République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

Pour l'obtention du diplôme de Master en Informatique

Option : *Système d'Information et de Connaissances (S.I.C)*

Thème

La détection automatique de l'ironie dans les tweets algériens

Réalisé par :

- **Chikh Mohammed Walid**

Présenté le 25 Juin 2020 devant le jury composé de :

- *Madame El Yebderi Zineb* *(Présidente)*
- *Monsieur Abderrahim Mohammed El Amine* *(Encadreur)*
- *Monsieur Boudefla Mohammed El Amine* *(Examinateur)*

Année universitaire: 2019-2020

Remerciements

Ce travail n’aurait jamais été possible sans le soutien et l’appui d’un ensemble de personnes que je tiens à remercier ici chaleureusement.

En premier lieu, je tiens à remercier chaleureusement mon encadreur, Docteur Abderrahim Mohammed El Amine, qui m’a fait confiance et m’a permis de développer un thème de recherche passionnant et d’actualité. Plus qu’un directeur de mémoire, il a su se montrer présent tout au long de ce semestre et me soutenir dans mes démarches. Je tiens ici à lui témoigner toute ma gratitude et ma reconnaissance pour sa générosité, le temps qu’il m’a accordé et la patience dont il a fait preuve. Mes nombreux échanges avec lui et ses remarques toujours très pertinentes ont permis d’approfondir et d’enrichir ce travail de projet de fin d’études malgré cette période difficile marquée par cette pandémie de Covid-19. Ce travail n’aurait en effet jamais été possible sans son implication et son dévouement. Qu’il trouve ici toutes les marques de mon profond respect et de mon admiration.

Je tiens à remercier aussi l’ensemble des membres du jury, Madame Elyebdri Zineb et Monsieur Boudefla Mohammed El amine pour m’avoir fait l’honneur d’accepter de juger ce travail de master.

Évidemment, les mots ne sauraient décrire l’immense gratitude envers tous mes enseignants du département d’informatique, qui m’ont permis d’avoir une formation de qualité durant toutes mes années d’étude à l’université de Tlemcen.

Enfin, un grand merci à mes parents, pour leurs encouragements pour affronter les difficultés présentes tout au long de ce dur chemin de mes études.

Dédicaces

A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,

A ma chère sœur Batoul pour son encouragement permanent, et son soutien moral,

A mes chers frère Firas et Mounib pour leur appui et leur encouragement,

A toute ma famille pour leur soutien tout au long de mon parcours universitaire,

Que ce travail soit l'accomplissement de vos vœux tant allégués, et le fuit de votre soutien infaillible,

Merci d'être toujours là pour moi.

Chikh Mohammed Walid.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

API Application Programming Interface

ROC Receiver Operating Characteristic

AUC Area Under the Curve

SVM Support Vector Machine

B2B Business to Bussiness

TPE Très Petites Entreprises

PME Petites et Moyennes Entreprises

Introduction générale

Actuellement, les institutions étatiques tout comme les entreprises, nous citons essentiellement les secteurs de la communication, des études de marché et du marketing, ces derniers se basent très souvent sur l'avis et l'opinion publique pour prendre des décisions et prévoir des pistes de réflexion sur des domaines stratégiques de grande envergure. Le traitement automatique des textes porteurs d'opinions a ainsi connu un déclic réel depuis l'apparition des réseaux sociaux comme Facebook et Twitter. Cette utilisation quotidienne des réseaux comme Twitter et Facebook a changé l'image du web 2.0 et lui a donné une nouvelle dimension et aussi de nouveaux défis. Les posts, les blogs et les commentaires publiés sur Twitter ou Facebook reflètent l'interaction d'utilisateurs avec les événements réels qui se déroulent partout dans le monde, comme les évènements politiques, sportifs, culturels ou sanitaires, etc. Ces évènements réels ont un impact direct sur la quantité de tweets mises en ligne. Devant cette grande quantité de données disponibles sur les réseaux sociaux et cette diversité de sources, la conception et l'implémentation d'outils pour extraire et analyser ces données sur un sujet particulier constitue un défi majeur pour le monde académique et de recherche. L'intérêt de ce type d'outils est considérable, pour les institutions et les entreprises qui visent à obtenir un retour client sur leurs produits ou leur image de marque comme pour les individus souhaitant se renseigner pour l'achat d'un produit ou l'organisation d'une sortie touristique.

L'importance donnée à ces évènements sur les réseaux sociaux généralement et sur Twitter plus précisément est un défi majeur pour le monde de la recherche, tout d'abord parce qu'un sujet sur Twitter est caractérisé par plusieurs termes (ces termes peuvent être des hashtags) qui peuvent changer avec le temps où certains peuvent devenir moins utilisés et d'autres peuvent apparaître. Cette contrainte, nous oblige à tenir compte de ces termes utilisés pendant le processus d'analyse et de traitement. Cela représente l'un de nos objectifs dans ce projet de fin d'étude. Lors de ce mémoire nous avons identifié au préalable les ensembles de tweets qui parlent du même thème, nous citons le domaine du service (Algérie-telecom), du média (télévision), du tourisme (découverte de l'algérie), santé (Covid 19) et politique (Hirak).

Aussi Pour connaitre les opinions d'une population donnée, il est difficile de lire tous les commentaires qui portent sur un sujet vu la grande quantité trouvée. Cette contrainte a encouragé beaucoup de chercheurs cités dans la littérature à proposer des travaux de recherche dont l'objectif principal est d'analyser les opinions exprimées par des internautes. Selon (Pak & Paroubek, 2010), une opinion peut être soit positive, négative, ou neutre, ce qui revient à un problème de classification en 3 classes.

Contributions

Dans le cadre de notre projet de fin d'études de master, nous avons proposé un outil qui permet d'analyser les tweets afin d'identifier l'ironie sur différents sujets dans le but de prédire les tendances et les préoccupations des utilisateurs des réseaux sociaux. Notons que notre travail entre dans le cadre d'un projet de recherche PRFU intitulé : analyse intelligente d'opinions sur les réseaux sociaux et le Web nouvellement agrée en 2020 sous le code C00L07UN130120200002.

Notre intérêt a été porté sur les tweets algériens. Les objectifs principaux de ce mémoire concernent les points suivants :

1. Collecte des tweets pour construire un corpus Algérien
2. Annotation manuelle du corpus
3. Prétraitement du corpus
4. Classification automatique des tweets du corpus (détection de l'ironie)

Organisation du mémoire

Ce mémoire est organisé en 04 chapitres, précédé par une introduction générale et suivi d'une conclusion générale :

- Le premier chapitre présente dans sa première partie des concepts de base et des définitions concernant les réseaux sociaux. Dans la deuxième partie, nous avons aussi montré l'intérêt des données massives (big data) existantes dans les réseaux sociaux et le besoin à des techniques de fouille de données pour les analyser. Nous avons terminé ce chapitre en expliquant ce phénomène d'ironie sur les réseaux sociaux et le besoin de l'identifier.
- Le deuxième chapitre explique la différence entre les opinions et les sentiments. Nous avons aussi donné un aperçu sur les approches d'opinion-mining, ces dernières

ont la capacité à corréler tous les attributs et les opinions des personnes sondées à leurs propriétés socio-démographiques. Nous avons aussi présenté dans ce chapitre le principe de la collecte des données pour constituer des corpus et l'intérêt croissant pour automatiser cette tâche vue la quantité énorme des données disponible sur les réseaux sociaux. A la fin de ce chapitre, nous avons cité les trois méthodes d'analyse des données, lexicale, automatique et hybride.

- Le troisième chapitre présente la première partie de notre travail de master et qui vise principalement à la constitution et l'annotation d'un corpus de tweets algériens.
- Dans le quatrième chapitre nous avons implémenté quatre types de classifieurs (SVM, Naivebyes, random forest et K-means) pour l'identification des tweets ironiques. Nous avons utilisé les trois modes d'apprentissage à savoir le mode supervisé, le mode non supervisé et le mode semi-supervisé. Nous avons discuté tous les résultats obtenus dans les différentes expérimentations réalisées.
- Enfin ce mémoire est clôturé par une conclusion générale, dans laquelle nous présentons nos principaux résultats obtenus, en mettant l'accent sur les avantages et les éventuelles limites de notre démarche le long de ce travail de master, ceci nous a permis de présenter nos perspectives pour des travaux futurs qui visent à améliorer d'avantage nos résultats en vue d'une implémentation réelle sur une plateforme appropriée.

Chapitre 1

1. Les concepts de base des réseaux sociaux

1.1. Introduction

Depuis deux décennies, le web 2.0 a cristallisé les liens et des contacts qui pouvaient exister dans une communauté donnée. Plusieurs mécanismes et plateformes sont apparus actuellement permettant ainsi des échanges personnalisés, sans tenir compte des facteurs espaces et temps. Nous comprenons ainsi que le réseau social web se base sur l'intelligence collective et la collaboration en ligne. Le web a encouragé la montée en puissance des réseaux sociaux, devenus pour certains de véritables médias sociaux, qui permettent aux internautes et aux professionnels de créer une page profil et de partager des informations, photos et vidéos avec leur réseau. Des espaces de partage qui se distinguent par leur utilité (personnel, professionnel, rencontres...), leur logo et leurs audiences.

Selon Wikipédia, l'expression « réseau social » dans l'usage habituel renvoie généralement à celle de « médias sociaux », qui recouvre les différentes activités qui intègrent technologie, interaction sociale entre individus ou groupes d'individus, et la création de contenu (Wikipedia, Réseau social, 2020).

Les auteurs Andreas Kaplan et Michael Haenlein définissent un réseau social comme un groupe d'applications en ligne qui se fondent sur la philosophie et la technologie du net et permettent la création et l'échange du contenu généré par les utilisateurs (Wikipedia, Réseau social, 2020)

Notons qu'aujourd'hui, en discutant de réseau social nous avons plutôt tendance à penser web, il faut savoir que réellement, un réseau social est un groupe de personnes qui maintiennent des liens IRL (In Real Life). Souvent, il s'agit d'individus partageant des idées, des pensées et des intérêts communs.

Le média social instaure une communication sociale qui permet aux individus de collaborer, créer, organiser, modifier ou commenter un contenu. Avec l'avènement du web, le réseau social prend alors une toute autre allure.

1.2. Différents types des réseaux sociaux

Nous vous proposons dans ce chapitre quelques exemples de réseaux sociaux par typologie.

1.2.1. Les réseaux sociaux à dominante visuelle

Les réseaux sociaux les plus cités dans cette catégorie concernent en particulier ceux qui privilégient les posts d'images, les photos et l'infographie.
Nous pouvons citer essentiellement dans cette catégorie :

- **Instagram :** C'est une application ou un service de partage de photos et de vidéos fondés et lancés en octobre 2010 par l'Américain Kevin Systrom et le Brésilien Michel Mike Krieger. Depuis 2012, l'application appartient à Facebook, elle est disponible sur plates-formes mobiles de type iOS, Android et Windows Phone et également sur ordinateurs avec des fonctionnalités réduites. Instagram revendique plus d'un milliard d'utilisateurs à travers le monde, dont 75 % d'utilisateurs en dehors des États-Unis, selon les chiffres officiels fournis en juin 2018. L'entreprise s'adresse à ses utilisateurs par la dénomination Igers. (Wikipedia, Instagram, 2020).
- **Pinterest :** C' est un site web américain mélangeant les concepts de réseautage social et de partage de photographies, lancé en 2010 par Paul Sciarra (en), Evan Sharp (en) et Ben Silbermann (en). Il permet à ses utilisateurs de partager leurs centres d'intérêt et passions à travers des albums de photographies glanées sur Internet. Le nom du site est un mot-valise des mots anglais pin et interest signifiant respectivement épingle et intérêt. La croissance du nombre de visiteurs s'accélère à partir de la fin de l'année 2011 selon différentes sociétés d'analyse de trafic. En décembre, Pinterest se classe 10^{e} parmi les réseaux sociaux les plus populaires aux États-Unis selon Experian Hitwise. D'après comScore, Pinterest attire 17,8 millions de visiteurs au mois de février 2012, contre 11,7 millions le mois précédent, et se classe 3^{e} en termes de croissance parmi les sites américains (Wikipedia, Pinterest, 2020).

1.2.2. Les réseaux sociaux professionnels

Les réseaux sociaux les plus connus dans cette catégorie sont Linkedin et Viadeo. Ces derniers sont les mieux placés pour donner de la crédibilité en tant qu'utilisateur professionnel. Ils permettent par ailleurs d'atteindre facilement une cible B2B (business to business).

- **LinkedIn :** C'est un réseau social professionnel en ligne créé en 2002 à Mountain View (Californie). En 2016, Microsoft annonce le rachat du réseau social pour un montant de 26,2 milliards de dollars américains soit 23,3 milliards d'euros. En 2018, LinkedIn atteint 546 millions d'utilisateurs (Wikipedia, LinkedIn, 2020).
- **Viadeo :** C'est un réseau social professionnel en ligne créé en 2004 à Paris qui permet de construire et d'agréger son réseau professionnel. Il se définit comme un réseau de connaissances qui facilite le dialogue entre professionnels. En 2018, il revendique 7,5 millions de membres en France. Pour ses membres, c'est aussi un outil de gestion de réputation en ligne et de marketing personnel. L'une des caractéristiques de Viadeo est de réunir des professionnels issus de TPE / PME puisque les profils présents dans des entreprises de moins de 50 employés représentent 45 % des inscrits. (Wikipedia, Viadeo, 2020).

  Les principaux usages de la plateforme Viadeo sont :
  - La création et la gestion de son profil professionnel (rédiger un curriculum vitæ, mettre à jour ses activités, ses compétences etc.).
  - La création et la gestion de son réseau (entrer en contact avec d'autres membres, recommander un utilisateur à un autre, etc.). Les utilisateurs, notamment les commerciaux, peuvent s'en servir pour trouver des prospects.
  - Une meilleure visibilité sur les moteurs de recherche.

1.2.3. Blogging et micro blogging

Nous citons dans cette catégorie deux réseaux sociaux à savoir Twitter et Tumblr.

- **Twitter :** C'est un réseau social de micro blogage géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés *tweets*, sur internet, par messagerie instantanée ou par SMS. Ces messages sont limités à 280 caractères. Twitter a été créé le 21 mars 2006 par Jack Dorsey, Evan Williams, Biz Stone et Noah Glass. Le service en ligne est rapidement devenu

populaire. Le 5 mars 2017, il compte 313 millions d'utilisateurs actifs par mois, 500 millions de tweets envoyés par jour et est disponible en plus de quarante langues. En 2018, Twitter annonce pour la première fois avoir fait du profit, notamment à la suite de restrictions budgétaires. Le siège social de Twitter Inc. se situe aux États-Unis à San Francisco (Wikipedia, Twitter, 2020).

Nous notons que ce type de réseau social constitue notre champ d'application et plus de détails seront présentés dans le chapitre trois (Résultats et analyse).

- **Tumblr :** C'est une plate-forme de microblogage créée en 2007 et permettant à l'utilisateur de poster du texte, des images, des vidéos, des liens et des sons sur son tumblelog. Elle s'appuie principalement sur le reblogage. La plate-forme a connu une augmentation rapide du nombre d'usagers. En janvier 2008, Tumblr comptait quelque 170 000 utilisateurs. Début août 2010, environ 6,6 millions de personnes l'utilisaient, selon Mark Coatney, employé de Tumblr (Wikipedia, Tumblr, 2020).

1.2.4. Facebook

C'est le réseau social le plus utilisé au monde avec 2,3 millions d'utilisateurs actifs dans le monde en 2018. (Statista, 2019)

Facebook permet aux utilisateurs de publier des images, des photos, des vidéos, des fichiers et documents, d'échanger des messages, joindre et créer des groupes et d'utiliser une variété d'applications. Facebook est fondé en 2004 par Mark Zuckerberg et ses camarades de l'université Harvard, Chris Hughes, Eduardo Saverin, Andrew McCollum et Dustin Moskovitz. D'abord réservé aux étudiants de cette université, il s'est ensuite ouvert à d'autres universités américaines avant de devenir accessible à tous en septembre 2006. (Wikipedia, Facebook, 2020)

1.3. Les données massives sur les réseaux sociaux

1.3.1. Relation entre réseaux sociaux et Big data

L'exploitation des données massives (Big data ou mégadonnées) dans différents domaines distincts est un véritable potentiel qui pourrait supporter les décideurs sur divers plans. Ces données massives sont nées de l'évolution de la technologie. Cette dernière rend possible la collecte systématique, la conservation et l'analyse d'informations de tout genre, notamment les données des réseaux sociaux.

Les réseaux sociaux et le Big Data représentent des opportunités majeures pour les acteurs du Web. Les réseaux sociaux génèrent des données massives qui forment le Big Data. C'est ce que l'on appelle le Big Social Data ou Social Big Data. Sachant que les réseaux sociaux présentent de très nombreuses données sur ses utilisateurs, nous citons notamment les messages échangés, les images, les vidéos etc. Ces données très appréciées par les décideurs se multiplient chaque année. D'où vient l'intérêt d'analyser cette masse d'informations pour en extraire que ce qui est utile pour les professionnels.

L'analyse des réseaux sociaux appelé aussi fouille de données est une problématique assez ancienne qui n'est pas née avec l'avènement des réseaux sociaux tels que Facebook ou Twitter. En effet, l'analyse des données issues des relations sociales a occupé depuis plusieurs décennies de nombreux domaines tels que les sciences humaines.

1.3.2. Les réseaux sociaux comme source d'information et de données potentielles

Actuellement les réseaux sociaux sont formés de groupe d'individus ou d'organisations reliées entre elles par des échanges sociaux. Depuis, ces derniers ne cessent d'être développés et rénovés, les utilisateurs de croître et les sphères de s'élargir. Les réseaux sociaux comme sources d'information espaces d'expression et de mise en relation pour plusieurs millions d'internautes, les réseaux sociaux ne pouvaient rester en dehors de l'activité journalistique (MERCIER & PIGNARD-CHEYNEL, 2018). Les usages journalistiques qu'il décrit sont nombreux : vecteur de diffusion de la presse, espace d'expression non contraint, lieux où se crée l'information, sources pour identifier des sujets et alimenter des articles. Les réseaux sociaux constituent un moyen pour les médias de cibler des communautés divers et variées et sont un lieu de production des idées, écrits, données qui peuvent intéresser les journalistes à la recherche de sources renouvelées, d'informations rapides. Du côté des utilisateurs, les plates-formes sociales s'imposent désormais comme sources d'information, d'après l'étude sur l'information numérique menée par l'institut britannique Reuters (Delcambre A., 2016) indique que cette prise de pouvoir est profondément corrélée à l'essor du smartphone. Les réseaux sociaux apparaissent donc, d'après les données récentes, comme sources d'information indispensable, pour les utilisateurs, comme pour les journalistes. Devenus incontournables, ils ont, comme le signale (SIGNORINO T., 2016), bouleversé le monde journalistique, ainsi que les utilisateurs de l'information. La rapidité, la personnalisation, la

variété, la liberté, le caractère participatif et partagé de l'information en font leurs principaux atouts.

1.4. Fouille de données sur Big Social Data

Le Big Social Data est une collection d'ensemble de données extrêmement volumineux avec une grande diversité dans les réseaux sociaux. Le Big Social Data est également un élément central de l'analyse de l'influence sociale et de la sécurité. Cependant, les travaux en cours sur les méga données de Big Social Data se concentrent sur le traitement de l'information, comme l'exploration et l'analyse de données.

Sachant qu'aujourd'hui, l'espace de stockage des données n'est plus une contrainte, tous les décideurs et les responsables à travers le monde veulent désormais tirer profit des grands volumes de données. Ces données peuvent les aider à connaitre et s'adapter avec leur environnement spécifique (ressources humaines, organisation, process etc…), leur environnement externe (type de clients, parcours clients, image de l'entreprise…) et à anticiper les phénomènes qui s'y rattachent. Les données deviennent actuellement une grande richesse si elles sont bien exploitées. C'est justement l'objectif du Data Mining.

A la frontière entre les statistiques, l'intelligence artificielle et l'informatique, le Data Mining – ou fouille de données – est une discipline qui vise à extraire les informations pertinentes d'un grand ensemble de données. Le principe réside dans la réussite dans la préparation, manipulation et l'analyse les données dans le but de les transformer en connaissance actionnable et en outil d'aide à la décision pour les entreprises.

L'apport du Data Mining pour les entreprises est d'une grande utilité, il touche en particulier :

- La stratégie marketing différenciée par types de clients grâce à l'élaboration d'une segmentation comportementale,
- L'optimisation de l'efficacité des actions marketing et commerciales grâce à une segmentation stratégique,
- L'efficacité accrue des campagnes marketing : e-mailing, sms réseaux sociaux… grâce au ciblage des clients à fort potentiel,
- L'investissement commercial adapté et optimisé grâce à une prédiction du potentiel de vente par zone géographique,

Le diagnostic de la relation clients grâce à une analyse Text Mining de posts sur les réseaux sociaux.

1.4.1. Des algorithmes de Data Mining

Beaucoup d'algorithmes issus de l'apprentissage artificiel permettent de réaliser des projets de Data Mining. La plus grande distinction pouvant être faite entre ces algorithmes se situe dans leur finalité. Il s'agit soit d'identifier sans a priori des similitudes ou des comportements analogues entre les individus ou les clients ; soit d'établir un modèle permettant de les classer dans des groupes bien déterminés.

Dans le premier cas, il s'agit des méthodes d'analyses de type non supervisé. Elles concernent particulièrement les méthodes descriptives ou exploratoires telles que l'analyse factorielle ou le cas des données non annotées. Dans le second cas nous parlons d'analyses supervisées et sont alors mises en œuvre dans des méthodes de prédiction appartenant souvent aux méthodes dites intelligentes, elles ont la caractéristique de pouvoir apprendre et donc de s'adapter et d'ajuster leurs paramètres internes.

1.4.2. Des outils indispensables à la fouille de données

Ils existent plusieurs techniques d'analyse de données dans la littérature. Ces dernières sont implémentées sous forme de logiciels libres ou payants, l'essentiel est de mettre en œuvre une technique ou un outil simple, optimisé (temps et architecture) et transparente.

Les caractéristiques majeures de ces outils résident dans :

- La facilité d'utilisation.
- Le temps de traitement nécessaire à l'exécution des algorithmes.
- La possibilité d'interagir avec un environnement Big Data.
- Les capacités à pouvoir préparer les données de manière simple et rapide (Data Management).

Faciliter à la restitution des résultats de manière visuelle et facilement interprétable (Data Visualisation).

