{i,j} * IDF_i \tag{3.3}$$

Cette méthode nous a permis de saisir que l'utilisation du log lors du calcul de la fréquence est une bonne idée, afin d'augmenter les petites valeurs.

Architecture de la méthode proposée : Pour la réaliser, nous avons suivi les étapes suivantes :

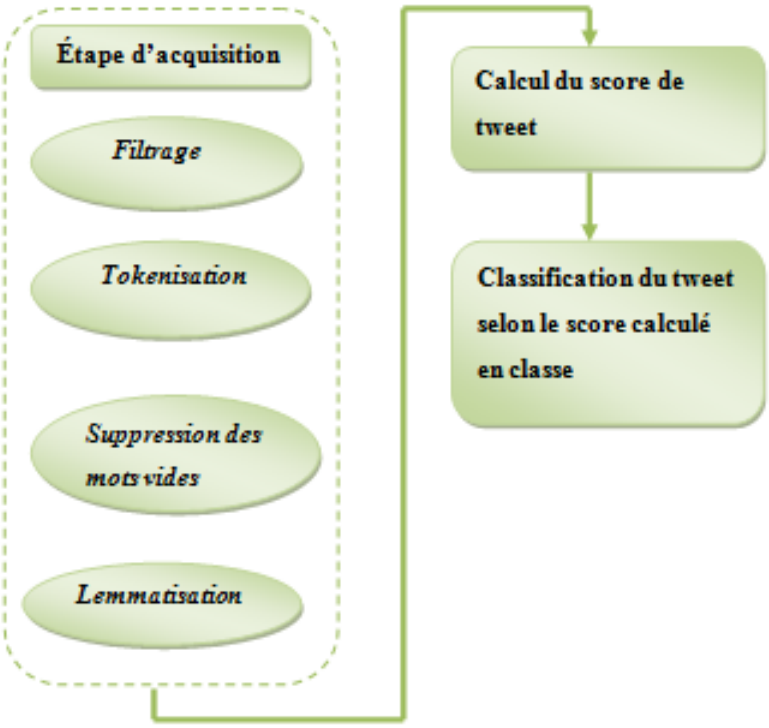

FIGURE 3.13: Architecture de notre proposition

Notre nouvelle méthode consiste à calculer le score du tweet par la somme des scores des mots de ce dernier comme suit :

$$Score(tweet) = \sum_{1}^{n} score(mot) \tag{3.4}$$

n : nombre de mots dans le tweet.

La nouveauté de cette méthode se résume dans le calcul du score du mot :

$$score(mot) = \frac{freq(mot)}{len(classe)} * [-\log(1 - P(mot))] \tag{3.5}$$

freq(mot) : c'est la fréquence du mot dans la classe (*républicain /démocrate*).

len(classe) : c'est le nombre de mots dans la classe (*républicain /démocrate*).

P(mot) : la probabilité du mot dans la classe (*républicain /démocrate*).

P(mot) = fréquence du mot dans la classe / le nombre total de mots dans la classe.

$$P(mot) = \frac{freq(mot)}{len(classe)} \tag{3.6}$$

On peut écrire la formule autrement :

$$score(mot) = P(mot) * [-\log(1 - P(mot))] \tag{3.7}$$

3.6.1 Aspect mathématique de la formule proposée par rapport à la méthode Naive Bayes

Nous allons considérer A comme le mot et B comme la classe :

La méthode Naive Bayes : théoréme de Bayes

$$P(B|A) * \frac{P(A)}{P(B)} = \frac{P(A \cap B)}{P(A)} * \frac{P(A)}{P(B)} \quad , \quad P(A \cap B) = P(A) * P(B)$$

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

$$= P(A)$$

Donc :

$$0 < P(A) < 1 \tag{3.8}$$

La score du mot est cerné entre 0 et 1 selon la méthode Naive Bayes.

La méthode proposée

$$P(A) * [- * \log(1 - P(A))]$$

$$\begin{aligned} 0 < P(A) < 1 \quad &\Longrightarrow 0 < 1 - P(A) < 1 \\ &\Longrightarrow P(A) * [- * \log(1 - P(A))] > 0 \\ &\Longrightarrow \log(1 - P(A)) < 0 \\ &\Longrightarrow - * \log(1 - P(A)) > 0 \\ &\Longrightarrow P(A) * [- * \log(1 - P(A))] > 0 \end{aligned} \tag{3.9}$$

Donc selon la formule proposée, le score du mot sera toujours supérieur à 0, ce qui permet aux résultats d'être considérablement meilleurs que la score de Naive Bayes qui est limité entre 1 et 0.

3.6.2 Implémentation de la solution proposée

Dans cette partie, nous présentons l'implémentation de notre proposition :
La mise en œuvre de notre méthode proposée nécessite comme entrées : un dataset qui contient des tweets politiques représentés en deux parties (*républicain /démocrate*), puis calculer le score d'un tweet ne figurant pas dans le trainset [4], selon ce dernier notre système sera capable d'extraire la classe de ce tweet.

3.6.3 Séparation de trainset et testset

Commençons par diviser notre dataset en test et train :

```
all_tweets = pd.read_csv("ExtractedTweets.csv")
tweeters = all_tweets.iloc[:,:2].drop_duplicates()
handles_train, handles_test = train_test_split(tweeters.Handle, stratify=tweeters.Party, test_size=0.3, random_state=0)
train = all_tweets[all_tweets.Handle.isin(handles_train)].reset_index().drop('index', axis=1)
test = all_tweets[all_tweets.Handle.isin(handles_test)].reset_index().drop('index', axis=1)
```

FIGURE 3.14: Trainset et testset

4. représente les données d'apprentissage sur les quelles l'algorithme ML s'entraine pour pouvoir prédire.