Les données démographiques et géographiques issues des réseaux sociaux permettent de mieux identifier les clients, ou du moins les personnes intéressées par la marque.

Dans le cas de Twitter et Facebook, nous pouvons apprendre beaucoup de choses sur les utilisateurs. Les Big Social Data fournissent des données générales (âge, sexe, localisation, …) comme des données personnelles beaucoup plus précises (préférences politiques,

religieuses, revenus, …). Grâce à cela, nous identifions mieux ces utilisateurs et nous agissons selon l'analyse de ces données.

1.4.3. Mesurer l'efficacité d'une stratégie Sociale Media

Actuellement, les données personnelles des utilisateurs de réseaux sociaux ne sont pas les seules à être exploitables mais le Big Social Data permet aussi de tenir compte toutes les conversations qui se sont déroulées sur les réseaux sociaux. Les commentaires, positifs comme négatifs, sont des informations précieuses pour les entreprises. Il existe aussi des outils d'analyse conversationnelle capables de déterminer quelles sont les attentes des clients pour pouvoir répondre en conséquence.

De même, le Big Social Data a fait émerger des outils de crowd innovation (innovation collaborative). L'idée est d'inclure le consommateur, à travers les réseaux sociaux, dans la création d'offres qui répondent à ses attentes.

Le Big Social Data est d'un grand intérêt pour les entreprises, Il faut juste savoir sélectionner les données pertinentes parmi cette masse de données existantes sur les réseaux sociaux. L'objectif visé est de savoir tirer la connaissance à partir de la donnée.

1.5. Langage figuratif

1.5.1. Présentation

Notons que le langage figuratif change le sens propre de la phrase pour lui donner un sens figuré. Ils existent plusieurs types de langage figuratif, nous citons en particulier, la métaphore, l'ironie, le sarcasme, la satire et l'humour.

L'ironie est définie en général comme une figure de rhétorique par laquelle on dit le contraire de ce qu'on veut faire comprendre, nous citons ci-dessous à titre d'exemple deux tweets ironiques :

- Au bled si tu respectes le code t'est foutu
- Dans le métro une fille à côté de moi n'arrête pas de tousser…smah binatna #corona

L'ironie est un phénomène complexe largement étudié depuis plusieurs décennies, notamment dans les domaines de philosophie et de linguistique (GRICE H. P., COLE P., & MORGAN J. L., 1975; SPERBER D. & WILSON D., 1981; UTSUMI A., 1996).

Du point de vue linguistique computationnelle, l'ironie est un terme générique. Nous l'utilisons pour indiquer un ensemble de phénomènes figuratifs comme le sarcasme (CLIFT R., 1999). La plupart des travaux de littérature en détection de l'ironie en traitement automatique du langage concerne des corpus de tweets car les auteurs peuvent explicitement indiquer le caractère ironique de leurs messages en employant des hashtags spécifiques, comme #sarcasme, #ironie, #humour. Ces hashtags sont alors utilisés pour collecter un corpus annoté manuellement, ressource indispensable pour la classification supervisée de tweets comme ironiques ou non ironiques. Malheureusement ce n'est pas le cas dans le cadre des tweets algériens. Dans ce contexte la littérature présentent globalement des tweets en anglais, mais des travaux existent également pour la détection de l'ironie et/ou du sarcasme pour l'italien, le chinois ou encore le néerlandais (FARIAS D. I. H., SULIS E., PATTI V, RUFFO G., & BOSCO C., 2015; JIE TANG Y. & CHEN H.-H., 2014; LIEBRECHT C., KUNNEMAN F., & VAN DEN B. A., 2013).

En général, les approches qui ont été proposées reposent presque exclusivement sur l'exploitation du contenu linguistique du tweet. Deux principales familles d'indices ont été utilisées :

- **Indices lexicaux :** (n-grammes, nombre de mots, présence de mots d'opinion ou d'expressions d'émotions) et/ou stylistiques (présence d'émoticônes, d'interjections, de citations, usage de l'argot, répétition de mots). (KREUZ R. J. & CAUCCI G. M., 2007; BURFOOT C. & BALDWIN C., 2009; TSUR O., DAVIDOV D., & RAPPOPORT A., 2010; GONZALEZ-IBANEZ R., MURESAN S., & WACHOLDE N., 2011; GIANTI A., BOSCO C., PATTI V., BOLIOLI A., & CARO L. D., 2012; LIEBRECHT C., KUNNEMAN F., & VAN DEN B. A., 2013; REYES A., ROSSO P., & VEALE T., 2013; BARBIERI F. & SAGGION H., 2014b)
- **Indices pragmatiques :** afin de capturer le contexte nécessaire pour inférer l'ironie. Ces indices sont cependant extraits du contenu linguistique du message, comme le changement brusque dans les temps des verbes, l'usage de mots sémantiquement éloignés, ou encore l'utilisation de mots fréquents vs. Mots rares (BURFOOT C. & BALDWIN C., 2009; REYES A., ROSSO P., & VEALE T., 2013; BARBIERI F. & SAGGION H., 2014b).

D'une manière générale, ces approches ont obtenu des résultats prometteurs, mais elles constituent une première étape et qu'il est indispensable de proposer d'autres approches

plus efficaces qui permettent d'inférer le contexte extralinguistique nécessaire à la compréhension de ce phénomène complexe.

1.6. Conclusion

Nous avons expliqué dans ce chapitre, ce phénomène de réseaux sociaux qui est apparue avec l'avènement du web 2.0 comme un moyen de communication entre les individus et aussi un espace pour exprimer les opinions et les avis des personnes. Ensuite nous avons cité les différents types de réseaux sociaux existants à savoir, les réseaux sociaux à dominante visuelle comme Instagram et Pinterest, les réseaux sociaux professionnels comme LinkedIn et Viadeo, les Blogging et micro blogging, où nous citons Twitter et Tumblr et enfin le dernier type de réseaux sociaux qui est Facebook. Nous avons aussi montré l'intérêt des données massives (big data) existantes dans les réseaux sociaux et le besoin à des techniques de fouille de données pour les analyser. Nous avons terminé ce chapitre en expliquant ce phénomène d'ironie sur les réseaux sociaux et le besoin de l'identifier.

Chapitre 2

2. Annotation et techniques d'analyse des données des réseaux sociaux

2.1. Introduction

Depuis l'apparition du web 2.0, les intentions ont été focalisées sur les opinions et les sentiments des utilisateurs des différents réseaux sociaux qui s'y expriment spontanément et en temps réel. Généralement cette immense quantité de données qui reflètent les opinions des internautes est manipulable avec des outils du webmining. Nous notons que cette collection d'informations est constamment mise à jour.

Actuellement plusieurs sites se sont spécialisés dans la collecte de ces opinions dans divers domaines (politique, sport, santé, achat de produit etc…) et les utilisateurs de l'étoile ont pris l'habitude de consulter les avis et commentaires déposés par les autres dès qu'il s'agit de prendre une décision bien spécifique (des élections des candidats pour un poste politique, solution pour lutter contre un virus, achat pour un produit, ou encore pour une réservation d'avion ou d'hôtel). Les avis et les opinions des réseaux sociaux intéressent deux catégories d'utilisateurs :

Les internautes d'une manière générale qui ont suscité des applications et services multiples, ce qui provoque un espace virtuel d'encouragement à donner son avis et même à se faire reconnaître comme donnant des avis pertinents et suivis par les autres.

Des propriétaires de la marque du produit et des bureaux d'études qui tentent de découvrir ce sentiment de consensus de la majorité. Souvent sensibles aux avis des fois imaginaires sachant qu'une réputation peut être détruite à cause d'un tweet. Le dilemme les marques se soucient de leur identité en ligne mais cherchent également à mieux connaître les attentes et critiques que les internautes leur adressent. D'où le besoin croissant de développer des techniques pour capter ces évaluations des internautes, allant du simple dénombrement de

tweets positifs ou négatifs à l'analyse plus détaillée des contenus de ces tweets en faisant appel aux techniques d'analyse de données ou du datamining.

2.2. Différence entre sentiment et opinion

Souvent les gens confondent entre les deux termes sentiments et opinions, dans cette section nous essayons de présenter d'une manière succincte les descriptions principales de ces deux termes.

2.2.1. Principe du sentiment

Le sentiment est la composante de l'émotion qui implique les fonctions cognitives de l'organisme, la manière d'apprécier. Le sentiment est à l'origine d'une connaissance immédiate ou d'une simple impression. Il renvoie à la perception de l'état physiologique du moment. Le sens psychologique de sentiment qui comprend un état affectif est à distinguer du sens propre de la sensibilité (Wikipedia, Sentiment, 2020). Le dictionnaire Larousse de Poche 2017 définit le sentiment comme étant un état affectif complexe et durable lié à certaines émotions ou représentations.

2.2.2. Principe d'opinion

Une opinion c'est un Jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense (Larousse, s.d.).

Une opinion comme un quintuple, (Ei, Aij, Sijkl, Hk, Tl), où Ei est le nom d'une entité, Aij est un aspect de Ei, Sijkl est le sentiment sur l'aspect Aij de l'entité Ei, Hk est détenteur de l'opinion, et Tl le temps où l'opinion a été exprimée par Hk. Sijkl est positive, négative ou neutre, ou peut être exprimé par des niveaux d'intensité (Liu, 2012).

2.3. Aperçu historique sur les Opinions

Dans le cadre de ce mémoire nous ciblons l'opinion des internautes. L'importance donner au phénomène d'opinions est née avant l'apparition de l'internet. Notons que ce que pensent et disent les gens est un sujet d'intérêt scientifique depuis au moins un siècle, politique depuis plus d'un demi-siècle et, dernièrement, un sujet d'intérêt commercial.

Nous sindons cette évolution en trois périodes distinctes :

2.3.1. Opinions basées sur la rumeur

Dès le début du vingtième siècle, plusieurs chercheurs (Stern W., 1902; Allport G.-W. & Postman L. J. , 1945). S'intéressent au phénomène de la rumeur et tentent de la caractériser et la formaliser. Via le principe de bouche à oreille, les informations peuvent être modifiées (réduites, simplifiées ou transformés). Les contenus peuvent être valorisés au détriment d'autres en fonction de la sensibilité ou du contexte social de l'auteur de message original. Dans le cas d'une rumeur crédible, cette dernière devient un phénomène identifiable, analysable et mesurable. Ensuite cette rumeur sera pris en charge par des experts dans la psychologie sociale, la psychanalyse et la sociologie (Morin E., 1969).Un débat est mené sur son caractère fondé (réelle information) ou non (simple projection de fantasme), et elle est entrée désormais dans le champ des études médiatiques et du management des marques. (Kapferer J.-N., 1987)

2.3.2. Opinions basées sur les sondages

Les sondages sont apparus depuis la fin de la seconde guerre mondiale. Partant du principe qu'il existe une opinion publique, les sondages consistent à créer des méthodes et des indicateurs pour la mesurer. Mais le premier inconvénient majeur des sondages réside dans le fait que nous obligeons les personnes interrogées à se poser des questions qu'elles n'ont peut-être jamais eu l'occasion de se poser ou à avoir une opinion sur un sujet qu'elles ignorent complétement. Le second inconvénient concerne leurs méthodes et notamment la prétention de représentativité, réalisée sous forme de quotas ou sous forme d'échantillons aléatoires, ce qui vaut là encore des critiques aux méthodes existantes. Mais par leur répétition, on peut considérer que les sondages détectent quelque chose des mouvements d'opinion.

2.3.3. Opinion mining

Les approches qui concernent l'opinion mining reprennent les mêmes principes cités auparavant sans s'interroger sur le statut de cette opinion, considérant que, à force de sondages, elle a fini par exister. Ces approches sont donc bien éloignées d'un souci de représentativité de la population, établie en fonction de critères sociodémographiques. Les études d'opinion ont la capacité à corréler tous les attributs et opinions des personnes sondées à leurs propriétés socio-démographiques. Ce lien personnel permet ensuite des agrégats qui prétendront à la validité statistique par rapport à une population de référence. Cependant, il sera tenu compte des intervalles de confiance, selon la taille de l'échantillon rapportée à celle

de la population, qui devraient relativiser les résultats mais qui sont très souvent oubliés dans les publications des médias, ce qui entraîne des confusions et des critiques. En réalité, bien d'autres opérations sont nécessaires pour maintenir cet effet de représentativité face à tous les problèmes de constitution de l'échantillon ou de récupération/exploitation des données : mais tous les redressements qui font l'art des sondeurs professionnels sont rarement explicités car la demande sociale/médiatique n'attend pas cette garantie scientifique mais seulement son approximation. C'est la même posture qui justifiera les approximations innombrables de la plupart des offres en opinion mining et sentiment analysis.

2.4. Collecte des données

2.4.1. Les caractéristiques majeures des données extraites des réseaux sociaux

Le matériau linguistique des commentaires recueillis sur les réseaux sociaux relève de plusieurs types d'énoncés :

- Un jugement
- Une évaluation
- Une opinion
- Un avis
- Un sentiment
- Un goût
- Un récit d'expérience
- Un récit de pratique

Tous ces types d'énoncés possèdent des statuts différents mais se retrouvent mêlés dans toute collecte de données.

2.4.2. Changement d'échelle des données

Aujourd'hui les données issues des réseaux sociaux sont des opinions, des avis, des commentaires et elles sont produites en grande quantité, par contre, dans le passé les contributions du public ciblé étaient limitées au courrier des lecteurs ou aux sondages et enquêtes qui sont en général fastidieuses et couteuses. Dès lors, ce changement d'échelle est comme toujours une augmentation (Eisenstein E. L. , 1991) potentielle tout à la fois :

- Des tendances participatives de la démocratie.
- De la visibilité d'opinions jusqu'ici marginales.

- De la puissance d'influence sur les esprits de tous les médias relayés par le public lui-même.
- De la focalisation collective immédiate sur des thèmes qui sont répliqués à grande vitesse.
- De la réflexivité d'une société sur son propre climat, son humeur (mood), ce qui n'était pas encore arrivé.
- D'une possibilité de traiter statistiquement des informations subjectives même non catégorisées ou indexées.
- Des outils de monitoring qui sont autant de tableaux de bord permanents de l'opinion.
- Des comparaisons entre types de données, émetteurs, propriétés des réseaux, des flux, etc. qui ouvrent des pistes infinies, d'autant plus qu'on ne sait pas vraiment ce qu'on pourrait y chercher.

2.4.3. Vers une automatisation de la collecte du corpus

Actuellement Les données issues des réseaux sociaux sont volumineuses et variées, une tendance à automatiser au maximum les procédures de collecte, de stockage, de constitution des bases de données, est imposée et qui donne une valeur ajoutée pour les masses de données disponibles. Aussi beaucoup de chercheurs dans la littérature travaillent sur le traitement linguistique automatisé de ces données, même si les résultats issus de ces travaux restent jusqu' à présent insuffisants. L'objectif principal des sociétés prestataires de service derrière l'automatisation de l'analyse elle-même, est de fournir des outils de pilotage direct par les clients qui désirent avoir de l'opinion mining.

2.4.4. Les APIs comme moyen d'accès aux données des réseaux sociaux

Pour faire face aux risques de blocage d'accès aux données issues des réseaux sociaux par des robots inconnus, plusieurs entreprises ont fait appel aux développeurs pour créer des applications basées sur les APIs (Application Programming Interfaces). A titre d'exemple une partie des données sous Twitter peut être réexploitée par des développeurs qui créent des APIs pour extraire les données de ce service (Twitter). Les données personnelles sur les réseaux sociaux sont devenues une opportunité qui fait l'objet de beaucoup de convoitises.

2.4.5. Les différents APIs dédiés à l'extraction des données des réseaux sociaux

Il existe plusieurs APIs pour extraire les données des réseaux sociaux. Les sociétés gérantes de ces derniers ont bien compris le besoin de fournir des services permettant d'extraire les

données générées, et proposent parfois même des outils pour exploiter et traiter automatiquement les données ciblées.

Dans cette section, nous nous concentrons sur les outils d'extraction permettant de collecter les données provenant de Twitter, Facebook et des blogs.

Twitter met en œuvre plusieurs plateformes (APIs REST), qui prennent en paramètre une requête et renvoient une réponse au format *JSON*. Le tout est accessible selon trois offres : STANDARD (gratuit), ENTERPRISE et PREMIUM. Dans chaque API, il est proposé plusieurs endpoints. Il en existe assez pour répondre à énormément de cas d'utilisation (stream, publier des tweets, récupérer les tendances etc.), nous présentons d'une manière succincte les outils permettant l'extraction (API Search Tweets, API Get Tweets Timelines) et le streaming : (Filter Realtime Tweets).

- **L'API Search Tweets :** Elle permet d'extraire des publications selon une recherche particulière, c'est-à-dire selon un ou plusieurs mots-clés, selon des hashtags/noms d'utilisateur ou encore sur une période donnée. Elle prend une requête de recherche et renvoie un JSON avec les données des publications correspondantes.
  Pour chaque publication, l'objet retourné contient un panel exhaustif d'informations. Outre le contenu du tweet, on retrouve diverses précisions sur l'auteur, la localisation, l'appareil utilisé pour la publication, sur le nombre de retweets/likes/abonnés, les mentions, les hashtags, les médias contenus et plus encore.
  Notons que l'API STANDARD permet d'effectuer jusqu'à 400 requêtes sur une fenêtre de 15 minutes. Sachant qu'une requête renvoie 100 publications maximum. Outre le nombre de requêtes maximum, la différence entre le service STANDARD et PREMIUM se trouve dans l'accès aux anciens tweets. L'offre gratuite ne permet de récupérer que les tweets récents (date de publication inférieure à 7 jours) alors que la version payante donne accès aux publications depuis 2006. A noter aussi que la recherche de l'offre gratuite n'est pas entièrement *fidèle* ni exhaustive, elle peut ignorer des tweets qu'elle ne considère pas pertinents. Mais malgré ces limitations avec une bonne utilisation des paramètres et de l'outil (mise en cache, curseur de page etc.), nous sommes quand même capables d'extraire gratuitement des informations sur des milliers de tweets en un temps record.
- **Stream API :** Twitter propose également un service de Stream en temps réel sur une requête donnée. Aussi L' API Twitter Streaming renvoie les mises à jour publiques de

l'état Twitter en filtrant les expressions de recherche, les ID utilisateur et par emplacement. L'API Twitter Streaming renvoie des statuts publics qui correspondent à un ou plusieurs prédicats de filtre. Plusieurs paramètres peuvent être spécifiés, ce qui permet à la plupart des clients d'utiliser une seule connexion à l'API Streaming.

- **Facebook Graph API :** L'API Facebook Graph nous permet de récupérer les données Facebook. Sachant que sur Facebook, les données sont protégées, dans le sens où il y a beaucoup plus de notions de publications, groupes, et profils privés que sur Twitter, où la grande majorité des tweets est publique. L'accès aux données est donc davantage contrôlé.

  Aussi, Facebook propose plusieurs services de communication et de partage (vidéos, groupes, pages, etc.), ce qui complique conceptuellement la donnée. Alors API Facebook Graph semble mieux adapté et bien organisé pour faire face à cette situation.
- **Scraping :** Le fait que les blogs ne sont pas fournis par un seul et même service (WordPress, Wix, etc.). L'extraction de données de ce dernier diffère de celle des autres réseaux sociaux. Aujourd'hui, la récupération automatisée des documents de blog se fait via l'API scraping. Nous notons aussi qu'il existe d'autres services ready-*to-use* pour scraper des blogs et sites en général comme Les produits payants (APIFY, SCRAPERAPI).

2.5. Les différentes méthodes d'analyses des données

L'état de l'art dans le domaine de l'analyse des données issues des réseaux sociaux est d'actualité et attire l'intention de plusieurs chercheurs dans le monde, les techniques et les méthodes adoptées sont aussi très variées, c'est pourquoi dans cette partie nous présentons quelques techniques utilisées dans la littérature pour la classification des opinions par polarité (positive, négative ou neutre) ou pour l'identification de l'ironie. Les grandes classes de méthodes utilisées sont :

- Méthode basée sur la classification automatique.
- Méthode basée lexique.
- Méthode combinant les deux méthodes précédentes

2.5.1. Méthode lexicale

Cette méthode utilise les outils du traitement automatique de la langue comme l'analyseur syntaxique, l'analyseur morphologique et l'analyseur sémantique, etc., et consiste aussi à

définir un ensemble de règles lexico-syntaxiques qui décrivent une expression régulière, formée de mots et de catégories grammaticales. Ces règles sont souvent porteuses d'un ou plusieurs marqueurs linguistiques (GHERSEDINE, BUCHE, DIBIE-BARTH EL EMY, HERNANDEZ, & KAMEL, 2012)

La méthode lexicale est basée sur un système d'extraction d'information, elle est fondée sur une analyse syntaxique du texte à l'aide d'un analyseur syntaxico-sémantique. Ce dernier contient un lexique de mots et utilise des règles de la grammaire pour construire des lexiques (ou des dictionnaires) d'opinions sur lesquels réagissent les règles. Notons que l'analyse du texte est faite phrase par phrase. Pour cette méthode, il est supposé que le corpus ne soit pas annoté au début. Donc il faut construire tout d'abord le vocabulaire de mots d'opinion initial, et utiliser des méthodes pour l'enrichir. Il existe trois méthodes pour construire des lexiques (dictionnaires d'opinions) lié à la méthode lexicale :

- Construction manuelle.
- Construction à base des dictionnaires.
- Construction automatique en utilisant des corpus.

2.5.1.1. La méthode manuelle

Cette méthode consiste à enrichir le lexique de mots d'opinions sans faire appel à aucun outil particulier, seulement les experts font la sélection de mots et expressions porteurs d'opinions. Cet ensemble de mots est appelé graine ou germe (en anglais, seedwords) construire une première liste de mots et d'expressions utilisée par la suite à trouver, répertorier et classer d'autres mots et expressions porteurs d'opinions. Mais nous notons que cette méthode est fastidieuse et consomme beaucoup de temps. Généralement cette méthode est utilisée conjointement avec d'autres approches dites automatisées telles que l'approche basée sur un dictionnaire ou sur le corpus. Elle peut être aussi exécutée après des approches à base d'apprentissage (décrits dans les sections suivantes) afin de vérifier les résultats et corriger les erreurs probables.

2.5.1.2. Méthode basée dictionnaire

Cette partie consiste à établir un ensemble de dictionnaires comportant des mots d'opinion avec la valeur qu'ils expriment suivis de leurs synonymes et antonymes.

L'utilisation de ces dictionnaires ne permet néanmoins pas de traitements orientés vers des contextes spécifiques (Vaithyanathan B. P., Sentiment classification using machine learning techniques., 2002).Il existe des outils permettant d'identifier le sentiment extrait d'un texte. Nous citons ci-dessous une liste contenant les outils les plus utilisés :

- Werfamous : outil d'analyse en ligne gratuit, donnant un score de sentiment sur une échelle de -100 à 100, ainsi qu'un niveau de confiance lié à ce score.
- AFINN : évalue-la positivité/négativité d'un mot à l'aide d'un dictionnaire contenu dans une archive.
- SenticNet : L'objectif principal de SenticNet est de rendre l'information conceptuelle et affective véhiculée par le langage naturel (destiné à la consommation humaine) plus facilement accessible aux machines (SenticNet, 2012)
- WordNet : permet de savoir à l'aide de groupe de synonymes si un mot est positif ou non.
- SentiWordNet : il s'agit d'une extension à WordNet ; il attribue à chaque groupe de synonymes provenant de WordNet, trois scores de sentiment : la positivité, la négativité, l'objectivité.
- SentiSense (Gerv´as J. C., 2014): il s'agit également d'un travail basé sur WordNet permettant de polariser les mots de façon plus précise.

2.5.1.3. Méthode basée corpus

Cette méthode consiste à établir un ensemble de dictionnaires comportant des mots d'opinion avec la valeur qu'ils expriment suivis de leurs synonymes et antonymes. L'utilisation de ces dictionnaires ne permet néanmoins pas de traitements orientés vers des contextes spécifiques (Vaithyanathan B. P., sentiment classification using machine learning techniques. Stroudsburg, 2002). Nous présentons ci-dessous la différence principale existante entre la méthode basée corpus et la méthode basée lexique.

Pour la première méthode, le corpus constitut un grand corps de texte en langage naturel, il est utilisé pour accumuler des statistiques sur le texte en langage naturel.

Aussi les corpus incluent souvent des informations supplémentaires comme l'annotation pour chaque mot et l'arbre d'analyse pour chaque phrase.

Par contre 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 contient d'autres informations

sur les rôles joués par le mot, tels que les informations sur les caractéristiques comme le cas d'un verbe :

- Il est transitif, ou intransitif etc.
- Participe présent, passé, etc…

Nous avons remarqué que malgré que la méthode lexicale suscite beaucoup d'intérêts dans l'analyse des données issues des réseaux sociaux mais néanmoins elle présente quelques limites citées ci-dessous :

- Les dictionnaires affectent une tonalité positive ou négative à un mot, sans tenir compte du contexte, c'est-à-dire du texte environnant, comme des paramètres de la communication située ;
- Les dictionnaires ont tendance à éliminer les mots à valence ambiguë a priori ;
- Le traitement des expressions ambiguës reste à faire et demande de faire appel à d'autres principes et à d'autres techniques ;
- Lorsque la négation n'est pas prise en compte (ce qui peut paraître étonnant mais qui existe encore, par exemple dans les méthodes basées sur les sacs de mots, qui calculent des fréquences d'occurrence dans un texte), le score de polarité peut être complètement faussé ;
- Ces dictionnaires ne permettent pas de traiter des figures de rhétorique qui peuvent changer entièrement la valence des expressions (le sarcasme, l'ironie).
  D'où le besoin de l'hybrider avec une méthode dite intelligente ou à base d'apprentissage.

2.5.2. Classification automatique

La classification automatique est une branche du grand domaine de la reconnaissance de formes. Elle est basée sur le principe de l'apprentissage artificiel. Ce dernier est inspiré de l'apprentissage naturel de l'être vivant. Ce mode de classification à base d'apprentissage a été largement utilisé dans divers domaines d'application entre autres la classification des opinions et des sentiments issues des différents réseaux sociaux.

2.5.2.1. Définition de l'apprentissage artificiel

L'apprentissage automatique (Machine Learning), apprentissage artificiel ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d' apprendre à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes (Wikipedia, Apprentissage automatique, 2020).

Nous distinguons trois types d'apprentissage : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage semi-supervisé.

2.5.2.2. L'apprentissage supervisé

L'apprentissage supervisé consiste à construire un modèle basé sur un jeu de données d'apprentissage et des étiquettes (nom de la classe) et le tester par la suite sur des nouvelles données (Ribeiro C. S., Inductive infeTence foT large scale text classification, 2010).

Nous présentons ci-dessous les algorithmes les plus utilisés dans le cas d'apprentissage supervisé :

- **Machine à vecteurs de support (SVM) :** Il s'agit d'un ensemble de techniques destinées à résoudre des problèmes de discrimination (prédiction d'appartenance à des groupes prédéfinis) et de régression (analyse de la relation d'une variable par rapport à d'autres) (Ribeiro C. S., Inductive infeTence foT large scale text classification, 2010). Quelques travaux dans la littérature ont montré l'efficacité des techniques de SVM dans l'analyse des sentiments et des opinions sur les réseaux sociaux. Nous citons dans ce cadre les travaux de Fang en 2015 (El-Beltagy S. R., 2006), la technique de SVM utilisée sur l'ensemble de données d'apprentissage des microblogs a amélioré nettement le taux de classification allant de 61% à 94%. Aussi Munir. A and al ont utilisé deux bases de données de sentiments distinctes (self-driving cars et apple), en appliquant la méthode SVM, ils ont déduit l'existence d'une grande dépendance entre les performances obtenues par SVM et la qualité des données des bases d'apprentissage utilisées. Ils ont obtenu un taux de précision de 59.91% pour la première base et un taux de 71.20% pour la deuxième base (Munir A., Shabib A., & Iftikhar A., 2017).
- **Réseaux de neurones (R.N)** : Un réseau de neurones artificiels, ou réseau neuronal artificiel, est un système dont la conception est à l'origine schématiquement inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques. Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage de type probabiliste, en particulier bayésien. Ils sont placés d'une part dans la famille des applications statistiques, qu'ils enrichissent avec un ensemble de paradigmes permettant de créer des classifications rapides (réseaux de Kohonen en particulier), et d'autre part dans la famille des méthodes de l'intelligence artificielle auxquelles ils fournissent un mécanisme perceptif indépendant des idées

propres de son concepteur, et fournissant des informations d'entrée
au raisonnement logique formel (Wikipedia, Réseau de neurones artificiels, 2020).
Plusieurs travaux dans l'état de l'art ont fait appels aux approches neuronales pour analyser les sentiments sur les réseaux sociaux (Ahmed Sulaiman M.Al, 2019) ont proposé un réseau neuronal convolutif (CNN) qui intègre des informations sur le comportement des utilisateurs dans un document donné (tweet), le réseau neuronal utilisé est évalué sur deux ensembles de données fournis par l'atelier SemEval-2016. Le modèle proposé surpasse les modèles de base actuels y compris Naïve Bayes et Support Vector Machines), ce qui montre que le fait d'aller au-delà du contenu d'un document (tweet) est bénéfique pour la classification des sentiments, car il fournit au classificateur une compréhension approfondie de la tâche.
(Attardi G., Sartiano D., & Alzetta C.) Ont réalisé une analyse des sentiments sur une base de tweets italienne. Ils ont utilisé des réseaux neuronaux convolutionnels dans la classification du sentiment twitter. Ils ont divisé leur approche en trois parties séparées.
La première concerne l'identification de la subjectivité des tweets, la deuxième cible la classification de polarité i.e. classer un tweet comme positif, négatif, neutre ou mixte (un tweet avec un sentiment positif et négatif) et la dernière tâche permet la détection d'ironie.

- **Naïve Bayes (N.B)** : est un type de classifieur bayésien probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses. Le classifieur bayésien naïf, ou classifieur naïf de Bayes, appartenant à la famille des classifieurs linéaires.
  Un classifieur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques.
  Les classifieurs bayésiens naïfs peuvent être entraînés efficacement dans un contexte d'apprentissage supervisé. Dans beaucoup d'applications pratiques, l'estimation des paramètres pour les modèles bayésiens naïfs repose sur le maximum de vraisemblance. Autrement dit, il est possible de travailler avec le modèle bayésien naïf sans se préoccuper de probabilité bayésienne ou utiliser les méthodes bayésiennes. Les classifieurs bayésiens naïfs ont fait preuve d'une efficacité plus que suffisante dans beaucoup de situations réelles complexes. L'avantage du classifieur bayésien naïf est qu'il requiert relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification, à savoir moyennes et variances des différentes

variables. En effet, l'hypothèse d'indépendance des variables permet de se contenter de la variance de chacune d'entre elles pour chaque classe, sans avoir à calculer de matrice de covariance. (Wikipedia, Classification naïve bayésienne, 2020)

- **Forêts d'arbres décisionnels (Random Forest)** : L'apprentissage par arbre de décision désigne une méthode basée sur l'utilisation d'un arbre de décision comme modèle prédictif. On l'utilise notamment en fouille de données et en apprentissage automatique.

  Dans ces structures d'arbre, les feuilles représentent les valeurs de la variable-cible et les embranchements correspondent à des combinaisons de variables d'entrée qui mènent à ces valeurs. En analyse de décision, un arbre de décision peut être utilisé pour représenter de manière explicite les décisions réalisées et les processus qui les amènent. En apprentissage et en fouille de données, un arbre de décision décrit les données mais pas les décisions elles-mêmes, l'arbre serait utilisé comme point de départ au processus de décision.

  C'est une technique d'apprentissage supervisé : on utilise un ensemble de données pour lesquelles on connaît la valeur de la variable-cible afin de construire l'arbre (données dites étiquetées), puis on extrapole les résultats à l'ensemble des données de test. (Wikipedia, Arbre de décision (apprentissage) , 2019).

  (Dubey K. & Agrawal S., 2018) Ont fait appel à la technique de Random Forest. Pour analyser une base d'opinions sous twitter sur le film Civil War. Les performances obtenues ont été très prometteuses.

2.5.2.3. Apprentissage non supervisé

Dans le domaine informatique et de l'intelligence artificielle, l'apprentissage non supervisé désigne la situation d'apprentissage automatique où les données ne sont pas étiquetées. Il s'agit donc de découvrir les structures sous-jacentes à ces données non étiquetées. Puisque les données ne sont pas annotées, il est impossible à l'algorithme de calculer de façon certaine un score de réussite. L'absence d'étiquetage ou d'annotation caractérise les tâches d'apprentissage non-supervisé et les distingue donc des tâches d'apprentissage supervisé. En général, des systèmes d'apprentissage non supervisé permettent d'exécuter des tâches plus complexes que les systèmes d'apprentissage supervisé, mais ils peuvent aussi être plus imprévisibles.

K-moyennes (KMeans) : Le partitionnement en k-moyennes (ou k-means en anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier k, le problème est de diviser les points en k groupes, souvent appelés clusters, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.

Il existe une heuristique classique pour ce problème, souvent appelée méthodes des k-moyennes, utilisée pour la plupart des applications. Le problème est aussi étudié comme un problème d'optimisation classique, avec par exemple des algorithmes d'approximation.

Les k-moyennes sont notamment utilisées en apprentissage non supervisé où l'on divise des observations en k partitions. Les nuées dynamiques sont une généralisation de ce principe, pour laquelle chaque partition est représentée par un noyau pouvant être plus complexe qu'une moyenne. Un algorithme classique de k-means est le même que l'algorithme de quantification de Lloyd-Max. (Wikipedia, K-moyennes, 2019)

- **Regroupement hiérarchique :** Dans l'exploration de données et les statistiques , le clustering hiérarchique (également appelé analyse de cluster hiérarchique ou HCA) est une méthode d' analyse de cluster qui cherche à construire une hiérarchie de clusters. Les stratégies de regroupement hiérarchique se divisent généralement en deux types :
    - Agglomération : il s'agit d'une approche ascendante : chaque observation commence dans son propre cluster, et des paires de clusters sont fusionnées au fur et à mesure que l'on monte dans la hiérarchie.
    - Diviseur : Il s'agit d'une approche descendante : toutes les observations commencent dans un seul cluster et les divisions sont effectuées de manière récursive à mesure que l'on descend dans la hiérarchie. En général, les fusions et les scissions sont déterminées de manière gourmande . Les résultats du clustering hiérarchique sont généralement présentés dans un dendrogramme .

  Georgiana I. and al. Ont présenté une méthode de détection des sujets dans les flux Twitter, basée sur le filtrage tweet/terme. Ils ont adopté un regroupement hiérarchique scindé en deux étapes, la première concerne le regroupement des tweets et le second concerne le regroupement des titres résultants de la première étape de clustering. Les résultats obtenus semblent encourageants et prometteurs, beaucoup d’entre eux étant publiés comme nouvelles dans les médias d’information traditionnels. Le point fort de

cette méthode est que l'utilisateur peut retracer la piste de retour à son tweet original. (Georgiana Ifrim, Bichen Shi, & Igor Brigadir)

2.5.2.4. 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. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non supervisé qui n'utilise que des données non étiquetées. Il a été démontré que l'utilisation de données non étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage.

Un autre intérêt provient du fait que l'étiquetage de données nécessite souvent l'intervention d'un utilisateur humain. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse et couteuses. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident.

Un exemple d'apprentissage semi-supervisé est l'auto-apprentissage, dans lequel un classifieur est entrainé sur une petite base de données annotées ensuite nous prenons un nouveau exemple non annotée et nous l'annotons avec le classifieur, nous l'ajoutons après à la petite base de données annotée, nous entrainons une autre fois le classifieur sur la nouvelle base de données annotées. Nous continuons cette procédure jusqu'à ce que toutes les données non annotées sont utilisées. (Wikipedia, Apprentissage semi-supervisé, 2019) (Apprentissage artificiel concepts et algorithmes, Antoine Cornuéjols et Laurent Miclet 2013)

Samah M. Alzanin and al ont abordé le problème de la détection des rumeurs dans les tweets arabes, sachant que la détection des rumeurs dans les réseaux sociaux en langue arabe a pris du retard par rapport aux travaux sur d'autres langues, en particulier en anglais. Les auteurs ont utilisé un ensemble de fonctionnalités extraites de l'utilisateur et du contenu. Ces caractéristiques ont été analysées pour déterminer leur importance. Une optimisation et des attentes semi-supervisées a été utilisée pour former le système proposé sur des sujets de tweets dignes d'intérêt. Une comparaison avec le Gaussian Naïve Bayes (NB) supervisé a montré que leur système à apprentissage semi-supervisé, utilisant une petite base de données étiquetées, surpasse le Gaussian NB avec une précision de 78,6%. Les performances de l'E-M non supervisé dépendent des valeurs initiales et ils ont atteint un F1 score de 80% dans une de leurs expériences. (Samah M. Alzanin & Aqil M., 2019)

Surendra Sedhai and al. Ont proposé un modèle de détection de spam en mode semi-supervisé (S3D) pour la détection de spam au niveau du tweet. Le modèle proposé se compose de deux modules principaux : un module de détection de spam fonctionnant en temps réel et un module de mise à jour fonctionnant en off line. Le module de détection de spam se compose de quatre détecteurs simples :

- Détecteur de domaine sur liste noire pour étiqueter les tweets contenant des URLs sur liste noire.
- Détecteur pour étiqueter les tweets qui sont presque des doublures de tweets pré-étiquetés avec certitude.
- Un détecteur fiable pour étiqueter les tweets publiés par des utilisateurs avec confiance, ils ne contiennent pas de spams.
- Un détecteur basé sur un classifieur multiple étiquettes des tweets restants. Les informations requises par le module de détection sont mises à jour en mode off line en fonction des tweets étiquetés precedemment. Des expérimentations sur un ensemble de données à grande échelle montrent que le modèle apprend de manière adaptative les nouvelles activités de spam et maintient une bonne précision pour la détection de spam dans un flux de tweet. (Sedhai S. & Sun A., 2017)

Valentina Sintsova and al. Proposent de construire des classifieurs d'émotions avec un minimum de connaissances initiales (par exemple un lexique d'émotions à usage général) et d'utiliser une méthode d'apprentissage semi-supervisée pour l'étendre à d'autres tweets afin de détecter correctement plus de tweets émotionnels dans un domaine spécifique. De plus ont montré que leur algorithme, le vote pondéré équilibré (Balanced Weighted Voting : BWV) est capable de surmonter la distribution déséquilibrée des émotions dans les données étiquetées initiales. Leurs expérimentations de validation montrent que l'algorithme BWV améliore les performances de trois classifieurs initiaux, au moins dans le domaine du sport. De plus, sa comparaison avec les deux autres stratégies d'apprentissage révèle sa supériorité en termes de F1-score, ainsi que des performances sont plus stables entre les différentes catégories d'émotions. (Sintsova V. , Musat C., & Pu P., 2014)

Dmitry Davidov and al. Ont utilisé un algorithme nommé SASI, c'est un algorithme robuste pour l'identification du sarcasme, ils l'ont testé sur une base de données twitter collectée ensuite ils ont comparé ses performances par rapport à une base de données Amazon product reviews. Après une phase d'apprentissage, ils ont pu obtenir des bonnes performances en

termes de précision, rappel et F-Score sur les deux bases de données. (Davidov D., Tsur O., & Rappoport A., 2010)

2.5.3. Méthode d'hybridation

Cette méthode combine entre l'approche basée lexique et l'approche basée apprentissage automatique et tente de corriger l'inconvénient du l'approche basée lexique (indépendance du domaine et du contexte et l'annotation manuelle). L'utilisation de la méthode hybride permet d'annoter automatiquement le corpus d'apprentissage avec la méthode basée lexique, et ensuite entrainer un classifieur sur ce corpus avec une méthode à base de l'apprentissage artificiel. Narayanan et al en 2009 ont effectué une fouille d'opinions au niveau phrases, ils déterminent la polarité des phrases par la méthode basée lexique en utilisant les mots d'opinions positifs et négatifs, ensuite, ils appliquent la technique SVM sur les phrases annotées automatiquement par la méthode basée lexique, cela a donné une précision de 75.6%. Notons qu'Ils ont travaillé sur un corpus de Tweets avec différentes requêtes, Li et Xu en 2011 ont annotés le corpus d'apprentissage avec la méthode basée lexique en utilisant POS (Part Of Speech) et les mots d'opinions, ensuite ils ont entrainé le classifieur sur ce corpus avec la technique SVM. Cette méthode a donné de bons résultats avec une précision de 85.4%. Nous remarquons qu'une partie des défauts de ces approches lexicales peut être corrigée par un enrichissement automatisé en utilisant les méthodes à base d'apprentissage.

2.6. Conclusion

Au début de ce deuxième chapitre, nous avons expliqué la différence entre les opinions et les sentiments. Nous avons aussi donné un aperçu historique sur l'évolution d'utilisation des opinions depuis le début du vingtième siècle en commençant par l'utilisation de la rumeur qui est devenue un phénomène analysable pour être utilisée dans le champ des études médiatiques. Ensuite, les sondages qui sont apparus depuis la fin de la seconde guerre mondiale, ces derniers ont pu aider à détecter des informations pertinentes concernant le mouvement de la société. Actuellement des approches d'opinion-mining sont nées, elles ont la capacité à corréler tous les attributs et opinions des personnes sondées à leurs propriétés socio-démographiques. Les résultats issus de ces approches sont très prometteurs. Nous avons aussi présenté dans ce chapitre le principe de la collecte des données pour constituer des corpus et l'intérêt croissant pour automatiser cette tâche vue la quantité énorme des données disponible sur les réseaux sociaux. A la fin de ce chapitre, nous avons cité les trois méthodes d'analyse des données, lexicale, automatique et hybride. Dans le chapitre suivant, nous

présentons l'essentiel de notre travail de master et qui vise principalement la constitution d'un corpus de tweets algériens.

Chapitre 3

3. Préparation et annotation du corpus

3.1. Introduction

Dans ce chapitre nous présentons l'environnement Anaconda et ses outils. Nous aussi avons présenté respectivement le langage de programmation Python avec ses bibliothèques et l'API twitter. Nos deux contributions majeures sont présentées par la suite avec discussion des résultats obtenus à savoir :

1. La collecte et l'annotation du corpus des tweets algériens
2. Le prétraitement et la classification des tweets (ironiques et non ironiques)

3.2. Les outils utilisés

Dans cette section, nous présentons les outils que nous avons utilisé durant la réalisation de notre projet de fin d'étude de master, nous citons en particulier l'environnement gratuit Anaconda, le langage de programmation Python et l'API Twitter.

3.2.1. Anaconda

Anaconda dans sa version 3.8 est une distribution gratuite et open-source des langages de programmation Python et R pour l'informatique scientifique (science des données, applications d'apprentissage automatique, traitement de données à grande échelle, analyse prédictive, etc.), qui vise à simplifier la gestion et le déploiement des paquets. Les versions de paquet sont gérées par le système de gestion de paquet « conda ». La distribution Anaconda comprend des paquets de data-science adaptés à Windows, Linux et macOS. (Wikipedia, Anaconda, 2020)

3.2.2. Python

Python dans sa version 3.8 est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée

objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire par ramasse-miettes et d'un système de gestion d'exceptions ; il est ainsi similaire à Perl, Ruby, Scheme, Smalltalk et Tcl.

Le langage Python est placé sous une licence libre proche de la licence BSD et fonctionne sur la plupart des plates-formes informatiques, des smartphones aux ordinateurs centraux, de Windows à Unix avec notamment GNU/Linux en passant par macOS, ou encore Android, iOS, et peut aussi être traduit en Java ou .NET. Il est conçu pour optimiser la productivité des programmeurs en offrant des outils de haut niveau et une syntaxe simple à utiliser. (Wikipedia, Python, 2020)

3.2.3. Twitter API

Twitter est le site de contenus le plus fréquemment mis à jour ; environ 500 millions de messages sont postés quotidiennement sur sa plateforme. Twitter met à disposition des développeurs des APIs publiques, simples d'utilisation qui permettent la collecte, le traitement et l'analyse de tweets.

Twitter dispose de plusieurs APIs permettant de requêter sa base de données, mais aussi de construire des services au-dessus de sa plateforme. Ces APIs sont particulièrement riches en retournant presque une centaine de variables par requête ; les données concernent les tweets (date de publication, le texte du message, etc.), l'auteur (date de création du compte, pseudo…), les entités contenues dans les messages (hashtags, mentions, urls…) et des informations de localisation (pays, timezone, longitude / latitude). (Twitter, 2020).