3.6.4 Fonction de classification

Nous avons programmé une fonction qui calcule le score de chaque mot du tweet, ainsi que la somme des scores de ces derniers, pour obtenir le score total du tweet et comme sortie, elle renvoie sa classe (*républicain /démocrate*).

```
def classify(given_tweet):
    kh= Nettoyer (given_tweet)
    ph=[word for word in kh if not word in stopwords]
    h=0
    w=0
    frequenceD={}
    for word in ph :
        for mot in democrat_tweets:
            if word ==mot :
                h=h+1
        frequenceD[word]=h
    frequenceP={}
    for word in ph :
        for mot in republican_tweets:
            if word ==mot :
                w=w+1
        frequenceP[word]=w
    ProbabiliteD={}
    Pdm=0
    ProbabiliteDemTweet=1
    ProbabiliteRepTweet=1
    ProbabiliteDemTweettotal=0
    ProbabiliteRepTweettotal=0
    for word in ph:
         for key , value in frequenceD.items():
               if word== key :
                 Pdm=float(float(frequenceD[word])/float(democrat))
                 ProbabiliteD[word]=Pdm
                 ProbabiliteDemTweet=float(np.log(1-frequenceD[word]/float(democrat)))

                 ProbabiliteDemTweet=-(float(float(ProbabiliteDemTweet)*float(float(ProbabiliteD[word]))))
         ProbabiliteDemTweettotal=ProbabiliteDemTweettotal+ProbabiliteDemTweet
    ProbabiliteP={}
    Pr=0
```

FIGURE 3.15: Code de la fonction de classification

```
for word in ph :
    for key , value in frequenceP.items():
        if word== key :
            Pr=float(float( frequenceP[word])/float(republique))

            ProbabiliteP[word]=Pr
            ProbabiliteRepTweet=float(np.log(1-frequenceP[word]/float(republique)))

            ProbabiliteRepTweet=-(float(float(ProbabiliteDemTweet)*float(frequenceP[word])))
    ProbabiliteRepTweettotal=ProbabiliteRepTweettotal+ProbabiliteRepTweet
if ProbabiliteDemTweettotal >= ProbabiliteRepTweettotal :
    predict="Democrat"
else :
    predict="Republican"
return predict
```

FIGURE 3.16: La suite du code de la fonction de classification

3.7 Conclusion

Nous avons présenté dans ce chapitre une nouvelle méthode que nous avons proposées, afin de classifier des tweets politiques obtenus de la source que nous avons citée. Nous avons commencé d'abord par dérouler un processus de prétraitement de données, pour enlever l'ambigüité dans le but d'améliorer la précision et d'avoir une meilleure analyse de sentiment. Ensuite, nous avons appliqué notre nouvelle méthode pour classifier les tweets. Dans le chapitre suivant, nous discuterons les résultats obtenus lors de cette application, nous comparons cette dernière par rapport à d'autres modèles.

Chapitre 4

Evaluation des performances

4.1 Introduction

Dans ce chapitre, nous présentons les outils et les langages utilisés pour implémenter La méthode proposée précédemment. Par la suite, nous montrons les résultats obtenus après la comparaison entre notre méthode et les autres méthodes de classification.

4.2 Environnement de Travail

4.2.1 Environnement matériel

Quand il s'agit de l'exigence de matériel, nous n'avons pas vraiment besoin de quelque chose de fantastique, un ordinateur qui marche correctement et sous n'importe quel système d'exploitation (*Windows, Linux, Mac Os*), est suffisant vu que dans notre cas, nous avons utilisé deux ordinateurs portable acer, Samsung doté d'un processeur I3 ayant 4gb de ram.

4.2.2 Environnement logiciel

Nous avons utilisé le langage de programmation Python la version 3.6.4. Python est un langage de programmation général, interprété, interactif, orienté objet et de haut niveau [**25**]. Il a été créé par Guido van Rossum entre 1985 et 1990. Comme Perl, le code source Python est également disponible sous licence GNU ou General Public License (*GPL*).

La *figure* 4.1 illustre le logo du langage Pyhton :

FIGURE 4.1: Logo de Python

4.3 Framework de programmation

Nous avons utilisé le framework Anaconda car c'est un distributeur libre et open source du langage de programmation Python appliqué au développement d'applications dédiées à la science de données et à l'apprentissage automatique (*traitement de données à grande échelle, analyse prédictive, calcul scientifique*) [**27**].

La *figure* 4.2 illustre le logo du framework de programmation Anaconda :

FIGURE 4.2: Logo d'anaconda

Nous avons utilisé l'environnement de développement Spyder (*nommé Pydee dans ses premières versions*). C'est un environnement de développement pour Python. Libre (*Licence MIT*) et multiplateforme (*Windows, Mac OS, GNU/Linux*) [**28**]qui contient nombreuses bibliothèques d'usage scientifique : Matplotlib, NumPy, SciPy et IPython.