3.2.4. Bibliothèques

Nous avons utilisé les quatorze bibliothèques suivantes :

- **Comma-separated values** (CSV) : Le module Csv met en œuvre des classes pour lire et écrire des données tabulaires en format Csv. Il permet aux programmeurs d'écrire ces données dans le format préféré par Excel, ou de lire les données de ce fichier qui a été généré par Excel, sans tenir compte des détails bien précis du format Csv utilisé par Excel. Les programmeurs peuvent également décrire les formats Csv compris par d'autres applications ou définir leurs propres formats Csv à usage spécial. (Python, s.d.)
- **Regular expression** (re) **:** Une expression régulière (RegEx), est une séquence de caractères qui forme un modèle de recherche. RegEx peut être utilisé pour vérifier si

une chaîne contient le modèle de recherche spécifié. Python a un package intégrer, appelé re, qui peut être utilisé avec des expressions régulières. (w3schools, s.d.)

- **Pandas :** Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. Pandas est un logiciel libre sous licence BSD (Berkeley Software Distribution License). (Wikipedia, Pandas, 2019)
- **Tweepy :** Tweepy est une bibliothèque Python pour accéder à l'API Twitter. Il est idéal pour une automatisation simple et une création de robots twitter. Tweepy a de nombreuses fonctionnalités (Collecter des tweets, Créer et supprimer les utilisateurs Tweets …). (Tweepy, s.d.)
- **NLTK :** NLTK est une plate-forme de premier plan pour la construction de programmes Python pour travailler avec les données du langage humain. Il fournit des interfaces faciles à utiliser à plus de 50 ressources corpora et lexicales telles que WordNet, ainsi qu'une suite de bibliothèques de traitement de texte pour la classification, la tokenisation, l'analyse et le raisonnement sémantique. (NLTK, 2019)
- **Demoji :** Demoji est une bibliothèque qui détecte et supprime avec précision les émojis dans le texte.
- **Langdetect :** Langdetect est une bibliothèque qui détecte le langage d'un texte précis, elle supporte plus que 55 langues.
- **Http.client :** Ce module définit les classes qui implémentent du côté client les protocoles HTTP et HTTPS (Client, 2018).
- **Spacy :** SpaCy est une bibliothèque open-source gratuite pour le traitement du langage naturel en python. Il dispose de NER, POS marquage, analyse de dépendance, vecteurs de mots et plus encore. (Spacy, 2020)
- **Farasa :** qui signifie "perspicacité" en arabe, est une boîte à outils de traitement de texte rapide et précise pour le texte arabe. Farasa peut faire la segmentation, la lemmatisation, le marquage POS, la diacritisation arabe, l'analyse de dépendance, l'analyse de circonscription, la reconnaissance entité- nommée, et la vérification orthographique.
- **Scikit-learn :** Scikit-learn est une bibliothèque libre Python destinée à l'apprentissage automatique. Elle est développée par de nombreux contributeurs. Elle comprend notamment des fonctions pour estimer des forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support.

Elle est conçue pour s'harmoniser avec d'autres bibliothèques libres Python, notamment NumPy et SciPy. (Wikipedia, Scikit-learn, 2019)

- **Matplotlib :** Matplotlib est une bibliothèque du langage de programmation Python 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. Matplotlib est distribuée librement et gratuitement sous une licence de style BSD. Sa version stable actuelle (la 2.0.1 en 2017) est compatible avec la version 3 de Python. (Wikipedia, Matplotlib , 2018)
- **Seaborn :** Seaborn est une bibliothèque de visualisation de données Python basée sur matplotlib. Il fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.
- **TextBlob correct :** bibliothèque assure la correction automatique des mots ex : miason → maison

3.3. Les caractéristiques des données sur les tweets

Nous avons adopté le même classement standard pour classer les tweets selon les caractéristiques suivantes :

- **La date de création :** la date de création du tweet. Ex :
```
"created_at": "Sun Mar 01 23:42:41 +0000 2020"
```
- **Le nom d'utilisateur** : L'utilisateur qui a posté ce Tweet. Ex :
```
"name": "Nouni Brahim"
```
- **Le pseudo de l'utilisateur** : identifiant de l'utilisateur sur twitter. Ex :
```
"screen_name": "AvocatNouni"
```
- **Le texte** : le commentaire posté par l'utilisateur. Ex :
```
"text": "Le MCA a perdu 2 pts contre le MCO quelle belle journée"
```
- **Le nombre de j'aimes par tweets** : Indique combien de fois ce Tweet a été aimé par les utilisateurs de Twitter. Ex :
```
"favorite_count": 3
```
- **Le nombre de retweets par tweets** : Nombre de fois ce Tweet a été partagé. Ex :
```
 "retweet_count": 8
```

3.4. La collecte des tweets

Pour la collecte des tweets, nous avons d'abord sélectionné un ensemble de cinq catégories discutées dans les médias pendant la période allant du 21 mars 2020 jusqu'à 15 avril 2020. Notre objectif derrière le choix de ces catégories est dicté par les évènements les plus discutés par un large public d'utilisateurs en Algérie sur l'espace twitter, ce qui nous a facilité la

reconnaissance de l'ironie et nous simplifie en même temps la tâche d'annotation de ces tweets par la suite.

Nous avons réparti les thèmes sur les cinq catégories choisies (services, politique, santé, tourisme et télévision).

Pour chaque thème, nous avons sélectionné un ensemble de mots-clés avec et sans hashtag : politique (#Hirak, tebboune, #algerie_libre_et_democratique), santé (#covid19, #corona, #كورونا), tourisme (#DiscoverAlgeria, #Travel), tv (#Netflix, #LacaseDePapel4), Services (Algérie Télécom, mobilis, ooredoo, djezzy).

Nous avons collecté 8178 tweets répartis sur les cinq catégories ciblées (voir tableau1)

Dans cette phase, nous avons supprimé (les doublons, les URLs et le mentions (ex : @walidchikh)).

Ensuite nous avons calculé le nombre de tweets dans chaque catégorie, le nombre de tweets dans tout le corpus puis le nombre moyen de mots par tweets.

Nous envisageons par la suite l'annotation des tweets comme ironique ou non ironique, d'une manière manuelle. Nous rappelons que pour la collecte du corpus, nous avons utilisé l'API de Twitter.

3.5. La catégorisation du corpus réalisé

Nous avons organisé et catégorisé notre corpus en cinq catégories distinctes qui représentent les cinq thèmes ciblés et qui sont affichés dans le tableau 1.

Thémes	Nombre de tweets	Mot/Tweet	Mot/Théme
Tv	207	24	5166
Politique	1744	28	49397
Sante	1903	22	42362
Service	4201	16	70120
Tourisme	124	29	3732
Résultat	8179	20	170777

Tableau 1 - 8179 tweets collectés et répartis sur les cinq catégories

3.6. Nettoyage des données

3.6.1. Détection automatique du langage des tweets

Avant de commencer notre procédure de nettoyage des données nous avons collecté 8179 tweets pendant la période du 21-03-2020 au 15-04-2020, ils sont triés selon leur langue en utilisant la bibliothèque « detect », nous avons constitué cinq catégories : arabe, français, anglais, mixte et des tweets qui réagit avec des emojis seulement.

Nous avons obtenu les résultats suivants : **3021** tweets arabes, **2709** tweets français, **809** tweets anglais, **719** tweets mixte (nous parlons des tweets avec dialecte algérien) et **921** tweets qui contiennent uniquement des emojis.

Ces résultats sont divisés sur les cinq thèmes ciblés dans notre travail :

	Tweets arab	Tweets franç	Tweets anglai	Tweets mixte	Tweets emo	Total	Taux
Politique	167	1144	194	101	138	1744	21%
Sante	837	686	151	65	164	1903	23%
Services	1942	662	449	536	612	4201	51%
Tv	42	135	15	11	4	207	3%
Tourisme	33	82	0	6	3	124	2%
Résultat	**3021**	**2709**	**809**	**719**	**921**	8179	

Tableau 2 - Distribution des thèmes selon leur langue

Nous remarquons du tableau 2 que pendant la période de collecte des tweets, peux d'internautes ont abordé le domaine touristique 2% et la télévision 3% en comparant avec les autres domaines sachant que ces derniers temps les évènements majeurs en Algérie concernent principalement le volet politique (hirak), le volet sanitaire (Covid 19) et la qualité de service (forte demande sur la connexion internet pendant la période de confinement et faible prestation de l'opérateur AlgérieTelecom)

3.6.2. Tokenisation et élimination des mots vides

3.6.2.1. Tokenisation

Nous avons appliqué dans cette étape la méthode de tokenisation pour diviser le tweet en sac-de-mots après nous avons éliminé les mots vides.

La tokenisation consiste à découper un texte en morceaux tels que des mots, des chiffres, des phrases, des symboles et d'autres éléments appelés tokens.

Dans le processus de tokenisation, certains caractères comme les signes de ponctuation sont ignorés. Les tokens deviennent l'entrée d'un autre processus à l'image de l'analyse syntaxique et la recherche de texte.

La tokinisation joue un rôle très important dans la phase de prétraitement. Par exemple pour le tweet :

- 103 nouveaux cas confirmés et 21 nouveaux décès en Algérie #COVID19

Le résultat de la tokinisation est :

['103', 'nouveaux', 'cas', 'confirmés', 'et', '21', 'nouveaux', 'décès', 'en', 'Algérie', '#COVID19'].

3.6.2.2. Élimination des mots vides

Les mots vides ou « stop-words » sont des mots qui n'apportent aucun sens lors de l'analyse lexicale d'un texte (les tweets dans notre cas). Ce sont donc des mots que nous les éliminons généralement lors de l'analyse de texte. Ces mots sont :

- Les conjonctions de coordination (et, ou, for, yet, و)
- Les déterminants (le, la, this, that, ال)
- Les prépositions (in, pour, avec, عَلَىٰ, مِنْ)

Nous avons utilisé notre propre dictionnaire qui contient tous les mots vides des trois langues français, arabe, anglais pour les éliminer.

3.6.3. Lemmatisation

La lemmatisation désigne un traitement lexical apporté à un texte en vue de son analyse. Ce traitement consiste à appliquer aux occurrences des lexèmes sujets à flexion (en français, verbes, substantifs, adjectifs) un codage renvoyant à leur entrée lexicale commune ("forme canonique" enregistrée dans les dictionnaires de la langue, le plus couramment), que l'on désigne sous le terme de lemme. (Wikipedia, Lemmatisation, 2020)

Dans un premier temps nous avons éliminé tous les emojis, après nous avons trié les tweets selon la langue pour appliquer le procédé de lemmatisation afin d'obtenir des résultats plus précis.

3.6.3.1. Lemmatisation des textes anglais

Nous avons utilisé dans l'analyse des textes anglais la bibliothèque « WordNetLemmatizer » pour la lemmatisation et le processus « pos_tag » du package NLTK pour la classification des mots selon leur emplacement dans la phrase et de leur étiquetage. (Figure -1)

Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

Figure 1-Table d'étiquète

3.6.3.2. Lemmatisation des tweets français

Nous avons utilisé dans l'analyse des textes français la bibliothèque « fr_core_news_md » pour la lemmatisation du package Spacy. Nous Notons que la lemmatisation et la classification des mots selon leur emplacement dans la phrase est réalisé par la bibliothèque elle-même.

3.6.3.3. Lemmatisation des tweets arabes

Nous avons utilisé dans l'analyse des textes arabes l'API « Farasa ». Cette dernière peut réaliser les tâches suivantes, la segmentation, la lemmatisation, le marquage POS, la diacritisation arabe, l'analyse de dépendance, l'analyse de circonscription, la reconnaissance d'entités- nommées, et la vérification orthographique. (Farasa, 2019)

3.6.3.4. Lemmatisation des tweets mixtes

Nous avons utilisé la lemmatisation des trois langues citées précédemment à savoir, arabe, français et anglais dans le but d'obtenir une meilleure précision.

3.6.4. L'analyse des mots

Dans cette partie de notre travail, nous sommes intéressés au calcul de la fréquence des mots cités dans chaque catégorie (politique, sante, etc …), ensuite dans tout le corpus.

Nous avons utilisé les résultats obtenus par l'opération de la lemmatisation pour calculer le nombre des mots les plus répétés. Nous avons trié les mots par leur fréquence dans un ordre décroissant. Nous avons utilisé par la suite les deux bibliothèques pandas et csv pour le stockage des résultats dans un fichier Excel. Le tableau suivant donne une partie des mots utilisés avec leurs fréquences selon leur domaine.

Politique	Fréquence	Sante	Fréquence	Service	Fréquence	Tv	Fréquence	Tourisme	Fréquence
حراك	21	ربي	163	جزائر	426	جزائري	33	سياحة	32
algerian	45	coronavirus	83	internet	76	casa	36	tourisme	76
algérien	158	cas	113	algérie	105	algérie	77	algérie	52
pouvoir	129	كورونا	135	télécome	62	papel	36	عائلي	15

Tableau 3 - Tableau des fréquences de mots

D'après l'analyse des fréquences, nous pouvons dire que :

- Les tweets de type Politique abordent beaucoup plus le gouvernement algérien comme sujet.
- Les tweets de type Sante abordent beaucoup plus la pandémie du coronavirus (Covid 19) comme sujet.
- Les tweets de type Service abordent beaucoup plus la société d'Algérie télécom comme sujet.
- Les tweets de type Tv abordent beaucoup plus la série de la casa de papel comme sujet.
- Les tweets de type Tourisme abordent beaucoup plus le tourisme en Algérie comme sujet.

Les résultats obtenus montrent que les mots (الله, حالة, كورونا, algérien, algerian, اتصال,جزائر) sont plus utilisés par la population algérienne dans cette période (figure 2).

Figure 2 - Nuages de mots les plus utilisés

3.6.5. L’analyse des hashtags

Dans cette partie nous avons calculé la fréquence des hashtags dans notre corpus (voir tableau 4).

Politique	Fréquence	Sante	Fréquence	Service	Fréquence	Tv	Fréquence	Tourisme	Fréquence
#Hirak	1287	#COVID19dz	947	#اتصالات_الجزائر_وكيلكم_ربي	1640	#LacaseDePapel4	29	#Tourisme	15
#Algerie	440	#ريح_في_داركم	564	#vive_algérie_télécom	678	#Algerie	24	#Algérie	14
#coronavirus	110	#confinement	99	#improve_dz_network	247	#netflix	9	#photography	10
#ريح_في_داركم	63	#كورونا	237	#لا_للإحتكار_إفتحو_المنافسة	66	#Hulu	4	#DiscoverAlgeria	4

Tableau 4 - Tableau des fréquences de hashtags

D’après l’analyse des fréquences des hashtags nous pouvons dire que :

- Les tweets de type Politique abordent le hashtag #Hirak et #Algérie et #coronavirus fréquemment ce qui montre qu’il y a une interaction entre le domaine politique et sante.
- Les tweets de type Sante abordent le hashtag #COVID19dz et #ريح_في_داركم fréquemment
- Les tweets de type Service abordent le hashtag #vive_algérie_télécom et #اتصالات_الجزائر_وكيلكم_ربي fréquemment ce qui indique une ironie
- Les tweets de type Tv abordent le hashtag #LacaseDePapel4 fréquemment

- Les tweets de type Tourisme abordent le hashtag #Tourisme et #Algérie fréquemment

Les résultats obtenus montrent que les hashtags (#Hirak, #COVID19dz, #اتصالات_الجزائر_وكيلكم_ربي, #ريح_في_داركم) sont les plus utilisés par la population algérienne utilisatrice du twitter durant cette période.

3.6.6. Traitement des fautes d'orthographe et abréviations

Pour le traitement des fautes d'orthographe nous avons utilisé la bibliothèque TextBlob correct qui assure la correction automatique des mots (figure 3), concernant le problème d'identification d'abréviation nous avons élaboré un dictionnaire qui contient une liste des mots associée associé avec leurs abréviations (figure 4).

```
df = pd.read_csv('Apprentissage/Corpusfinal.csv',sep=';',encoding='utf-8')
for Texte, Ironie in df.itertuples(index=False):
    print('\nAvant')
    print('--------------------')
    print(Texte)
    text = TextBlob(Texte)
    txt = text.correct()
    print('\nApres')
    print('--------------------')
    print(txt)
    csvWriter.writerow([txt,Ironie])
```

```
Avant
--------------------
1)Le match qiu a contaminé 40.000 sepctateurs #AtlantaValencia #COVID_19

Apres
--------------------
1)Le match qui a contaminé 40.000 spectateurs #AtlantaValencia #COVID_19

Avant
--------------------
2)73 nwe confirmed cases of coronavirus (Covid-19) nad 4 new daeths have ben recorded durin th last 24 hours in Algeria, bringi
ng the number of confirmed cases to 584 and that of deaths to 35, announced on Monday in Algiers. #Algeria #covid19dz #coronavi
rusalgerie

Apres
--------------------
2)73 new confirmed cases of coronavirus (Covid-19) and 4 new deaths have been recorded during the last 24 hours in Algeria, bri
nging the number of confirmed cases to 584 and that of deaths to 35, announced on Monday in Algiers. #Algeria #covid19dz #coron
avirusalgerie
```

Figure 3 – Exemple de code python pour la correction automatique des faute d'orthographe

```
mots = ['Beaucoup','Pour quoi','Coucou','Salut']
abbreviations = ['slt','cc','bcp','pq']
```

```
B [('bcp', 100), ('slt', 0), ('cc', 0), ('pq', 0)]
Pq [('pq', 100), ('bcp', 67), ('slt', 0), ('cc', 0)]
C [('cc', 100), ('bcp', 100), ('slt', 0), ('pq', 0)]
S [('slt', 100), ('cc', 0), ('bcp', 0), ('pq', 0)]
Beaucoup - ('bcp', 100)
Pour quoi - ('pq', 100)
Coucou - ('cc', 100)
Salut - ('slt', 100)
```

Figure 4 - Exemple de code python pour la détection d'abréviation

3.7. Phase d'annotation du corpus

3.7.1. Les types de textes dans le corpus

Nous présentons à titre d'exemple quatre types de tweets algériens de notre corpus réalisé, un tweet en langue arabe classique écrit en caractère arabe (exemple 01), un tweet en langue arabe classique écrit en caractère latin (exemple 02), un tweet en langue arabe dialectal écrit en caractère arabe (exemple 03), un tweet en langue arabe dialectal écrit en caractère latin (exemple 04).

(01) ماذا أفعل بالرصيد المجاني والنت ضعيييييف 😂

(02) mobilis fal tadhabi ila ljahim 😂😂😂😂😂😂

(03) السميد اللي ملهوفين عليه اليوم، غدوة يفتلوه في جنازتكم الا قعدتو تخرجو هكذا☺

(04) Uuugh ghalqou lejwama3 bch tji direlna nta khotba ta3 joumou3a ? #covid19dz #hirak #tebboun #dz

3.7.2. Annotation manuelle des tweets

Dans le cadre de notre projet de fin d'études, nous nous intéressons à la classification des tweets selon deux classes (voir figure 5) à savoir :

- **Ironique :** Un tweet est ironique s'il exprime une ironie verbale, ironie situationnelle, sarcasme, satire ou humour (e.g. تالمون تقيلة تقول الجزائرين يكونيكتوا قاع بمودام واحد 😂 #اتصالات_الجزاير_وكيلكم_ربي)
- **Non ironique :** Un tweet est dit non ironique s'il ne correspond à aucune forme d'ironie (e.g. #Coronavirus : le confinement total commence-t-il à porter ses fruits à #Blida ? #COVID19dz #RESTEZCHEZVOUS)

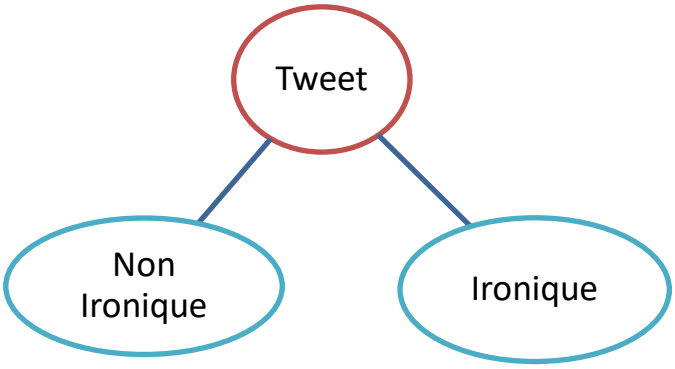

Figure 5 - Schéma d'annotation

Nous avons affecté pour les tweets ironiques la valeur '1' et pour les tweets non ironiques la valeur '0'.

Nous avons obtenu les résultats suivants (voir tableau 5) :

	Ironique	Non ironique
Arabe	515	2506
Français	382	2327
Anglais	60	749
Multi langues	160	1480
Total	1117	7062

Tableau 5 - Résultat de l'annotation des tweets

Nous avons recensé **3231** Tweets écrits en arabe partagé en deux groupes (arabe classique, arabe dialectal) et dans les deux groupes nous rencontrons des tweets écrits soit en arabe soit en latin ce qui représente un taux de **39.5%** du corpus général. Dont **2468** sont écrits avec l'arabe classique avec un taux de **30.17%** du corpus général, et **743** sont écrits en arabe dialectal (Algérien) avec un taux de **9.09%** du corpus général, le reste des tweets sont écrits avec leurs langues respectives ce qui représente un taux de **61.5%** du corpus général. (Tableau 6)

	Ironique	Non ironique	Taux/corpus
Arabe classique écrit en caractère arabe	266	2217	30,36%
Arabe classique écrit en caractère latin	2	3	0,06%
Arabe dialectal écrit en caractère arabe	187	299	5,94%
Arabe dialectal écrit en caractère latin	56	201	3,14%
Total	511	2720	39,50%
Total Arabe dialectal (algérien)	243	500	9,09%
Total Arabe classique	268	2220	30,42%

Tableau 6 - Résultat de l'annotation des tweets arabe

3.7.3. La nature des tweets algériens

Nous avons recensé **743** Tweets écrit en dialecte algérien ce qui représente un taux de **8.33%** du corpus général dont **257** sont écrits en caractère latin avec un taux de **2.88%** du corpus global et **486** sont écrits en caractères arabes avec un taux de **5.45%** du corpus global (Tableau 6)

3.8. Conclusion

Dans ce chapitre, nous avons décrit le contenu de notre corpus réalisé en montrant les cinq thèmes utilisés dans ce travail (politique, santé, service, tourisme, télevision). Nous avons aussi montré les quatre types de la langue arabe utilisée (arabe classique ou dialectal qui utilisent soit des caractères arabe ou latin). Nous avons présenté les opérations principales du prétraitement du corpus à savoir la tokenisation, l'élimination des mots vides et la lemmatisation selon la langue utilisée. Le chapitre suivant présente l'essentiel de la phase de classification des tweets selon la classe ironique ou non-ironique.