La *figure* 4.3 présente le logo de l'environnement de développement Spyder :

FIGURE 4.3: Logo de l'environnement de développement Spyder

4.4 Bibliothèques utilisées

- **NLTK** [1] Natural Language ToolKit (*NLTK*) est une plate-forme leader pour la construction de programmes Python pour travailler avec des données de langage humain [**29**]. Elle fournit des interfaces faciles à utiliser pour plus de 50 ressources corporelles et lexicales telles que WordNet. Ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, tokenization, stemming, étiquetage, analyse et raisonnement sémantique.

  NLTK est disponible pour Windows, Mac OS X et Linux. C'est un projet libre, open source. Il a été appelé "un outil merveilleux pour enseigner, et travailler dans, la linguistique computationnelle en utilisant Python".

- **Package Regular expressions** [2] **(re)** ce module fournit des opérations correspondant aux expressions régulières. Qui utilisent le caractère barre oblique inverse ('\') pour indiquer des formes spéciales ou pour permettre l'utilisation de caractères spéciaux sans invoquer leur signification particulière [**30**].

- **Package word_tokenize** [3] Suite à l'aide de NLTK word_tokenize permet de diviser

1. Traitement des textes pour la classification, tokenization, stemming, étiquetage, analyse et raisonnement sémantique.
2. Motif de recherche d'une séquence de caractères.
3. Diviser une phrase à une liste de mots sépares.

une phrase à une liste de mots séparés [**31**].

- **NumPy** [4] est le paquet du traitement de tableau [**32**] pour les nombres, les chaines de caractères, les enregistrements et les objets.
- **Pandas** [5] est une bibliothèque libre, écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données [**33**]. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques.

4.5 Evaluation

Matrice de confusion

Une matrice de confusion [**29**] ou tableau de contingence sert à évaluer la qualité d'une classification. Elle est obtenue en comparant les données classées avec des données de référence qui doivent être différentes de celles ayant servi à réaliser la classification. Elle ne doit pas être confondue avec la table de performance qui sert à évaluer l'homogénéité thématique des zones-test servant à réaliser une classification assistée.

Valeur de prédictions			
		Républicain	Démocrate
Valeurs réelles	Républicain	VP	FN
	Démocrate	FP	VN

TABLE 4.1: Matrice de confusion

Les paramètres d'évaluation sont définis comme suit :

— Vrai positive VP : classe positive classée comme étant positive.

— Vrai négative VN : classe négative classée comme étant négative.

— Faux positive FP : classe négative considérée positive.

— Faux négative FN :classe positive considérée négative.

4. Traitement de tableau pour les nombres, les chaines de caractères, les enregistrements et les objets.
5. Fournissant des structures et des outils d'analyse de données hautes performances.

Nous avons utilisé : Accuracy, précision, et le rappel comme paramètres d´évaluation de notre nouvelle méthode.

— **Rappel** (*Recall*) : permet de déterminer la proportion de résultats positifs réels qui ont été identifiés correctement :

$$Rappel = \frac{VP}{VP + FN} \tag{4.1}$$

— **Précision** : permet de déterminer la proportion d'identifications positives qui ont étés effectivement correctes :

$$Précision = \frac{VP}{VP + FP} \tag{4.2}$$

— **Accuracy** : permet de déterminer la proportion de prédictions correctes.

$$Accuracy = \frac{VP + VN}{VP + FP + VN + FN} \tag{4.3}$$

4.6 Résultats d'évaluation

Commençons, tout d'abord par comparer notre méthode de classification aux différentes approches existantes, en terme de Accuracy :

Nous avons déroulé les algorithmes suivants : MultinomalNaiveBayes, DecisionTrees , k-nearestneighbors *knn*(k=1,k=3), AdaBoostClassifier, RandomForest. Avec la taille du trainset égale à 0.7 et la taille du testset égale à 0.3 (*qui est la division la plus fréquente*). Nous avons calculé l'Accuracy des approches par rapport à notre dataset. Les résultats obtenus sont résumés dans cette illustration :

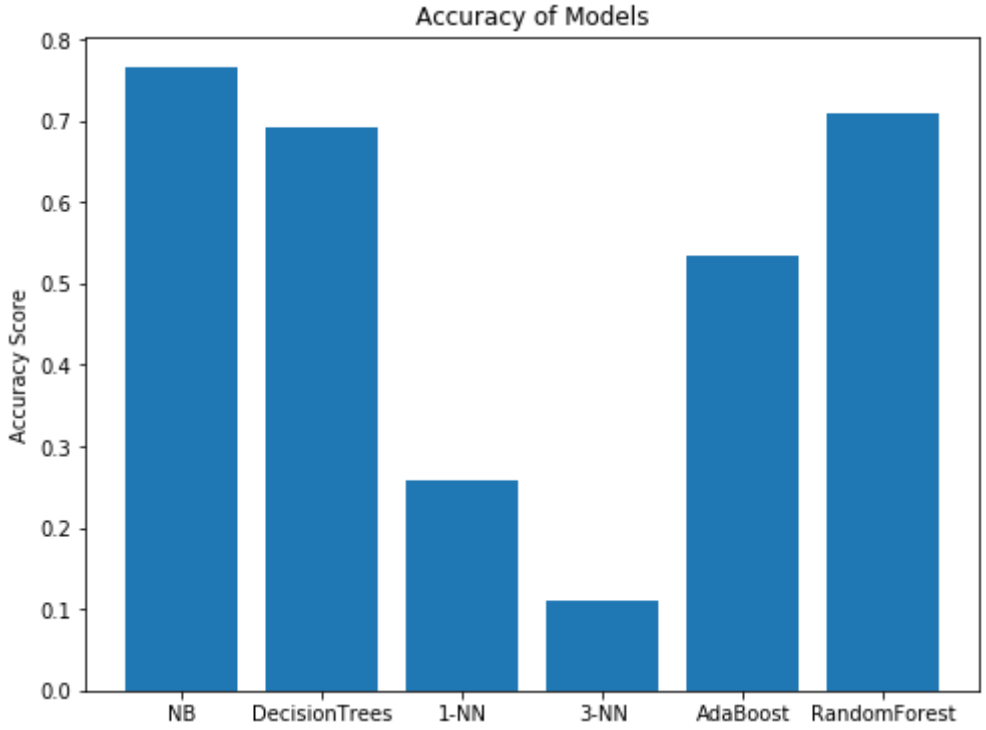

FIGURE 4.4: Accuracy des approches par rapport à notre dataset

La *Figure* 4.4 met en évidence que la méthode la plus performante c'est Naive Bayes, suivit par RandomForest après DecisionTrees. Quant aux deux méthodes k-nearestneighbors *knn*(k=1,k=3) et AdaBoostClassifier, leur performance n'était suffisamment satisfaisable. Ensuite, nous avons déroulé le code de notre méthode proposée avec les mêmes données (*trainset=0.7* ,*testset=0.3*), les résultats sont comme suit :

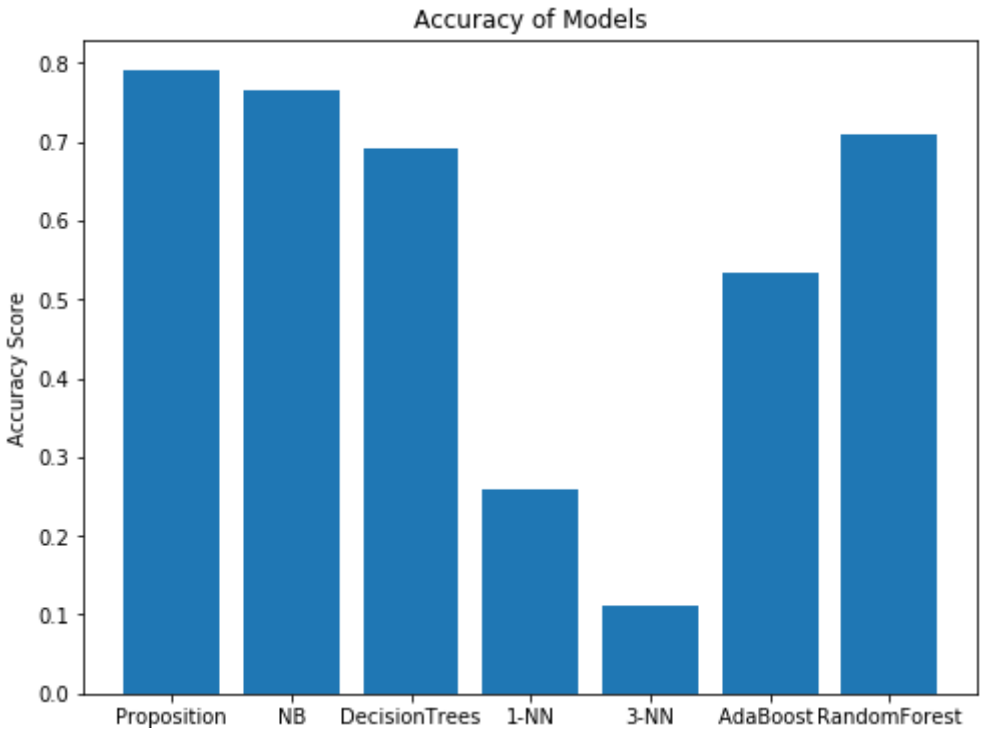

FIGURE 4.5: Accuracy des approches et proposition par rapport à notre dataset

La *Figure* 4.5 montre que notre méthode a atteint la plus grande valeur de Accuracy suivant la taille 0,7 pour le train et 0,3 pour le test qui est 79%.

Sachant que le classificateur MNB convient à la classification avec des caractéristiques discrètes (*par exemple, la classification de texte*) ce qui est exactement ce qu'on cherche puisque les tweets ce sont un ensemble de mots. Pour s'assurer que l'algorithme MNB est le plus performant dans le cas de notre étude, nous avons tenté de dessiner un graphe illustrant les différentes valeurs de Accuracy de notre modèle (*voir la figure 4.6*) :