Chapitre 4

4. Classification automatique des tweets

4.1. Introduction

Nous ciblons dans ce chapitre la classification de notre corpus en tweets ironiques et non ironiques dans toutes nos expérimentations réalisées lors de ce projet de fin d'étude. Pour cela nous avons utilisé les différents modes de classification existants dans la littérature. Nous rappelons que chaque mode de classification est lié avec un type d'apprentissage i.e. supervisé, non supervisé et semi-supervisé.

4.2. La classification des tweets

Nous avons utilisés trois algorithmes très sollicités dans la littérature en mode supervisé i.e Support Vector Machine, Random Forest et Naive Bayes. Aussi Nous avons fait appel à l'algorithme K-means en mode non supervisé et l'algorithme Support Vector Machine une autre fois en mode semi-supervisé. Nous avons réalisé quatre expérimentations différentes pour le développement des trois modes de classification. Notons que pour chaque expérimentation, nous avons scindé le corpus correspondant en deux parties (2/3 pour la phase d'apprentissage et 1/3 pour la phase de test). Nous avons utilisé quatre métriques pour évaluer les résultats de test dans chaque expérimentation. Nous rappelons que les quatre expérimentations sont liées avec l'organisation de la base d'apprentissage.

4.2.1. Métriques utilisées :

La première métrique concerne le rappel ou Recall, ce dernier permet de donner la proportion des tweets ciblés identifiés correctement.

$$Recall = TP/TP+FN$$

La deuxième métrique concerne la précision, elle permet de donner la proportion des tweets ciblés qui sont effectivement correctes.

$$Précison = TP/TP+FP$$

La troisième métrique concerne le F1-score, cette métrique combine la précision et le rappel est leur moyenne harmonique.

F1-score = 2*[(précision*rappel) / (précision+rappel)]

La quatrième métrique concerne le taux (AUC- Area Under The Curve) de classification, cette métrique mesure le taux de classification par rapports au nombre total des tweets

Taux : (TP+TN) / (TP+TN+FN+FP)

4.2.2. Courbes utilisées :

Nous avons aussi utilisé deux types de courbes pour enrichir la partie discussion de nos résultats obtenus comme la courbe AUC (Area Under The Curve) ROC (Receiver Operating Characteristic) qui trace le taux de vrais positifs en fonction du taux de faux positifs et la courbe Rappel-Précision.

4.2.2.1. AUC-ROC :

Pour l'apprentissage artificiel, la mesure des performances est une tâche essentielle qui indique le succès ou l'échec de cette tâche. Nous utilisons aussi la courbe AUC-ROC pour évaluer et visualiser les performances d'un classifieur multi-classes.

- **AUC d'un classifieur idéal :** Un classifieur idéal ne fait aucune erreur de prédiction. Cela signifie que le classifieur peut parfaitement séparer les classes de façon que le modèle atteint un taux positif (ironique) réel de 100% avant de produire des faux positifs (non ironique). Ainsi, l'AUC du classifieur égal 1. (Figure 6)

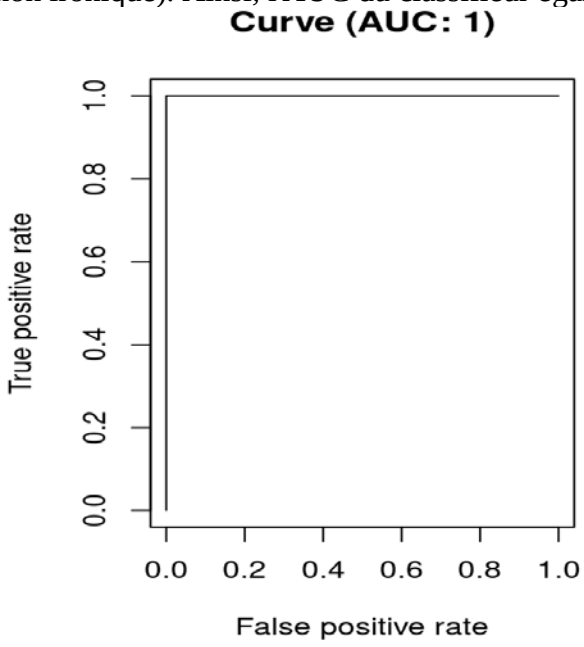

Figure 6 - Courbe AUC-ROC d'un classifieur idéal

- **AUC d'un bon classifieur :** Un classifieur qui sépare bien les deux classes, mais pas parfaitement. (Figure 7)

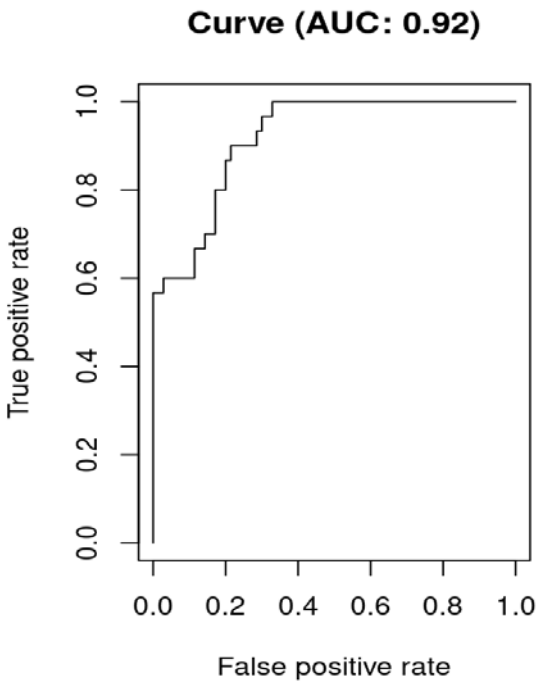

Figure 7 - Courbe AUC-ROC d'un bon classifieur

- **AUC d'un mauvais classifieur :** Un mauvais classifieur affichera des scores dont les valeurs ne sont que légèrement associées au résultat. Le classifieur n'atteindra un TPR (True Positive Rate) élevé qu'après un FPR (False Positive Rate) élevé. (Figure 8)

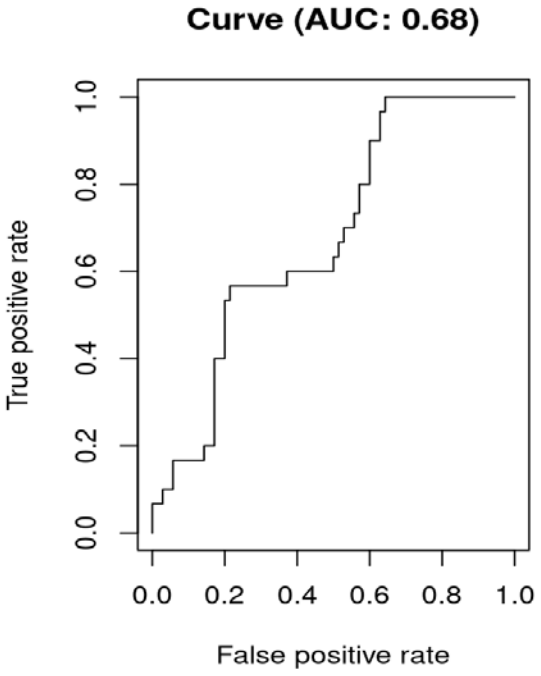

Figure 8 - Courbe AUC-ROC d'un mauvais classifieur

- **AUC d'un classifieur aléatoire :** Un classifieur aléatoire aura un AUC proche de 0,5. C'est facile à comprendre : pour chaque prédiction correcte, la prochaine sera incorrecte. (Figure 9)

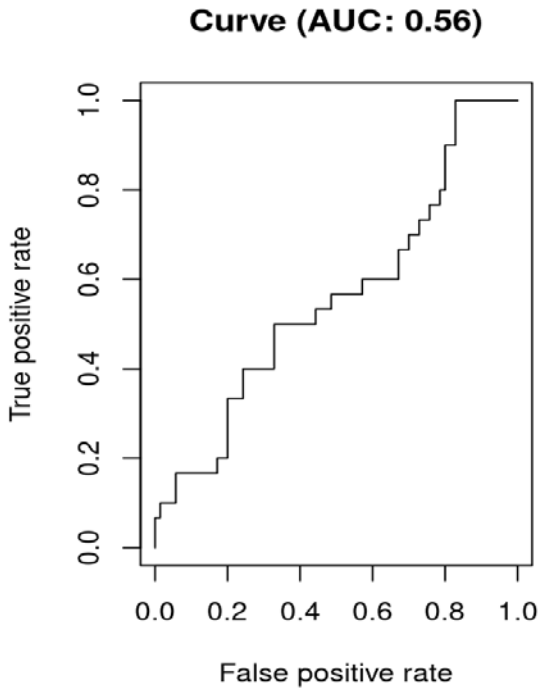

Figure 9 - Courbe AUC-ROC d'un classifieur aléatoire

4.2.3. Précision-Rappel :

La courbe Précision-Rappel trace la valeur prédictive positive (PPV, y-axe) par rapport au taux positif réel (TPR, x-axe).

- **AUC d'un classifieur idéal :** Un classificateur idéal ne fait aucune erreur de prédiction. Ainsi, il obtiendra un AUC-PR de 1. (figure 10)

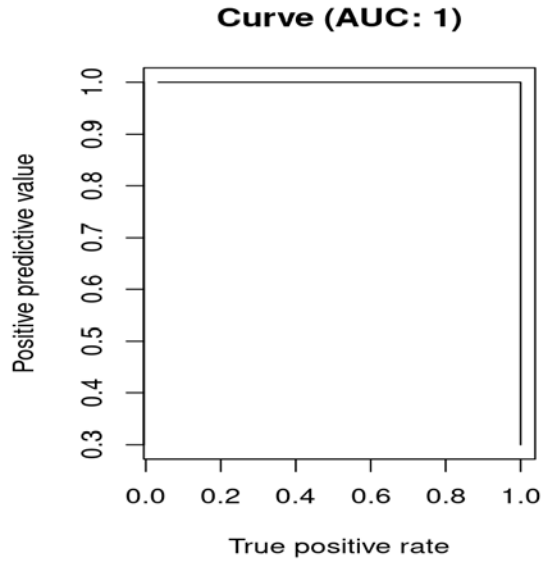

Figure 10 - Courbe Précision-Rappel d'un classifieur idéal

- **AUC d'un mauvais classifieur :** Un mauvais classificateur affichera des scores dont les valeurs ne sont que légèrement associées au résultat. Un tel classificateur n'atteindra un rappel élevé qu'avec une faible précision. (figure 11)

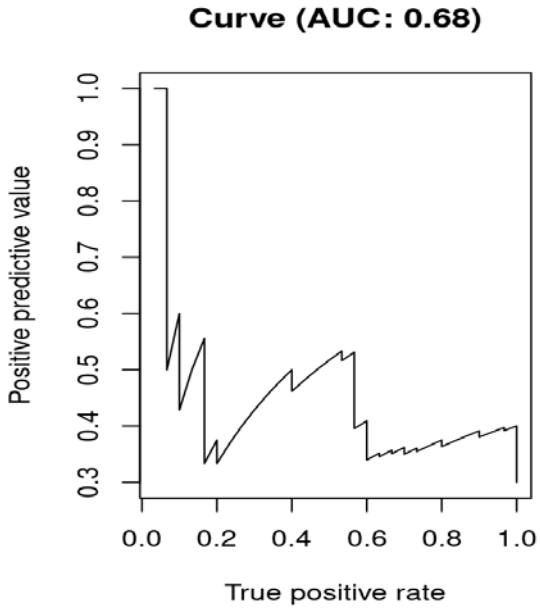

Figure 11 - Courbe Précision-Rappel d'un mauvais classifieur

- **AUC d'un classifieur aléatoire :** Un classifieur Un classificateur aléatoire a un AUC-PR proche de 0,5. C'est facile à comprendre : pour chaque prédiction correcte, la prochaine prédiction sera incorrecte. (figure 12)

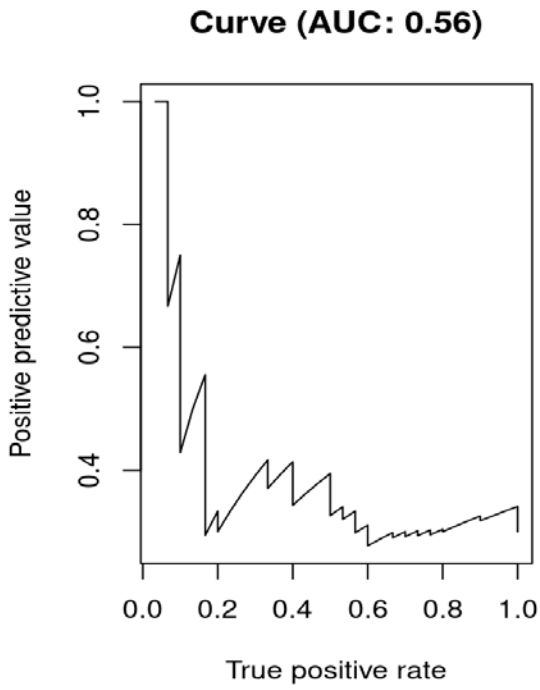

Figure 12 - Courbe Précision-Rappel d'un classifieur aléatoire

Nous avons utilisé les trois modes d'apprentissage (supervisé, non-supervisé et semi-supervisé) pour entrainer les cinq classifieurs en faisant appel aux quatre algorithmes cités auparavant (SVM, Naivebayes, random-forest et K-means), nous avons aussi choisi quatre expérimentations distinctes, ces dernières sont liées avec la distribution des exemples ironiques et non ironiques dans la base d'apprentissage. Nous avons par contre discuté dans ce chapitre les résultats qui présentent des difficultés lors de l'apprentissage avec le maximum d'exemples à savoir la première expérimentation (corpus global) et la deuxième expérimentation (corpus multi-langue) les autres résultats sont reportés aux annexes. Après cette phase d'apprentissage nous avons évalué ces derniers sur la base de test.

4.2.4. Apprentissage supervisé

4.2.4.1. Première expérimentation :

Nous avons utilisé le corpus global de 8178 tweets. Notons que les tweets non-ironiques présents dans ce corpus sont majoritaires avec un taux de 86.35 % par rapport aux tweets ironiques qui représentent seulement 13.65 %, il s'agit dans ce cas d'une base d'exemples non équilibrée.

La figure 13 présente un histogramme que montre la distribution des tweets du corpus global. Ironiques et non-ironiques :

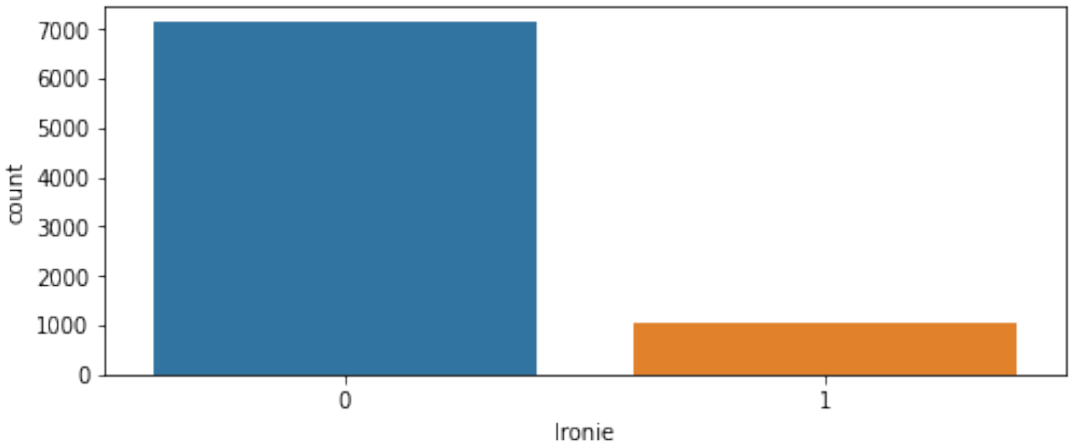

Figure 13 - Distribution des tweets du corpus globale

Nous avons obtenu les taux de classification suivants : 89.487% pour le classifieur NB (Naïve Bayes), 90.273% pour le classifieur SVM et 90% pour le classifieur RF (Random Forest).
Le tableau ci-dessous récapitule les résultats obtenus en termes des quatre métriques pour chaque classifieur. (Tableau 7)

	Précision	Rappel	F1-score	Accuracy
Naive Bayes	0.805	0.64	0.713	0.90
SVM	0.735	0.68	0.706	0.89
RFC	0.835	0.66	0.737	0.91

Tableau 7 - Première expérimentation

Nous avons dressé les deux courbes ROC et précision-rappel des trois classifieurs.

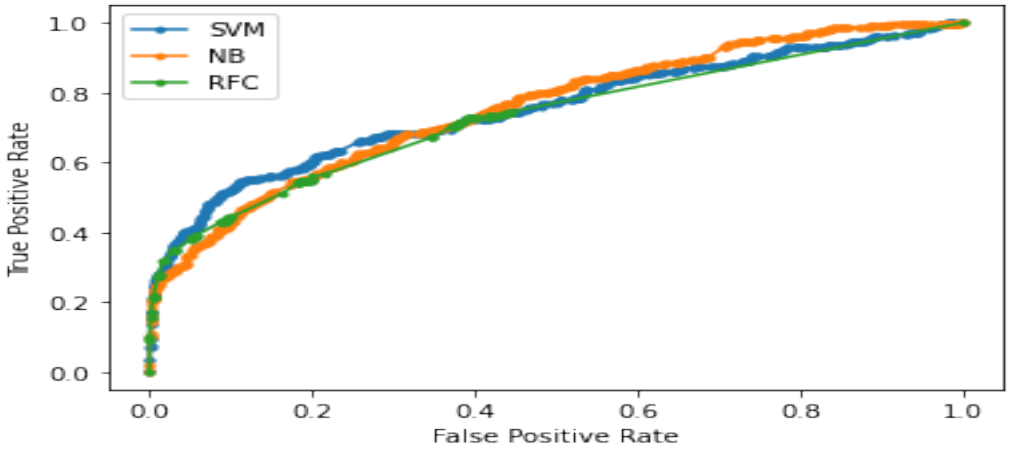

Figure 14 - La courbe ROC du corpus global

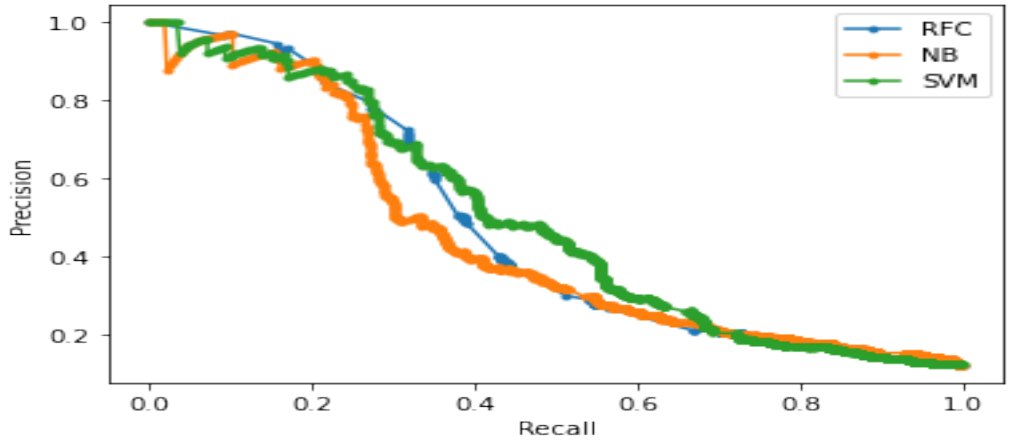

Figure 15 - La courbe Précision/Rappel du corpus global

Nous remarquons à partir des deux figures (figure 15 et figure 16) que les trois classifieurs ont obtenu des performances similaires avec une légère supériorité du premier classifieur à base de l'algorithme SVM (ROC AUC=0.755) en comparant au deuxième classifieur NB (ROC AUC=0.756), et au troisième classifieur RFC (ROC AUC=0.731).

Nous présentons à titre d'exemple quatre types de tweets après prétraitement identifiés par le meilleur classifieur-SVM :

TP : un tweet ironique et détecté ironique par le classifieur

TN : un tweet non-ironique et détecté non-ironique par le classifieur

FP : un tweet non-ironique et détecté ironique par le classifieur

FN : un tweet ironique et détecté non-ironique par le classifieur

(1) **TN**: good people silence option we continue raise voice matter repression dictatorship hirak

(2) **FN** : connexion tortue étre corona.

(3) **TP** : كورونا انترنت نتاع سلحفاة ريقلونا شبكة رحم والد

(4) **FP**: karim tabbou militanteak eta kazetariak besteak beste sufritzen dute aljeriako erregimenaren ankerkeria hirak mugimendua covid 19

Nous remarquons que la présence des FPs et des FNs sont dus essentiellement aux problèmes standards d'apprentissage (le choix des paramètres initiaux du classifieur et la vitesse d'apprentissage), nous mentionnons ici la difficulté d'ajuster tous des paramètres du classifieur pendant la phase d'entrainement ce qui représente en même temps un axe de recherche à part entière. Et un problème de caractérisation qui est engendrée avec beaucoup de difficultés comment a été mentionné dans les exemples de tweets cités ci-dessus où l'exemple de FP concerne un tweet qui a utilisé le dialecte kabyle écrit en latin (problème de désambiguïsation). Aussi l'exemple FN d'un tweet utilisé avec le français mais il a été mal détecté, deux causes possibles pour cette erreur soit un problème de caractérisation ou un problème d'apprentissage par contre les deux exemples de TN et de TP sont représentés respectivement en anglais avec des caractères latin et en arabe dialectal avec des caractères arabe, ces deux exemples ont été reconnus facilement par notre classifieur.

4.2.4.2. Deuxième expérimentation :

Nous avons scindé le corpus en quatre sous-corpus selon la langue utilisée (arabe, français, anglais, multi-langues). Nous présentons dans cette partie uniquement l'expérimentation qui concerne le quatrième sous-corpus, sachant que ce dernier présente beaucoup de contraintes liées principalement avec la présence simultanée de plusieurs langues dans le même tweet

1er sous corpus : 3181 tweets arabe

2ème sous corpus : 2723 tweets français

3ème sous corpus : 539 tweets anglais

4ème sous corpus : 1735 tweets multi-langues

Pour ce 4ème sous corpus, nous avons entrainé les trois classifieurs, après une phase d'apprentissage et nous avons testé ces derniers sur la base de test.