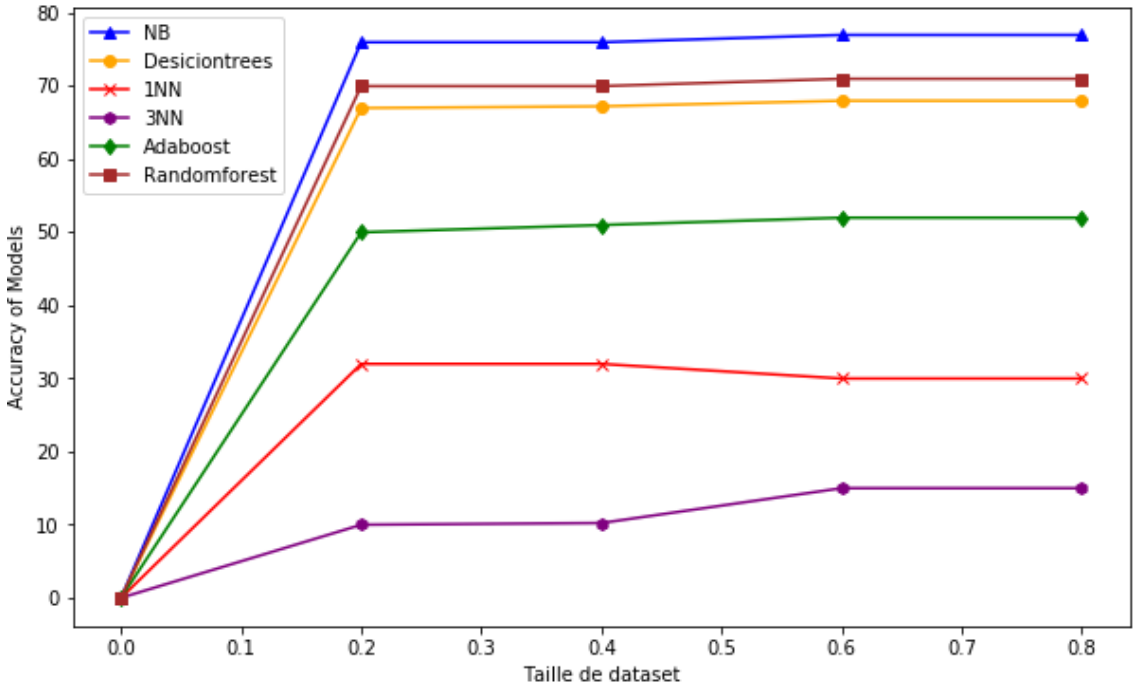

FIGURE 4.6: La variation de l'Accuracy en fonction de la taille du dataset

La *Figure* 4.6 décrit la variation du pourcentage de prédictions correctes (*Accuracy*) en fonction de la taille du dataset qui se varie de 10% jusqu'à 80% de la taille globale de ce dernier. La courbe bleu représente l'algorithme Multinomial Naive Bayes. Quant à la courbe marron elle représente la méthode RandomForest, la courbe orange concerne Decision Trees, la courbe violet concerne la méthode AdaBoostClassifier, la courbe verte représente la méthode k-nearest neighbors *knn*(k=1), et enfin la courbe rouge concerne la méthode k-nearest neighbors *knn*(k=3).

La méthode Multinomial Naive Bayes est la plus performante tout au long des différentes tailles de notre dataset. Dans cette optique, nous nous sommes lancé un défi de proposer une méthode qui sera plus performante que Multinomial Naive Bayes.

4.6.1 Les résultats obtenus par la méthode proposée par rapport à Multi Nominal Naive Bayes

- **Accuracy** : Afin de comparer notre méthode proposée à la méthode Multinominal Naive Bayes, nous traçons ce graphe montrant précisément les différences entres ces dernières en terme d'Accuracy .

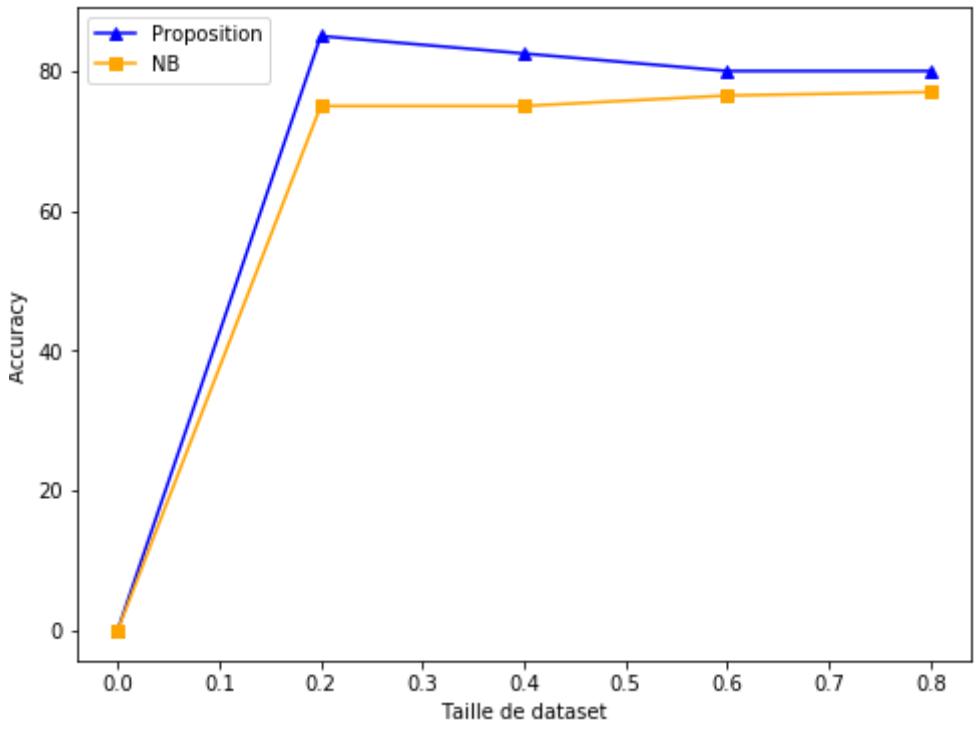

FIGURE 4.7: la variation du Accuracy des deux méthodes en fonction de la taille du dataset

La *Figure* 4.7 décrit la variation du pourcentage de prédictions correctes *Accuracy* en fonction de la taille du dataset qui se varie de 10% jusqu'à 80% de la taille globale de ce dernier, la courbe bleu représente l'Accuracy de notre méthode proposée quant à la courbe orange, elle représente l'Accuracy de Multinomial Naive bayes, ainsi nous remarquons que les valeurs cernées entre 0,0 et 0.05 des deux modèles sont égaux en terme de Accuracy.

Après à partir de 0.05 nous remarquons que notre méthode a une meilleure Accuracy que MNB tout au long des abscisses. En effet, les deux modelés atteignent le

pourcentage de prédictions correctes maximal qui est 85% pour notre méthode et 72% pour l'autre modèle au point 0.2 qui est 20% de la taille du dataset. Après bien évidement ça baisse pour les deux modelés. Enfin ça devient constant.

- **La Précision** : Nons traçons cette figure, pour mettre en évidence les différences en terme de précision suivant la taille du dataset, entre la méthode que nous avons proposée et MNB.

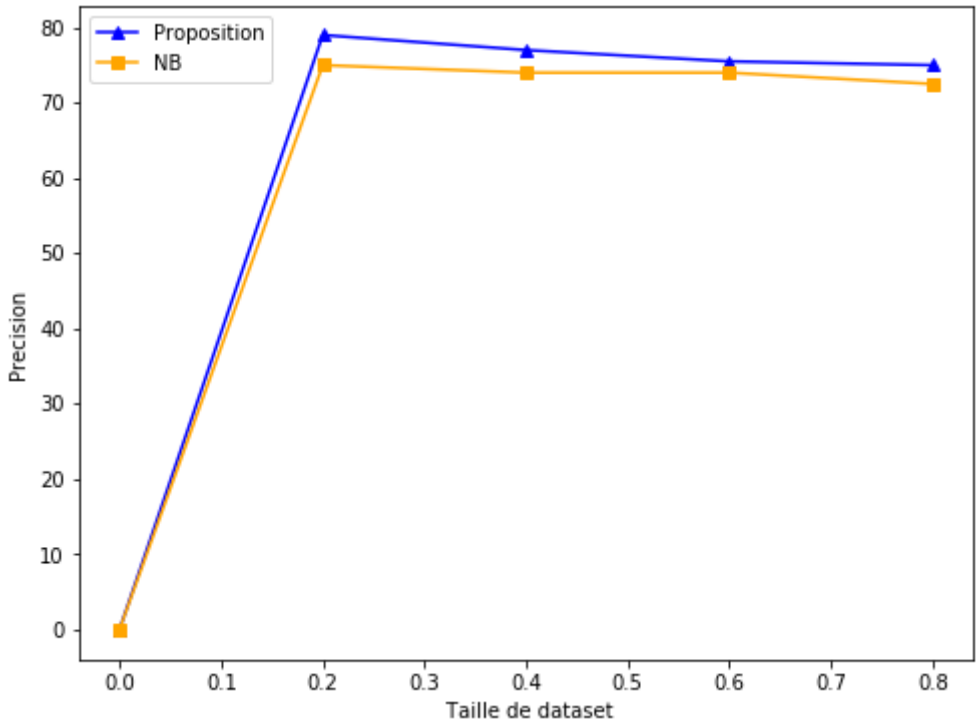

FIGURE 4.8: Variation de la Précision des deux méthodes en fonction de la taille du dataset

La *Figure* 4.8 décrit la variation de la Précision des deux modèles en fonction de la taille du dataset qui se varie de 10% jusqu'à 80% de la taille globale de ce dernier, ainsi nous remarquons que les valeurs de la précision des deux modèles sont égales au niveau des abscisses cernées entre 0.0 et 0.1. Après à partir de 0.1, nous remarquons que notre méthode a une meilleure précision que Multinomial Naive bayes tout au long des abscisses.

En effet, les deux modelés atteignent la précision maximale qui est 79% pour notre

méthode et 75% pour Multinomial Naive bayes au point 0.2 qui est 20% de la taille globale du dataset. Après la précision baisse légèrement pour les deux modèles, mais ça reste au-dessus de 70%. Enfin ça tend vers une précision constante.

- **Le Rappel** : Nous traçons cette figure, dans le but de visualiser la différence en terme de rappel(*recall*) entre notre méthode proposée et MNB.