Nous avons obtenu les taux de classification (accuracies) suivants : un taux de 88.798% pour le classifieur NB (Naïve Bayes) et un taux de 88.145% pour le classifieur SVM et un taux de 90.145% pour RF (Random Forest).
La figure 16 présente un histogramme que montre la distribution des tweets arabe ironique et non-ironique :

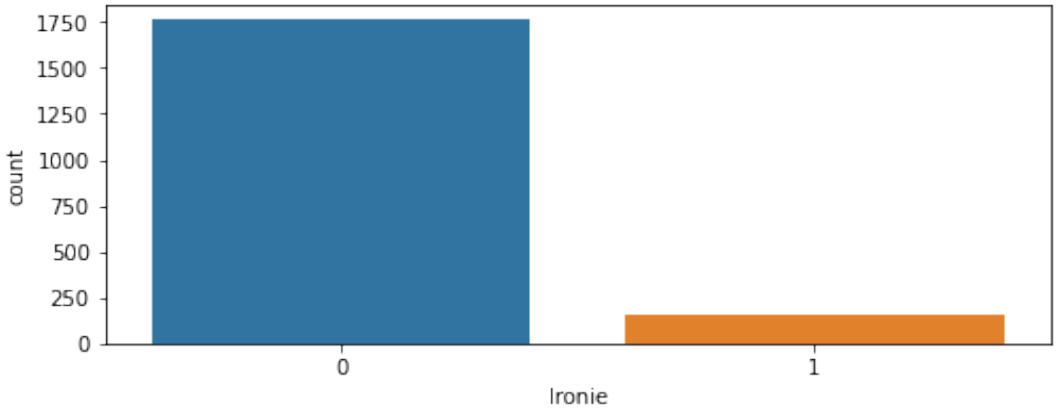

Figure 16 - Distribution des tweets du 4ème sous corpus

Les tableaux récapitulent les résultats obtenus en termes des quatre métriques pour chaque classifieur. (Tableau 8)

	Précision	Rappel	F1-score	Accuracy
Naive Bayes	0.70	0.545	0.6128	0.89
SVM	0.665	0.58	0.6195	0.88
RFC	0.76	0.545	0.6347	0.90

Tableau 8 - Deuxième expérimentation 4ème sous corpus

Nous avons présenté les deux courbes ROC et précision-rappel des trois classifieurs dans les deux figures (figure 17 et figure 18)

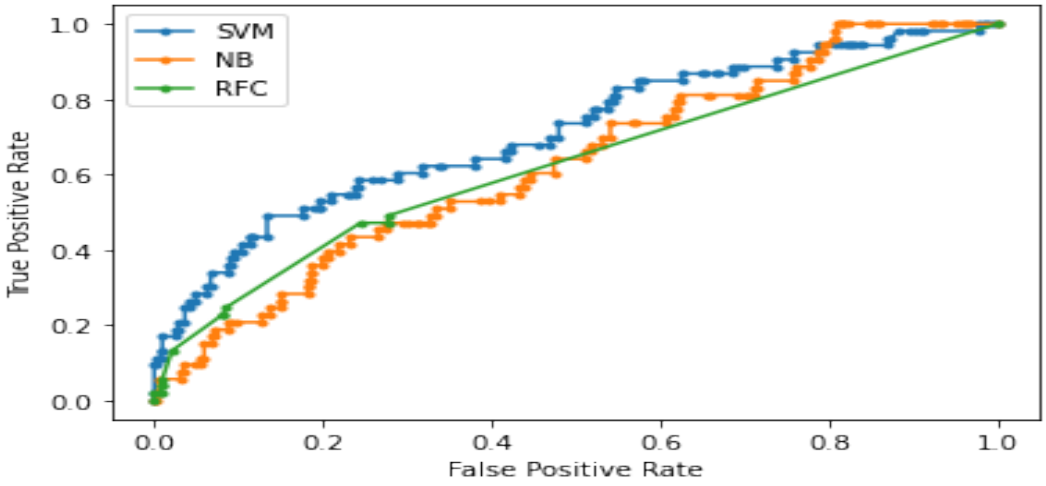

Figure 17 - La courbe ROC du 4ème sous corpus

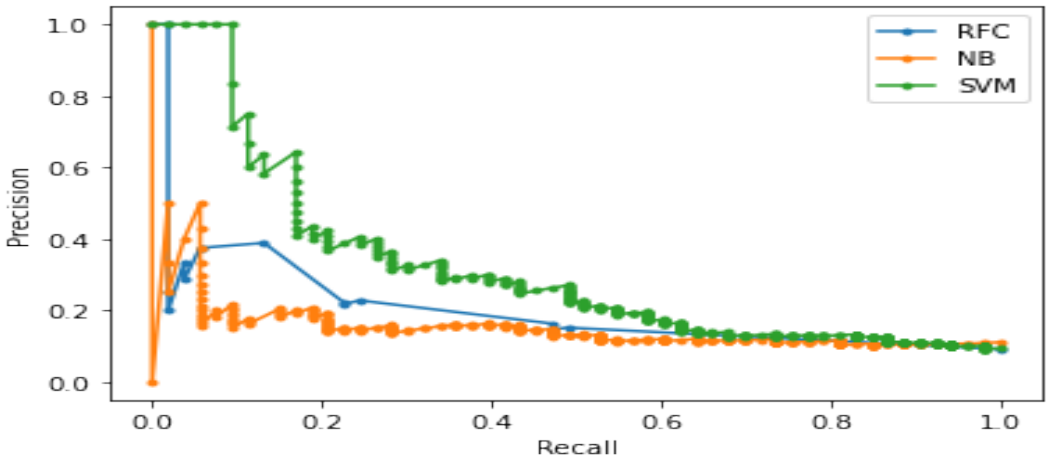

Figure 18 - La courbe Précision/Rappel du 4ème sous corpus

Nous remarquons aussi à partir des deux figures que les trois classifieurs ont obtenu des performances similaires avec une légère supériorité du premier classifieur à base de l'algorithme SVM (ROC AUC=0.713) comparant aux deux autres classifieurs à base de NB (ROC AUC=0.628), et de RFC (ROC AUC=0.624).

Nous présentons quatre exemples après prétraitement de type TP, TN, FP et FN obtenus avec le classifieur-SVM

(1) **FN** : l'algerie تربح راح 2-1 واش تقولو alg-nig sur la chaine 6

(2) **TP** : جميل netflix شحال حلقة algerie

(3) **FP** : netflix ntouma tani kima اتصال جزاير

(4) **TN** : algeria wind of change ثورة نظام أراد نخبة.

Nous remarquons toujours que la présence des FPs et des FNs sont dus essentiellement aux problèmes standards d'apprentissage déjà cités auparavant. Aussi le problème de

caractérisation se pose toujours comme a été remarqué dans les exemples de tweets cités ci-dessus où l'exemple FP concerne un tweet qui a utilisé un arabe dialectal écrit en latin et un arabe classique écrit en caractères arabe. Aussi l'exemple FN concerne un tweet qui a utilisé un arabe dialectal écrit en caractères arabe et français écrits en caractères latin. Idem pour les deux exemples du TP et TN, le premier est représenté en arabe dialectal écrit en caractères arabe et en même temps il est représenté par des termes en français écrits en caractères latin, le deuxième exemple est représenté en anglais avec des caractères latins et en même temps il est représenté en arabe classique avec des caractères arabe mais ces deux derniers exemples sont correctement reconnus par notre classifieur.

4.2.5. Apprentissage semi supervisé

Nous avons développé dans cette section un classifieur à base de l'algorithme SVM (Support Vector Machine) en mode semi supervisé. Ce mode d'apprentissage est déjà présenté dans le deuxième chapitre. Nous avons réalisé les quatre expérimentations citées auparavant.

Nous avons choisi une base d'exemple équilibrée constitué de 200 tweets représentées par des tweets ironiques et non ironiques. Nous avons entrainé notre classifeur SVM. Ensuite pour chaque expérimentation, nous avons appliqué le principe du mode semi supervisé pour développer le classifieur final à base de l'algorithme SVM

4.2.5.1. Première expérimentation :

Nous avons obtenu un taux de classification de 90.273% pour le classifieur SVM.
Le tableau 9 récapitule les résultats obtenus en termes des quatre métriques pour chaque classifieur.

	Précision	Rappel	F1-score	Accuracy
SVM	0.735	0.68	0.705	0.90273

Tableau 9 - Première expérimentation

Nous avons présenté les deux courbes ROC et précision-rappel de classifieur kNN dans les deux figures (figure 19 et figure 20).

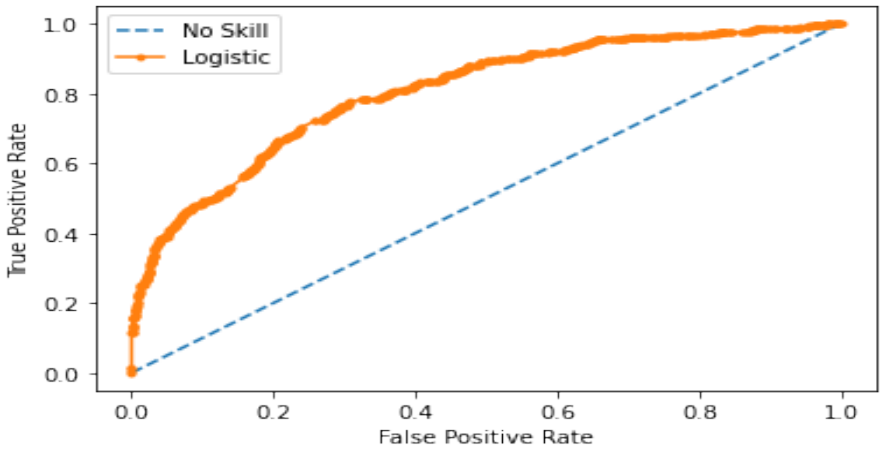

Figure 19 - La courbe ROC du corpus global

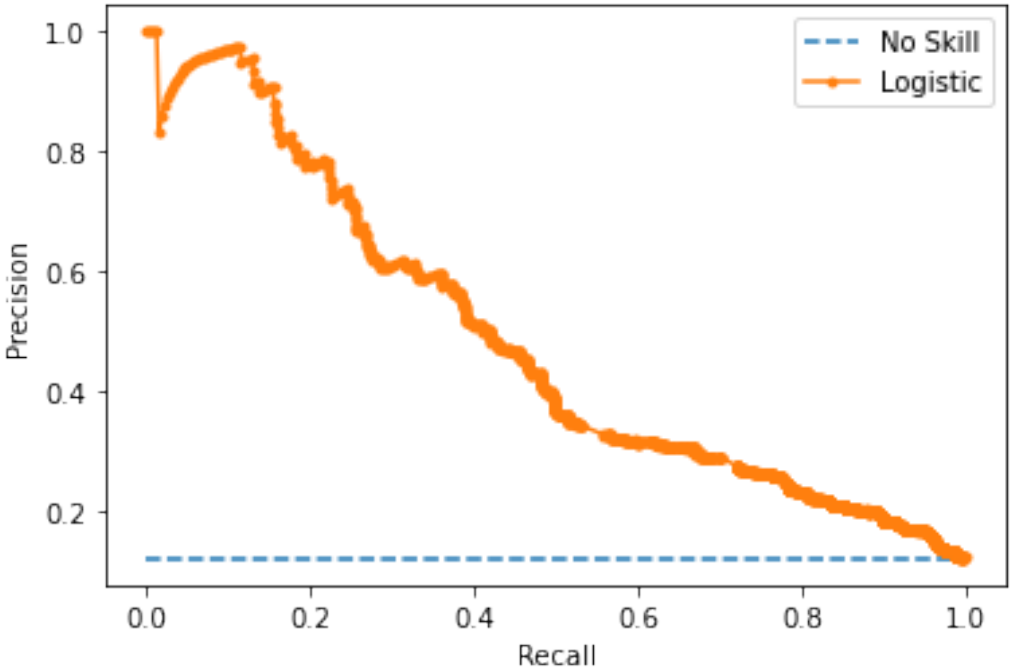

Figure 20 - La courbe Précision/Rappel du corpus global

Nous remarquons à partir des deux figures que le classifieur SVM a obtenu un résultat de (ROC AUC=0.695).

Nous présentons quatre exemples de tweets annotés par le classifieur-SVM.

(1) **FP** : راعي رسمي حجر صحي covid 19

(2) **FN** : liberté detenu hirak

(3) **TP** : خلاص فراها بوصبع زرق دوك قدر أدار حجر ليلي متنساش بوصبع كي نسحقوك فرينا

(4) **TN** : suiver instagram challenge attendre faire passer temp beauté

L'exemple FP concerne un tweet qui a utilisé l'arabe dialectal écrit en caractères arabe et le français écrite en latin. Aussi l'exemple FN concerne un tweet qui a utilisé le français écrit en caractères latin. L'exemple TP est représenté en arabe dialectal écrit en caractères arabe et l'exemple TN est représenté en français écrit en caractères latin. Ces deux derniers exemples ont été reconnus facilement par notre classifieur.

4.2.5.2. Deuxième expérimentation :

Dans le 4ème sous corpus, nous avons obtenu un taux de classification de 88.145%

Le tableau 10 récapitule les résultats obtenus par notre classifieur en termes des quatre métriques.

	Précision	Rappel	F1-score	Accuracy
SVM	0.665	0.58	0.60	0. 88145

Tableau 10 - Deuxième expérimentation 4ème sous corpus

Nous avons présenté les deux courbes ROC et précision-rappel de classifieur kNN dans les deux figures (figure 21 et figure 22).

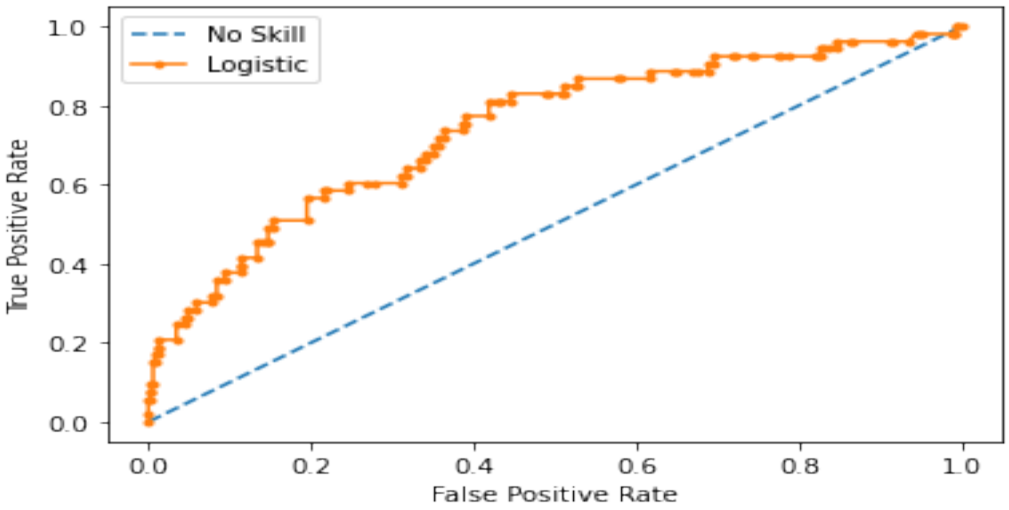

Figure 21 - La courbe ROC du 4ème sous corpus

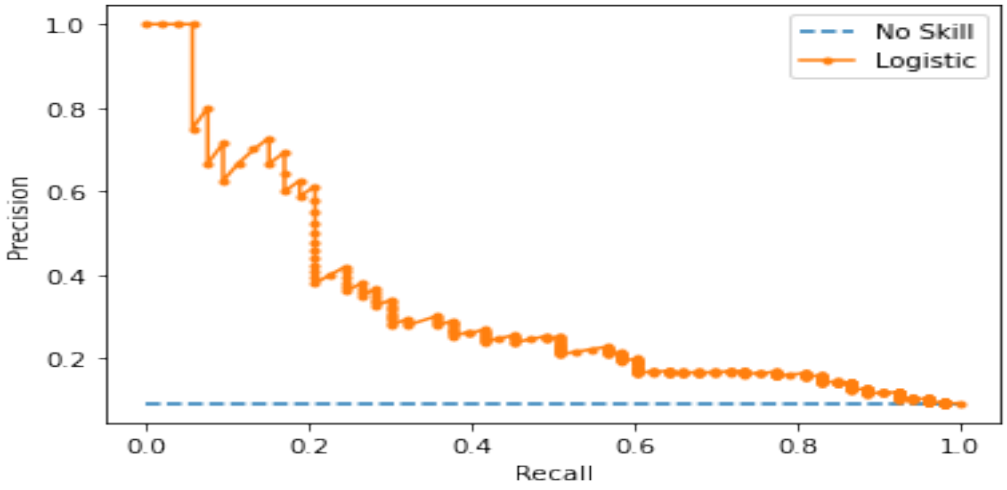

Figure 22 - La courbe Précision/Rappel du 4ème sous corpus

Nous remarquons à partir des deux figures que le classifieur SVM a obtenu le résultat (ROC AUC=0.54)

Nous présentons le tableau des TP, TN, FP et FN avec le classifieur-SVM

(1) **FN** : bütün dünyanin sevdigi bir dizide bu zamanda türk karekteri osman ve bu osman dizide isgenceci rasistlik bu casadepapel pay envoyé aide turquie casa papel osman turk algerie militant islamiste

(2) **FP** : نهائي كأس كورونا جزائر تمنراست سعيد

(3) **TN:** stay safe stayhome covid 19

(4) **TP :** مشكلة جزائر حكومة شجع سياحة دليل فيزا صعب

L'exemple FP concerne un tweet qui a utilisé l'arabe dialectal écrit en caractères arabe. Aussi l'exemple FN concerne un tweet qui a utilisé le turk écrit en caractères latin. L'exemple TP est représenté en arabe dialectal écrit en caractères arabe et le deuxième exemple TN est représenté en anglais écrit en caractères latin. Ces deux derniers exemples ont été reconnus correctement par notre classifieur.

4.2.6. Apprentissage non supervisé

Nous avons déjà mentionné en introduction de ce rapport que nous avons fait appel au classifieur à base de l'algorithme kNN (kNeighborsClassifier), ce dernier a été largement cité dans la littérature. Nous avons réalisé quatre expérimentations différentes pour le développement de ce classifieur. Notons que pour chaque expérimentation, nous avons scindé le corpus correspondant en deux parties (2/3 pour la phase d'apprentissage et 1/3 pour la

phase de test). Nous avons utilisé les mêmes métriques pour évaluer les résultats de test pour chaque expérimentation

Nous mentionnons toujours que les mêmes corpus sont utilisés dans les trois modes d'apprentissage.

4.2.6.1. Première expérimentation :

Nous avons obtenu un taux de classification (accuracies) de 64.9 % pour le classifieur K-means.
Le tableau 11 récapitule les résultats obtenus en termes des quatre métriques pour chaque classifieur.

	Précision	Rappel	F1-score	Accuracy
K-means	0.482	0.475	0.456	0.649

Tableau 11 - Première expérimentation

Nous présentons à titre d'exemple quatre types de tweets classés par le classifieur-K-means comme TP, TN, FP et FN

(1) **FN** : لاميزار نتوم vive algérie télécom
(2) **TN** : merci choisir mobilis votre satisfaction objectif mobilishelp
(3) **FP** : اتصال جزاير كيل رب fdsfss
(4) **TP** : habit nchuf film ma9dertch xd roht insta nchuf live chose vive algérie télécom

Nous remarquons que les erreurs de classification (FPs et FNs) sont dues essentiellement aux problèmes d'apprentissage et de caractérisation déjà mentionnés auparavant. L'exemple du tweet FP a utilisé un arabe classique écrit en caractères arabe et a utilisé des caractères latins écrits au hasard. Aussi l'exemple FN concerne un tweet qui a utilisé un arabe dialectal écrit en caractères arabe et le français écrit en latin. L'exemple TP a été représenté en arabe dialectal écrit en caractères latin plus le français écrit en latin et le deuxième exemple TN est représenté en français écrit avec des caractères latin. Ces deux derniers exemples ont été reconnus correctement par notre classifieur.

4.2.6.2. Deuxième expérimentation :

Dans ce 4 ème sous corpus, nous avons obtenu un taux de classification (accuracies) de 42.32% pour le classifieur K-means.

Le tableau 12 récapitule les résultats obtenus en termes des quatre métriques pour ce classifieur.

	Précision	Rappel	F1-score	Accuracy
K-means	0.41	0.43	0.41	0.423

Tableau 12 - Deuxième expérimentation 4ème sous corpus

Nous présentons les quatre exemples classés par le classifieur K-means :

(1) **FN** : Utiliser zit zitoun pour corona khawti.
(2) **TN** : Discover Algeria cette photo de lune alger
(3) **TP** : دعم حساب خون فيروس كورونا hirak
(4) **FP** : el gobierno argelia juzgó condenó año prisión tabbou sin presencia sus abogados el condenado sufrió una presión arterial cual dejó boca los miembros exteriores paralizados los argelinos creen quiere matar algerie hirak جزائر

L'exemple FP concerne un tweet qui a utilisé l'espagnole écrit en caractères latin, un mot en français écrit en caractères latin et un autre mot en arabe classique écrit en caractères arabe. Aussi l'exemple FN concerne un tweet qui a utilisé un arabe dialectal écrit en caractères latin et le français écrit en caractères latin. L'exemple TP est représenté en arabe dialectal écrit en caractères arabe plus le français écrit en caractères latin. Le dernier exemple TN est représenté en français et en anglais et écrit avec des caractères latin. Ces deux derniers exemples ont été reconnus correctement par notre classifieur.

Une remarque pertinente concernant les quatre expérimentations réalisées en mode non supervisé est que les performances du classifieur K-means sont nettement inférieures par rapport aux autres classifieurs.

4.3. Conclusion

Durant ce quatrième chapitre nous avons présenté la phase d'annotation des huit mille cent soixante-dix-neuf (8179) tweets de notre corpus où 12.53 % du total des tweets sont ironiques le reste est non-ironique. Nous avons réalisé quatre expérimentations où nous avons examiné des situations différentes concernant la base d'apprentissage des classifieurs. Nous avons remarqué que l'organisation et la taille et la proportion des tweets influent sur les résultats globaux de classification. Ce travail nous a permet d'ouvrir plusieurs pistes de recherche et des perspectives en particulier le problème de caractérisation des tweets sachant que les

tweets issus du dialecte Algérien ont été mal identifiés en comparant avec le cas des autres tweets où une phase de désambiguïsation est primordiale pour remédier à ce problème.

Conclusion générale

Les informations personnelles mises en ligne par les utilisateurs des réseaux sociaux ont une valeur ajoutée, pour les décideurs politiques et les dirigeants des entreprises en général.