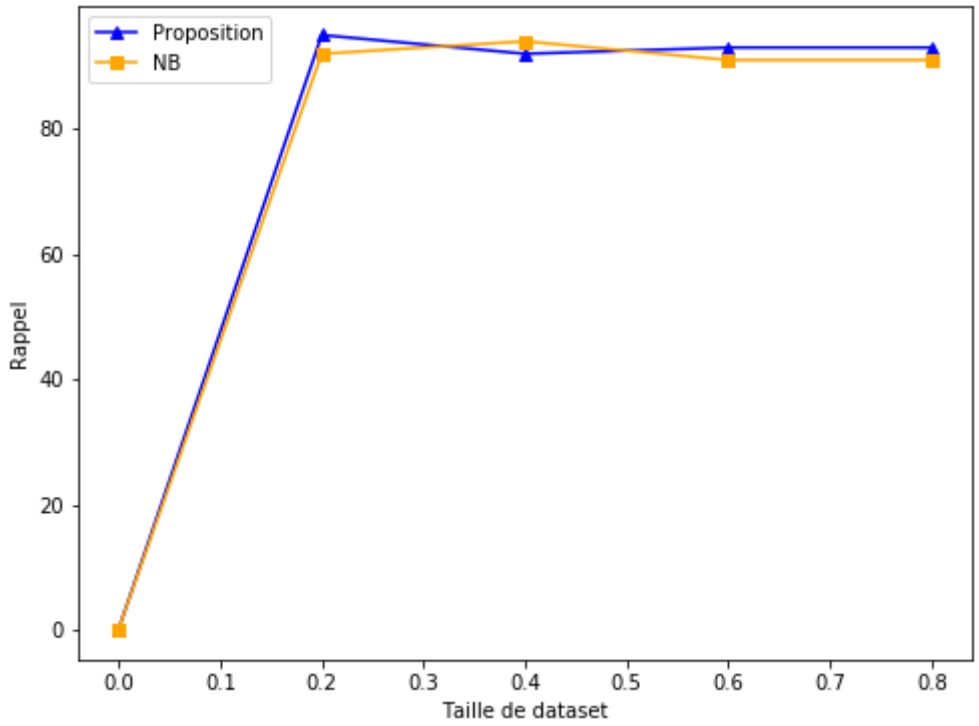

FIGURE 4.9: La variation du Rappel des deux méthodes en fonction de la taille du dataset

La *Figure* 4.9 décrit la variation du Rappel des deux modèles en fonction de la taille du dataset qui se varie de 10% jusqu'à 80% de la taille globale de ce dernier, ainsi nous remarquons que les valeurs du rappel des deux modèles sont égales au niveau des abscisses cernées entre 0.0 et 0.1. Après de 0.1 à 0.3, nous remarquons que notre méthode a un meilleure rappel que Multinomial Naive bayes, en effet, notre méthode atteint le rappel maximal qui est 95% au point 0.2 qui est 20% de la taille globale du dataset. Ensuite, de 0.2 à 0.4, notre méthode subit une baisse en terme de rappel, quand à la méthode Multinomial Naive bayes subit une hausse jusqu'à ce qu'elle atteigne son rappel maximal qui est 94%. Après de 0.4 jusqu'à 0.6 , notre méthode

subit une hausse, contrairement à la méthode Multinomial Naive bayes, qui subit une baisse en terme de rappel. Enfin, à partir de 0.6, le rappel des deux modelés tend vers un une valeur constante, sachant que celui de notre méthode est plus élevé que celui de la méthode MNB.

4.7 Conclusion

Dans ce dernier chapitre nous avons mentionné le langage et les principaux outils, qui ont été implémentés dans notre programme, afin de savoir quel niveau d'exactitude le programme donne. Nous avons testé la proposition en appliquant les paramètres d'évaluation par rapport à plusieurs méthodes existantes essentiellement Multinominal NB qui a été la plus performante. Les résultats ont été globalement satisfaisants que ça soit en Accuracy, rappel ou précision.

Conclusion générale

L'analyse d'opinion est de différencier et de classer les points de vue ou des sentiments ou des évaluations dans le contenu composé .Plusieurs recherches s'intéressent à la tâche de l'analyse d'opinion, en particulier dans le domaine de la politique.

Afin d'atteindre notre objectif, nous avons essaye de bien comprendre et analyser les techniques de classification des opinions qui existent déjà.

Notre travail s'intègre dans ce même axe de recherche, nous proposons un système de classification des opinions politiques des utilisateurs du réseau social Twitter, concernant les élections présidentielles républicaines américaines en deux parties : républicain, démocrate.

Notre contribution consiste à calculer le score républicain et le score démocrate de chaque tweet, en se basant essentiellement sur la fréquence des mots composant ce dernier dans le dataset que nous avons divisé en deux : les tweets ayant le Party républicain et le reste ayant le Party démocrate, ainsi sur la probabilité de chaque mot du tweet dans chaque partie. Après, la classification se fait en prenant en compte le score le plus élevé.

D'après la comparaison entre notre méthode proposée et les méthodes existantes, essentiellement Multinominal Naive Bayes qui s'est avéré la plus performante. Nous avons déduit que notre méthode a donné de meilleurs résultats, ce qui était notre objectif dès le commencement de ce travail.

Nous envisageons comme perspectives du travail réalisé dans ce mémoire :

— Prise en charge de la sémantique des mots plutôt que leur occurrence.

— L'application de notre méthode sur des datasets appartenant à d'autres domaines que la politique.

— L'application de notre méthode sur des datasets contenant des tweets arabes ou plutôt français.

— L'enrichissement de la partie train *dictionnaire* avec d'autre termes du domaine politique.

Bibliographie

[1] [Comment mettre en lumière les réseaux sociaux dans les corpus historiques numériques. Jose Miguel Vieira. 2020]

[2] [Réseaux Sociaux Numériques : Essai de catégorisation et de cartographie des controverses. NISRINE ZAMMAR. février 2017]

[3] [https ://megganeangellotti.wordpress.com/titre-5/ Consulté le 23 /11/ 2019]

[4] [Représentation de soi et identité numérique. Une approche sémiotique et quantitative de l'emprise culturelle du web 2.0. Georges Fanny. Avril 2009]

[5] [Sentiment Analysis. TiejianLuo, Su Chen, Guandong Xu, and Jia Zhou. juin 2013]

[6] [Analysis of community in social networks. Anh Dang. 2017]

[7] [Did Big Data Win the Election for Trump. Katie Wozniak. Jessica Cote, Noah Steinfeld, Shweta Ramdas. Avril 2017]

[8] [Dictionnaires Larousse français]

[9] [Natural Language Processing, Second Edition. N.Indurkhya. Damerau .2010]

[10] [Détection d'opinions à partir de Twitter.HASNI Bachir, GOUDJIL Ayyoub. aout 2010]

[11] [Traitement automatique des langues. Taha Zerrouki. Aout 2020]

[12] [Inductive infeTencefoTlargescaletext classification. berlinheidelberg : springer-verlag . Ribeiro. Aout 2020]

[13] [https ://projeduc.github.io/intro_apprentissage_automatique/arbres.html ?fbclid=IwAR22VP2nvPjJ40i2AJfNgCcAZ3_t8BuTC_RiaZyH3jVbdFW_CV12vcpPaaw consulté le 24/10/2020]

[14] [https ://fr.audiofanzine.com/techniques-du-son/forums/t.662434,ces-appareils-audio-etonnants,p.11.html consulté 24/10/2020]

[15] [Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information. Guillaume Cleuziou. Juillet 2006]

[16] [Classification par réseaux de neurones dans le cadre de la scattérométrieellipsométrique. SabitFawzi , Philippe Zaki. 2016]

[17] [K-means algorithm for the detection and delineation of QRS-complexes in Electrocardiogram. Hanjie Chen ,Koushik Maharatna. February 2020]

[18] [http ://medium.com/@julien.wiesel/lart-subtil-du-clustering-ebac591e9ebd consulté 24/10/2020]

[19] [méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information. Julien Ah-Pine. 2019]

[20] [Modèle de mélange Gaussien : Application sur image cytologique.Koudri Mohammed. septembre 2011]

[21] [Maximum Likelihood from Incomplete Data via the EM Algorithm. A.P. Dempster, N.M. Laird, Donald Rubin. Avril 2017]

[22] [Les Traitement du signal. marie tahon. 2015]

[23] [Digital Image Processing.Rafael C. Gonzalez, Richard E. Woods. 2008]

[24] [Sentiment analysis algorithms andapplications. HodaKarashyWalaaMedhat, Ahmed Hassan. mai 2014]

[25] [Detection d'opinion à partir de Twitter. GoudjilAyyoubHasni Bachir. 2018]

[26] [https ://www.ionos.fr/digitalguide/web-marketing/analyse-web/analyse-tf-idf/ : :text=Le%20TF%2DIDF%20est%20une,projet%20Web%20dans%20son%20ensemble.]

[27] [Python - Overview. s.d, sur Tutorials Point SimplyEasyLearning : https ://www.tutorialspoint.com/python/python_overview.htm Consulté le 07/ 05/2020

[28] [Open Source Community. Anaconda. Avril 2019]

[29] [introduction à Python. Antony Lesage. janvier 2018]

[30] [Natural LanguageToolkit. N. (Éd.)]

[31] [Neural Networks. Joseph Awange, BelaPalancz, LajosVolgyesi 2020]

[32] [https ://kite.com/python/docs/nltk.word_tokenize Consulté le 10 /06/ 2020]

[33] [http ://pandas.pydata.org/pandas-docs/stable/overview.html#license Consulté le 14 /06/ 2020]

Résumé

Les réseaux sociaux sont une excellente source d'information et d'extraction d'opinion ou la majorité des internautes utilisent ces plateformes pour partager leurs sentiments et opinions. L'exploitation de ces opinions ne peut être que fructueuse. Dans notre travail, nous avons exposer le problème de l'analyse des sentiments sur les réseaux sociaux en présentant une nouvelle méthode de classification supervisé des opinions politiques concernant les élections présidentielles républicaines américaines en deux parties : républicain, démocrate faites dans ce contexte sur des Tweets.

Mots-clés : Fouille des opinions, Twitter, Classification supervisé, Elections.

Abstract

Social networks are an excellent source of information and opinion extraction where the majority of Internet users use these platforms to share their feelings and opinions. The exploitation of these opinions can only be fruitful. In our work, we exposed the problem of sentiment analysis on social networks by presenting a new method of supervised classification of political opinions concerning the American Republican presidential elections in two parts : Republican, Democrat made in this context on Tweets.

Keywords : Opinion mining, Twitter, Supervised classification, Elections.

الملخص

تعد الشبكات الاجتماعية مصدرًا ممتازًا للمعلومات واستخلاص الآراء حيث يستخدم غالبية مستخدمي الإنترنت هذه المنصات لمشاركة مشاعرهم وآرائهم. إن استغلال هذه الآراء لا يمكن إلا أن يكون مثمرا. في عملنا ، كشفنا عن مشكلة تحليل المشاعر على الشبكات الاجتماعية من خلال تقديم طريقة جديدة لتصنيف الآراء السياسية المتعلقة بالانتخابات الرئاسية الجمهورية الأمريكية في جزأين: جمهوري ، وديمقراطي صنع في هذا السياق على التغريدات .

الكلمات المفتاحية: التنقيب عن الرأي ، تويتر ، التصنيف المراقب .