L'analyse des données issues des réseaux sociaux est un thème de recherche actif et d'actualité. Il peut être utilisé dans une variété de domaines d'application comme le domaine politique, le marketing, le sport, le tourisme et la santé.

En effet, jusqu'à présent, aucun outil dit intelligent n'est encore arrivé au point de faire une classification parfaite de données dans le but de détecter une ironie sur les tweets, même pas les êtres humains à cause de la subjectivité et l'incohérence de certains commentaires.

Le sujet abordé dans le cadre de ce master a un double objectif :

1. Proposer un corpus de tweets algériens annoté manuellement.
2. Effectuer un prétraitement sur le corpus et implémenter des classifieurs afin d'identifier l'ironie sur ces tweets.

Nous rappelons que dans le cadre de notre projet de fin d'études, nous avons proposé trois modes de classification des tweets (ironiques ou non ironiques),où nous avons fait appel aux trois types d'apprentissage, supervisé, non-supervisé et semi-supervisé. Nous avons utilisé 04 algorithmes distincts :

- Support vector machine
- Naïve Bayes
- Random forest
- K-means

Nous avons constitué un corpus de 8178 tweets dont 13.65% sont ironiques et le reste non ironiques.

Les résultats de classification obtenus sont très encourageants. Le résultat optimal a été obtenu avec le clssaifieur SVM en mode supervisé avec un taux de précision de 90%. Par contre les mauvaises performances ont été obtenues avec le classifieur Naïve Bayes avec le mode supervisé. Les difficultés rencontrées lors de l'implémentation de ces classifieurs sont

liées principalement avec la phase d'apprentissage où les proportions des exemples ironiques et non ironiques dans la base d'exemples d'entrainement influent énormément sur les résultats de ces derniers.

Aussi la caractérisation des tweets constituent un autre handicap pour la classification, sachant qu'une bonne partie des commentaires sur les tweets est non structurée et des fois même incohérente.

A la fin de ce mémoire, nous dégageons quelques perspectives futures permettant d'améliorer et d'enrichir nos différentes propositions.

Une première perspective consiste à faire appel à d'autres techniques de caractérisation des tweets pour représenter ces derniers avec des caractéristiques pertinentes.

Une deuxième perspective est liée à l'annotation des tweets du corpus, où nous proposons d'élargir l'opération d'annotation à d'autres personnes pour augmenter la fiabilité de cette tâche.

Une troisième perspective à nos travaux concerne la phase d'apprentissage où nous proposons l'appel aux principes de l'apprentissage profond (deep learning).

Une quatrième perspective concerne l'élargissement de la méthode de classification avec un seul classifieur à une méthode de classification collaborative ou distribuée où plusieurs classifieurs peuvent intervenir ensemble pour un meilleur taux de classification.

Enfin nous proposons une dernière perspective où nous mettons notre classifieur sur un site pour assurer un apprentissage en ligne avec des nouveaux tweets.

Nous rappelons que les résultats obtenus dans le cadre de ce projet de fin d'études font l'objet d'un projet d'article de recherche qui sera soumis prochainement à une revue scientifique.

Bibliographie

Dubey K., & Agrawal S. (2018). A Critical Analysis of Twitter Data for Movie Reviews Through ‘Random Forest’ Approach. *International Conference on Information and Communication Technology for Intelligent Systems.*

Ahmed Sulaiman M.Al. (2019). Analyse des sentiments sur Twitter avec un réseau de neurones profond: une approche améliorée utilisant les informations comportementales des utilisateurs. *Cognitive Systems Research*, 50-61.

Allport G.-W. , & Postman L. J. . (1945). The Basic Psychology of Rumor. *In Transactions of the New York Academy of Sciences, 8*.

Attardi G., Sartiano D., & Alzetta C. (s.d.). Convolutional Neural Networks for Sentiment Analysis on Italian.

BARBIERI F., & SAGGION H. (2014b). Modelling irony in twitter : Feature analysis and evaluation. *In Proceedings of Language Resources and Evaluation Conference (LREC)* , 4258–4264.

BURFOOT C., & BALDWIN C. (2009). Automatic satire detection: Are you having a laugh ? *In Proceedings of the ACL-IJCNLP 2009 conference*, 161–164.

Client, H. (2018). Récupéré sur https://docs.python.org/3/library/http.client.html

CLIFT R. (1999). Irony in conversation. *Language in Society*(28), 523–553.

Davidov D., Tsur O., & Rappoport A. (2010). Semi-Supervised Recognition of Sarcastic Sentences. *the Fourteenth Conference on Computational Natural Language Learning the Fourteenth Conference on Computational Natural Language Learning*, 107–116.

Delcambre A. (2016). Les réseaux sociaux prennent une place croissante dans l’accès à. *Le Monde économie*.

Eisenstein E. L. . (1991). La Révolution de l’imprimé dans l’Europe des premiers temps modernes. *La Découverte.*

El-Beltagy S. R. (2006). A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic.

Farasa. (2019). *http://alt.qcri.org*. Récupéré sur http://alt.qcri.org/farasa/

FARIAS D. I. H., SULIS E., PATTI V, RUFFO G., & BOSCO C. (2015). Valento : Sentiment analysis of figurative language tweets with irony and sarcasm. *SemEval-2015*, 694.

Georgiana Ifrim, Bichen Shi, & Igor Brigadir. (s.d.). Event Detection in Twitter using Aggressive Filtering. *Insight Centre for Data Analytics*.

Gerv´as J. C. (2014). An easily scalable concept-based affective lexicon.

GHERSEDINE, A., BUCHE, P., DIBIE-BARTH EL EMY, J., HERNANDEZ, N., & KAMEL, M. (2012). Extraction de relations n-aires inter phrastiques guidée par une RTO. *IRIT-IC3* .

GIANTI A., BOSCO C., PATTI V., BOLIOLI A., & CARO L. D. (2012). Annotating irony in a novel italian corpus for sentiment analysis. *In Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, 1-7.

GONZALEZ-IBANEZ R., MURESAN S., & WACHOLDE N. (2011). Identifying sarcasm in twitter : a closer look. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers, 2*, 581–586.

GRICE H. P., COLE P., & MORGAN J. L. (1975). Syntax and semantics. *Logic and conversation*, 41–58.

JIE TANG Y., & CHEN H.-H. (2014). Chinese Irony Corpus Construction and Ironic Structure Analysis. *In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, 1269-1278.

Kapferer J.-N. (1987). Rumeur. *Le plus vieux média du monde*.

KREUZ R. J., & CAUCCI G. M. (2007). Lexical influences on the perception of sarcasm. *In Proceedings of the Workshop on computational approaches to Figurative Language*, 1-4.

Larousse. (s.d.). *Opinion*. Récupéré sur https://www.larousse.fr/dictionnaires/francais/opinion/56197

LIEBRECHT C., KUNNEMAN F., & VAN DEN B. A. (2013). The perfect solution for detecting sarcasm in tweets# not. *In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 29-37.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*.

MERCIER , A., & PIGNARD-CHEYNEL, N. (2018). #Info.Commenter et partager l'actualité sur Twitter et Facebook. *Le bien commun*.

Morin E. (1969). La Rumeur d'Orléans.

Munir A., Shabib A., & Iftikhar A. (2017). Sentiment Analysis of Tweets using SVM. *International Journal of Computer Applications* .

NLTK. (2019). Récupéré sur https://www.nltk.org/

Python. (s.d.). *Csv*. Récupéré sur docs.python.org

REYES A., ROSSO P., & VEALE T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 239–268.

Ribeiro C. S. (2010). Inductive infeTence foT large scale text classification.

Ribeiro C. S. (2010). Inductive infeTence foT large scale text classification. *heidelberg: springer-verlag.*

Samah M. Alzanin , & Aqil M. (2019). Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization. *Knowledge-Based Systems*.

Sedhai S., & Sun A. (2017). Détection de spam semi-supervisée dans le flux Twitter. *IEEE Transactions on Computational Social Systems*.

SenticNet. (2012). Récupéré sur http://sentic.net/about/

SIGNORINO T. (2016). Comment les réseaux sociaux changent-ils le journalisme ? Les réseaux sociaux renforcent-ils l'accessibilité à l'information ?

Sintsova V. , Musat C., & Pu P. (2014). Semi-Supervised Method for Multi-category Emotion Recognition in Tweets. *l'atelier d'exploration de données.* Shenzhen.

Spacy. (2020). Récupéré sur https://spacy.io/

SPERBER D. , & WILSON D. (1981). Irony and the use-mention. *Radical pragmatics*, 295–318.

Statista. (2019, juin 14). *Facebook*. Récupéré sur https://fr.statista.com/statistiques/565258/facebook-nombre-d-utilisateurs-actifs-mensuels-dans-le-monde/

Stern W. (1902). Zur Psychologie der Aussage. Experimentelle Untersuchungen über Erinnerungstreue. *Zeitschrift für die gesamte Strafrechtswissenschaft, XXII*.

TSUR O., DAVIDOV D., & RAPPOPORT A. (2010). Icwsm-a great catchy name : Semi-supervised recognition of sarcastic sentences in online product reviews. *In ICWSM*.

Tweepy. (s.d.). Récupéré sur www.tweepy.org

Twitter. (2020). *developer*. Récupéré sur https://developer.twitter.com/

UTSUMI A. (1996). A unified theory of irony and its computational formalization. *In Proceedings of the 16th conference on Computational linguistics-Volume 2*, 962–967.

Vaithyanathan B. P. (2002). Sentiment classification using machine learning techniques. *EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10*.

Vaithyanathan B. P. (2002). sentiment classification using machine learning techniques. Stroudsburg. *EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10*.

w3schools. (s.d.). *Python regex*. Récupéré sur https://www.w3schools.com/python/python_regex.asp

Wikipedia. (2018, juillet 23). *Matplotlib* . Récupéré sur Wikipedia: https://fr.wikipedia.org/wiki/Matplotlib

Wikipedia. (2019 , août 6). *Apprentissage semi-supervisé*. Récupéré sur https://fr.wikipedia.org/wiki/Apprentissage_semi-supervisé

Wikipedia. (2019, novembre 28). *Arbre de décision (apprentissage)* . Récupéré sur https://fr.wikipedia.org/wiki/Arbre_de_décision

Wikipedia. (2019, décembre 9). *K-moyennes*. Récupéré sur https://fr.wikipedia.org/wiki/K-moyennes

Wikipedia. (2019, septembre 20). *Pandas*. Récupéré sur https://fr.wikipedia.org/wiki/Pandas

Wikipedia. (2019, décembre 10). *Scikit-learn*. Récupéré sur Wikipedia: https://fr.wikipedia.org/wiki/Scikit-learn

Wikipedia. (2020, Avril 1). *Anaconda*. Récupéré sur https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)

Wikipedia. (2020, février 5). *Apprentissage automatique*. Récupéré sur https://fr.wikipedia.org/wiki/Apprentissage_automatique

Wikipedia. (2020, mars 30). *Facebook*. Récupéré sur https://fr.wikipedia.org/wiki/Facebook

Wikipedia. (2020, mars 29). *Instagram*. Récupéré sur https://fr.wikipedia.org/wiki/Instagram

Wikipedia. (2020, février 8). *Lemmatisation*. Récupéré sur Wikipedia: https://fr.wikipedia.org/wiki/Lemmatisation

Wikipedia. (2020, mars 27). *LinkedIn*. Récupéré sur https://fr.wikipedia.org/wiki/LinkedIn

Wikipedia. (2020, mars 25). *Pinterest*. Récupéré sur https://fr.wikipedia.org/wiki/Pinterest

Wikipedia. (2020, avril 1). *Python*. Récupéré sur https://fr.wikipedia.org/wiki/Python_(langage)

Wikipedia. (2020, janvier 2020). *Réseau de neurones artificiels*. Récupéré sur https://fr.wikipedia.org/wiki/Réseau_de_neurones_artificiels

Wikipedia. (2020, février 20). *Réseau social*. Récupéré sur https://fr.wikipedia.org/wiki/Réseau_social

Wikipedia. (2020, février 27). *Sentiment*. Récupéré sur https://fr.m.wikipedia.org/wiki/Sentiment

Wikipedia. (2020, janvier 27). *Tumblr*. Récupéré sur https://fr.wikipedia.org/wiki/Tumblr

Wikipedia. (2020, mars 22). *Twitter*. Récupéré sur https://fr.wikipedia.org/wiki/Twitter

Wikipedia. (2020, février 21). *Viadeo*. Récupéré sur https://fr.wikipedia.org/wiki/Viadeo

Annexe 1

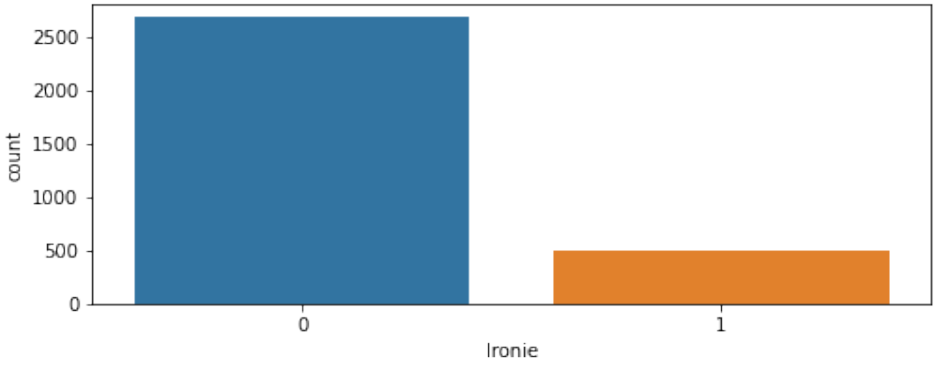

Figure 23 - Distribution des tweets du 1er sous corpus supervisé

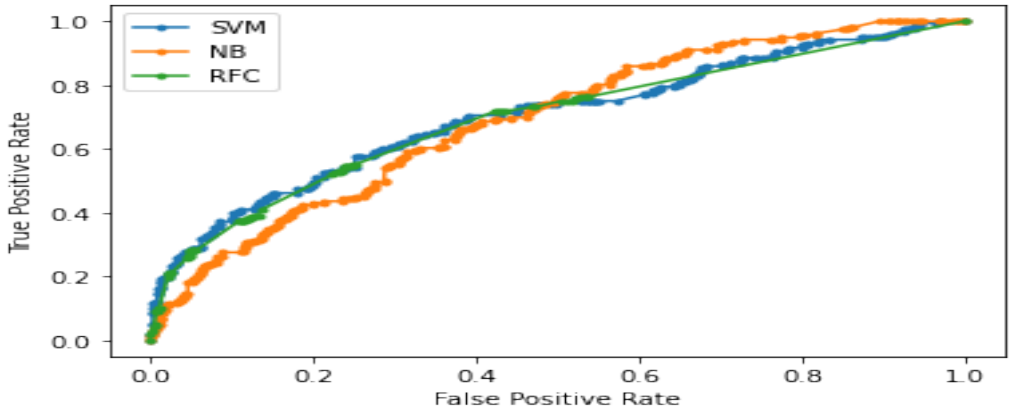

Figure 24 - La courbe ROC du 1er sous corpus supervisé

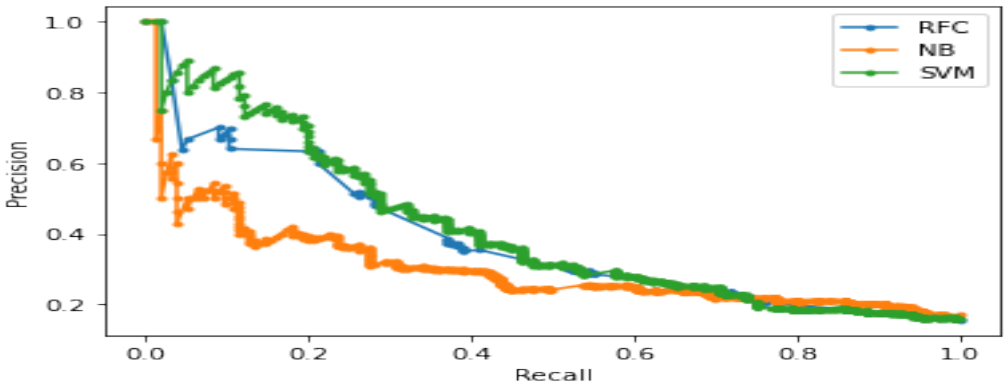

Figure 25 - La courbe Précision/Rappel du 1er sous corpus supervisé

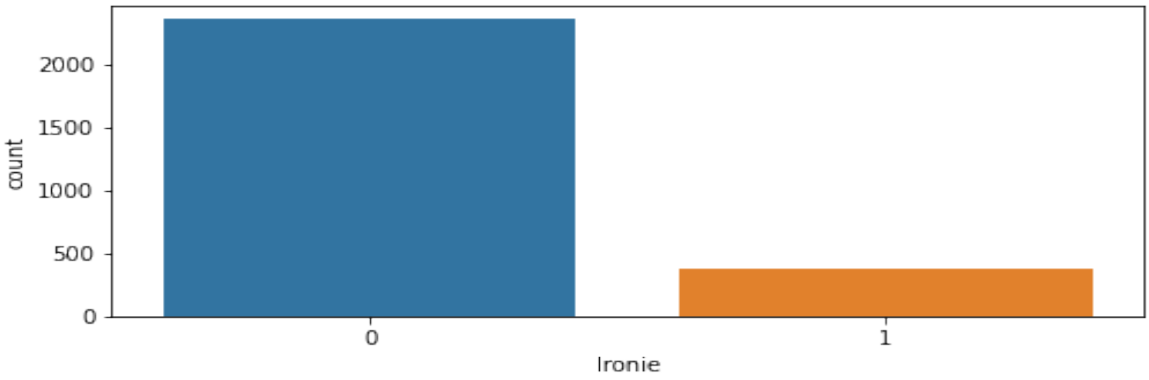

Figure 26 - Distribution des tweets du 2ème sous corpus supervisè

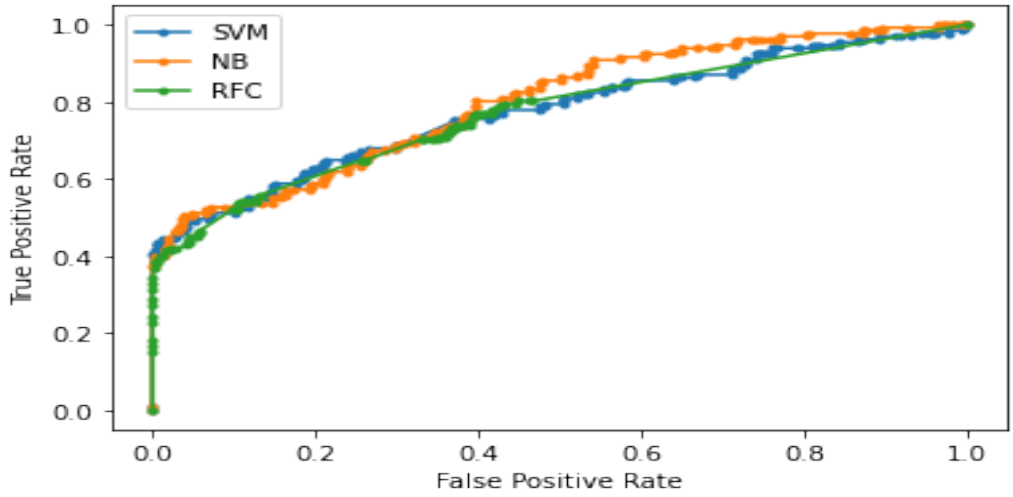

Figure 27 - La courbe ROC du 2ème sous corpus supervisé

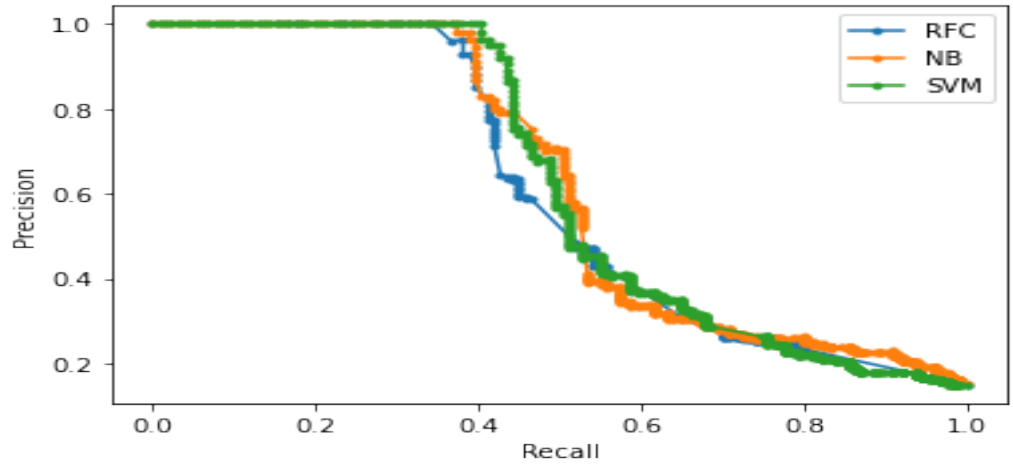

Figure 28 - La courbe Précision/Rappel du 2ème sous corpus supervisé

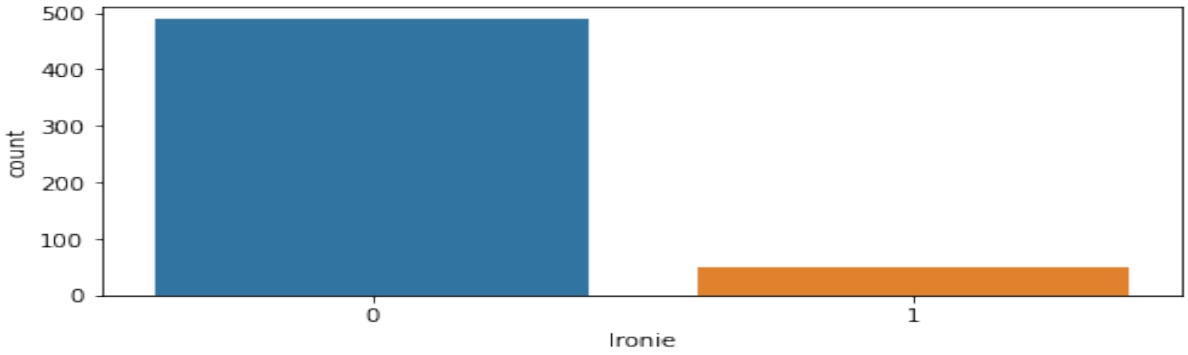

Figure 29 - Distribution des tweets du 3ème sous corpus supervisé

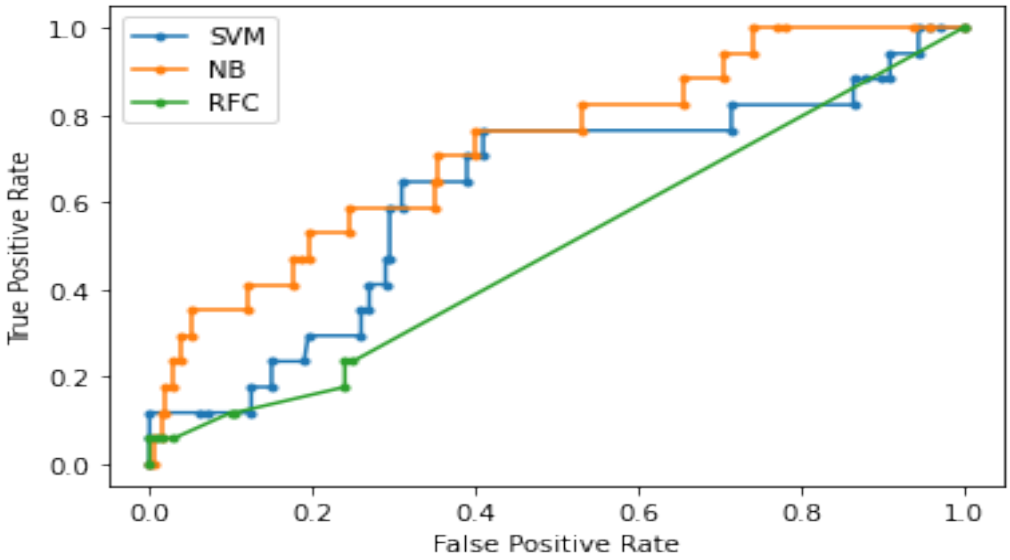

Figure 30 - La courbe ROC du 3éme sous corpus supervisé

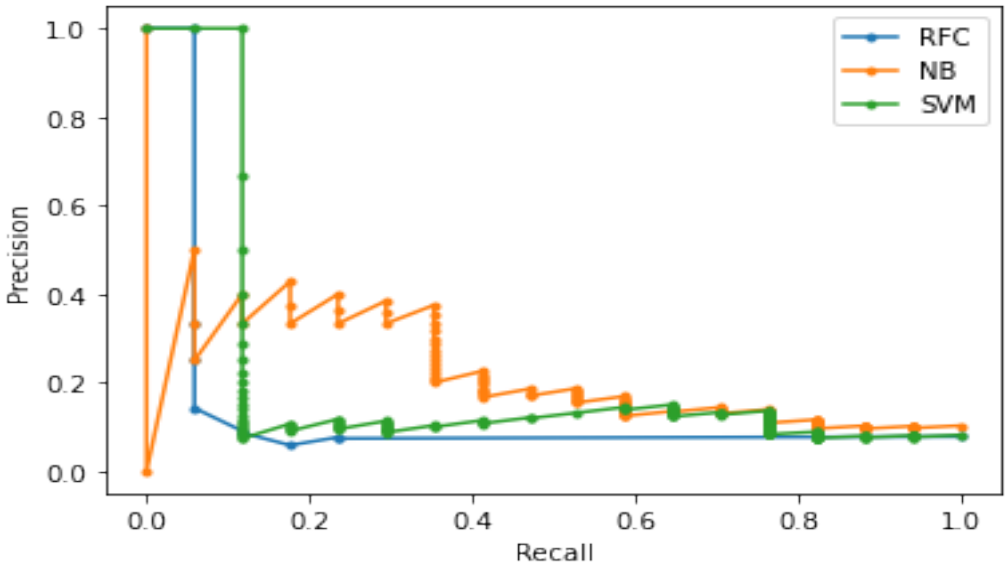

Figure 31 - La courbe Précision/Rappel du 3ème sous corpus supervisé

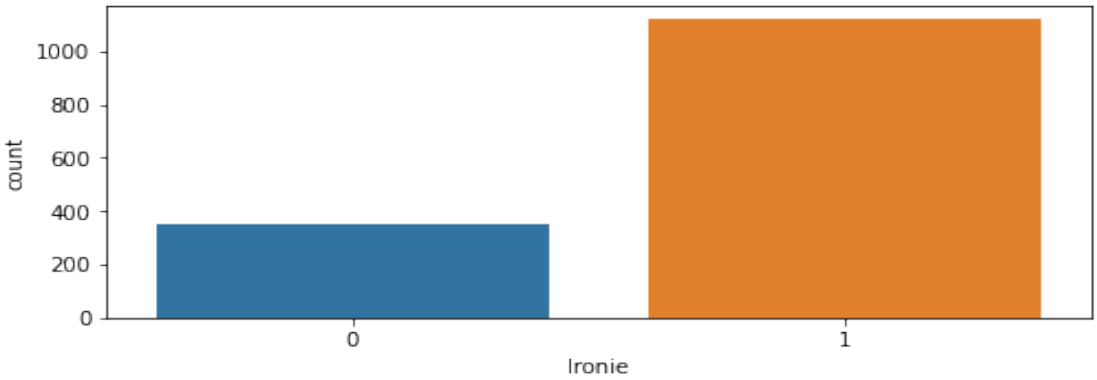

Figure 32 - Distribution des tweets du corpus Ironique supervisé

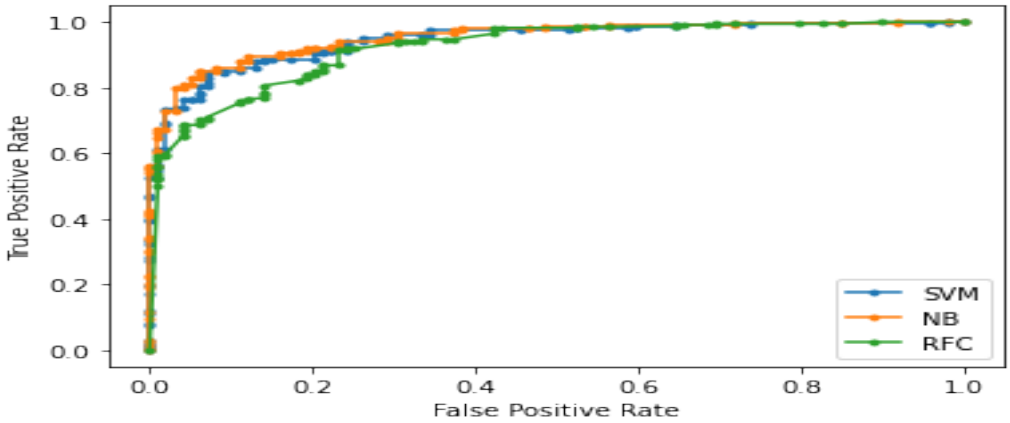

Figure 33 - La courbe ROC du corpus Ironique supervisé

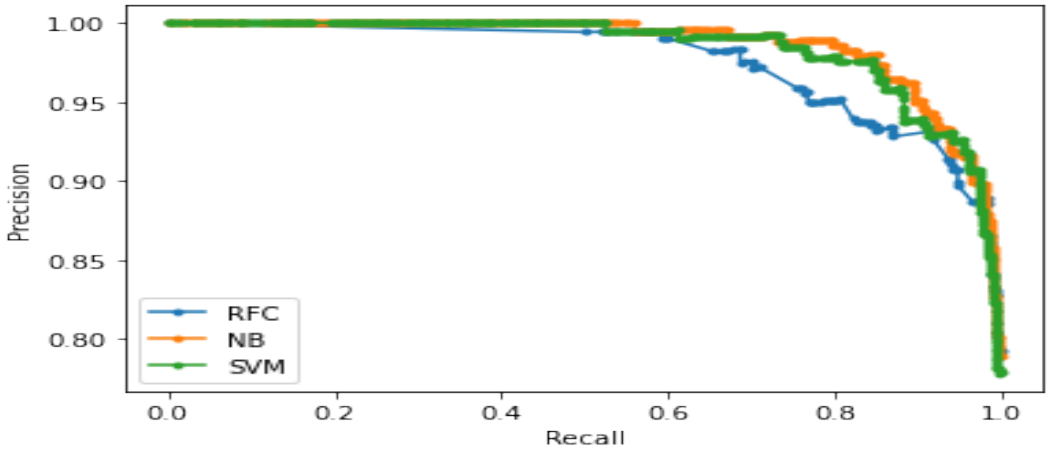

Figure 34 - La courbe Précision/Rappel du corpus Ironique supervisé

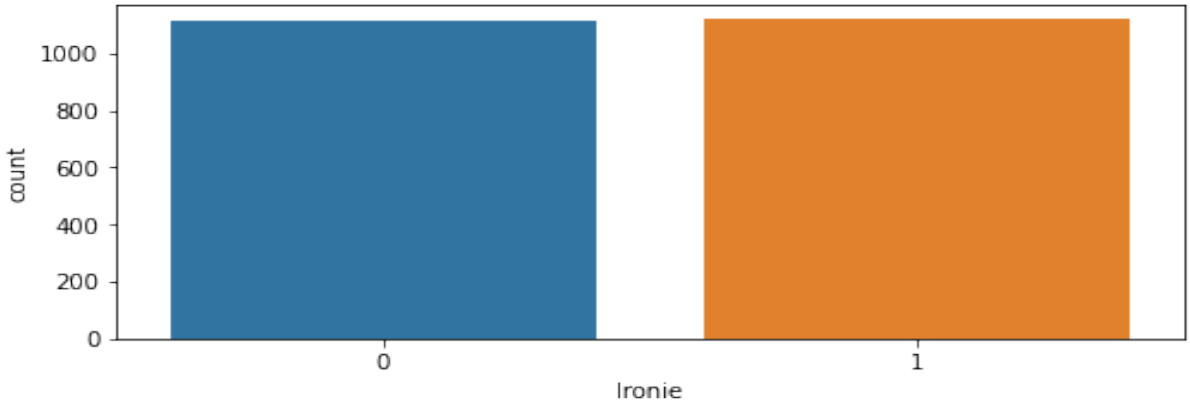

Figure 35 - Distribution des tweets du corpus globale supervisé

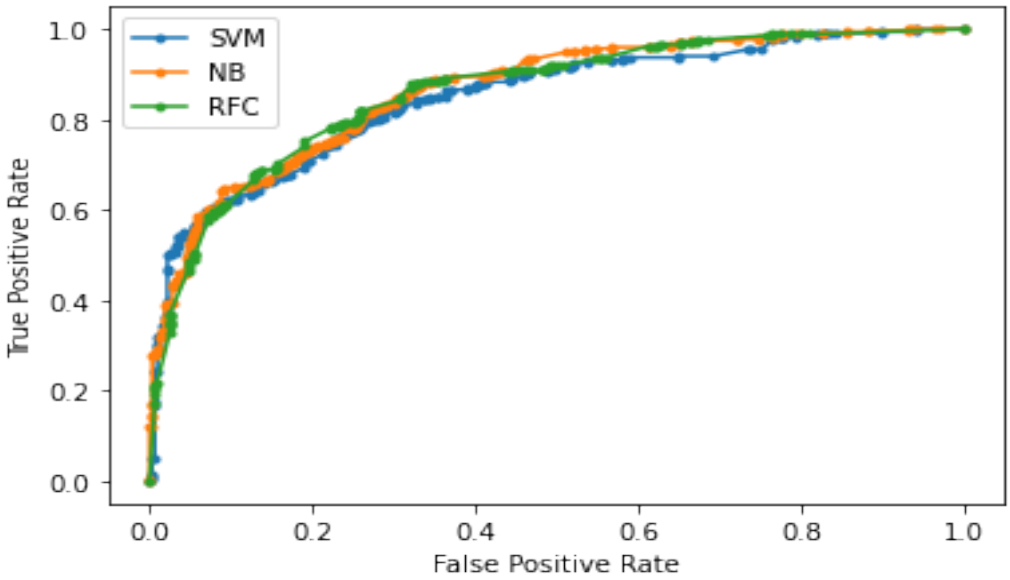

Figure 36 - La courbe ROC du corpus global supervisé

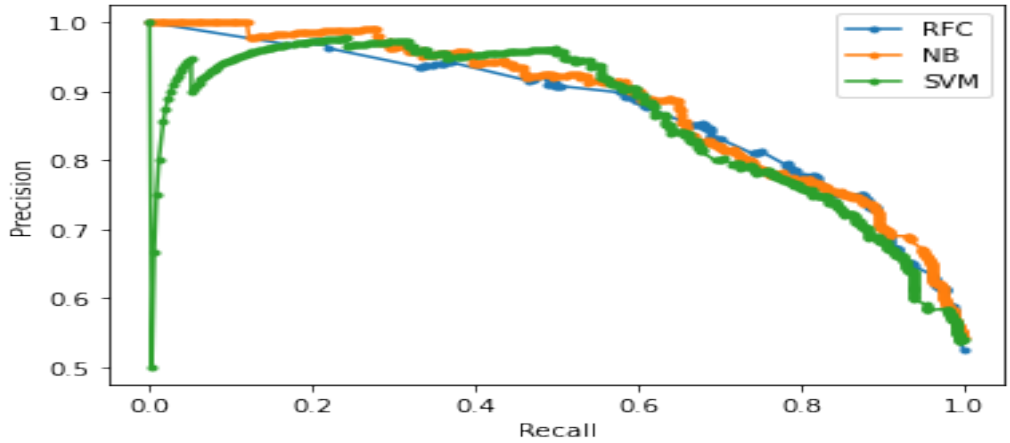

Figure 37 – la courbe Précision/Rappel globale supervisé

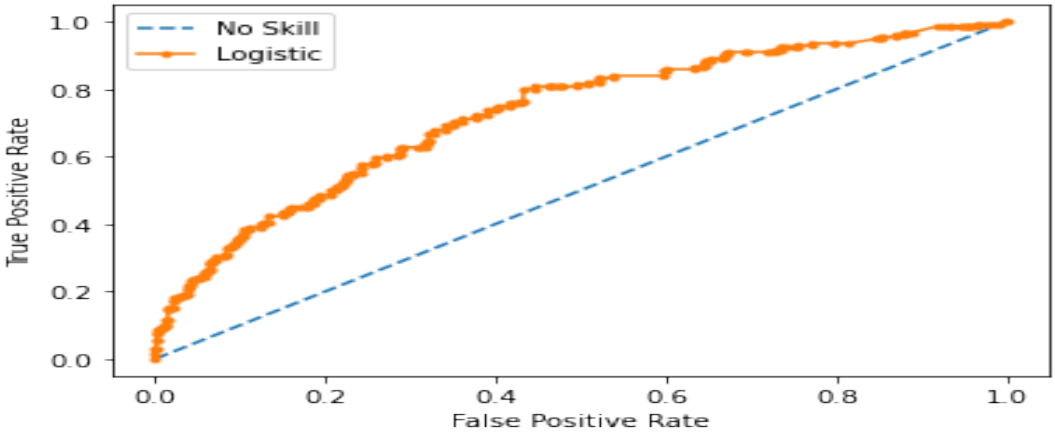

Figure 38 - La courbe ROC du 1er sous corpus

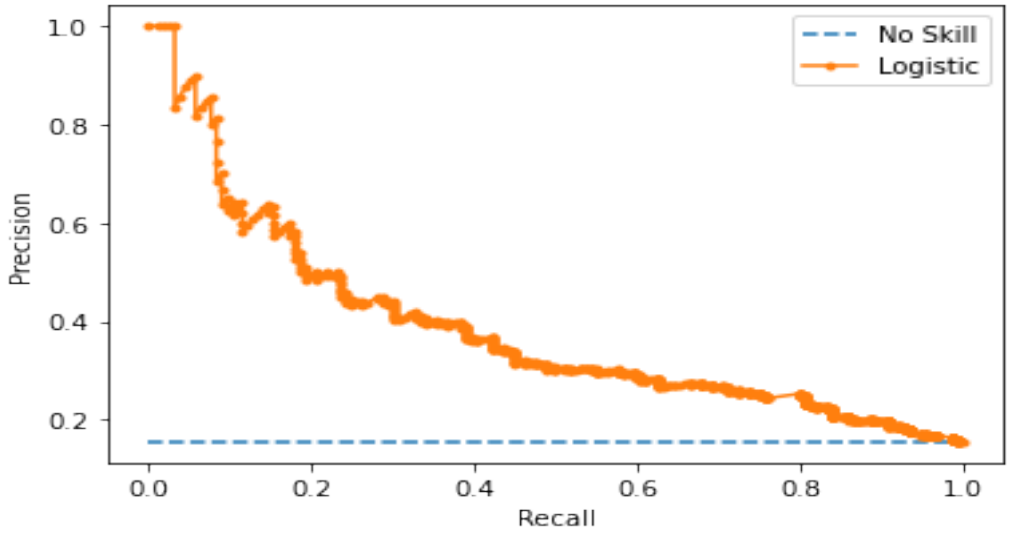

Figure 39 - La courbe Précision/Rappel du 1er sous corpus

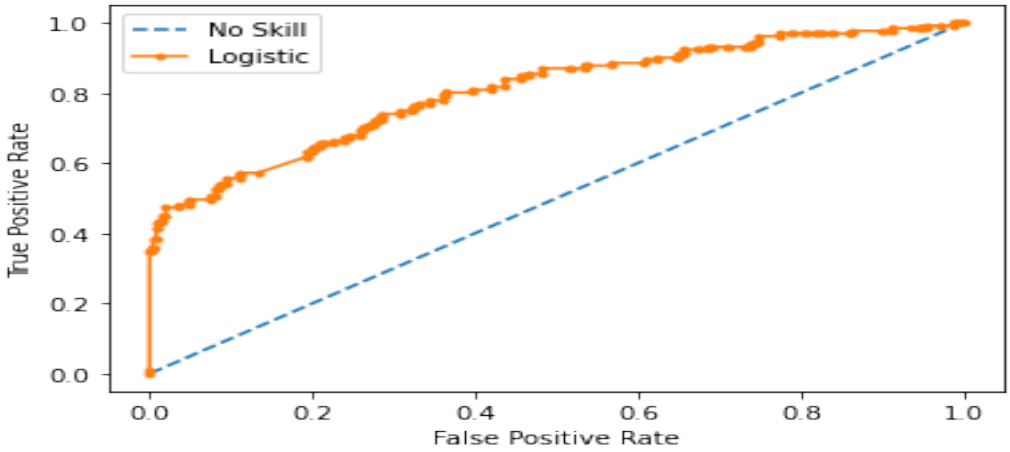

Figure 40 - La courbe ROC du 1er sous corpus

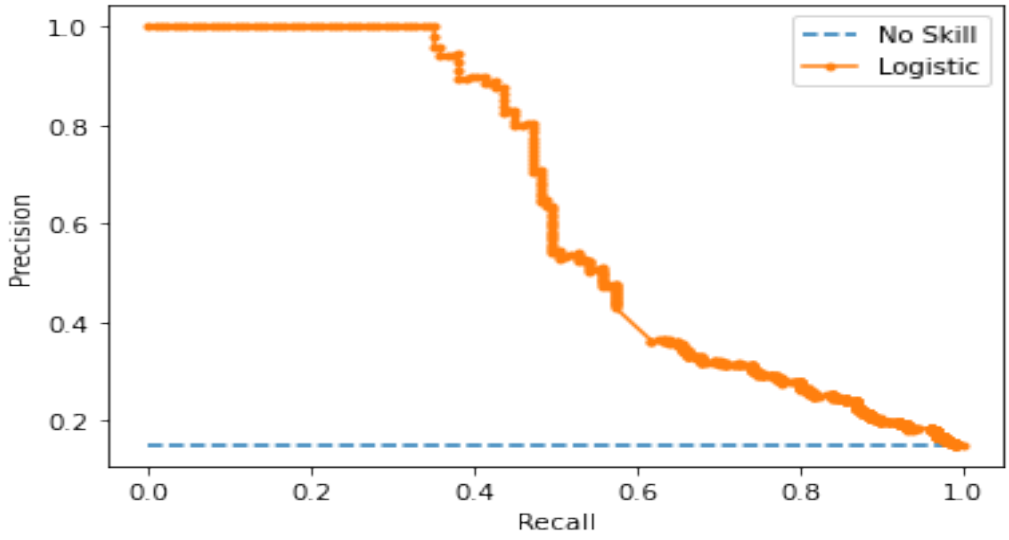

Figure 41 - La courbe Précision/Rappel du 1er sous corpus

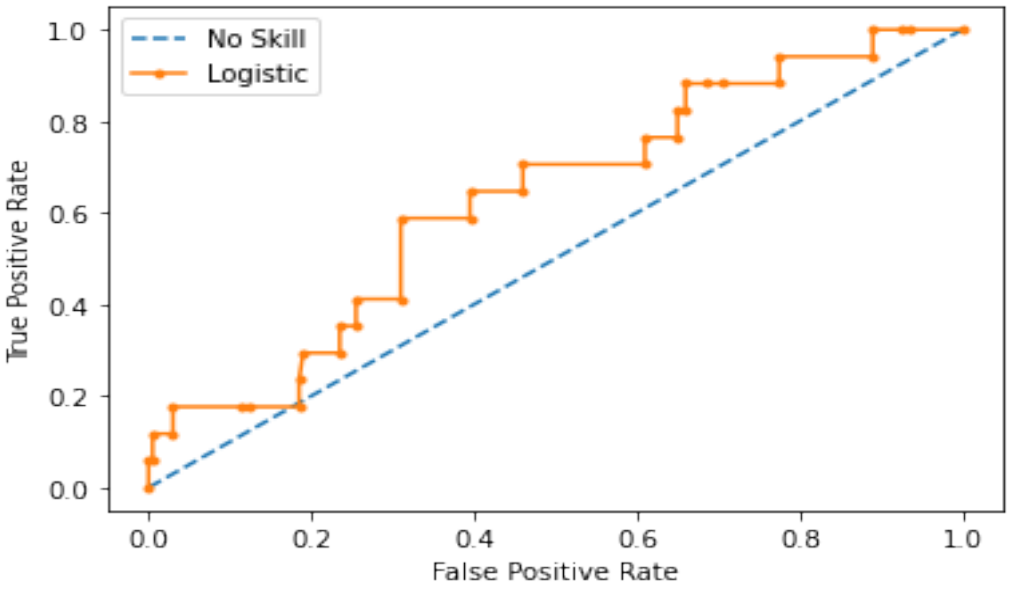

Figure 42 - La courbe ROC du 1er sous corpus

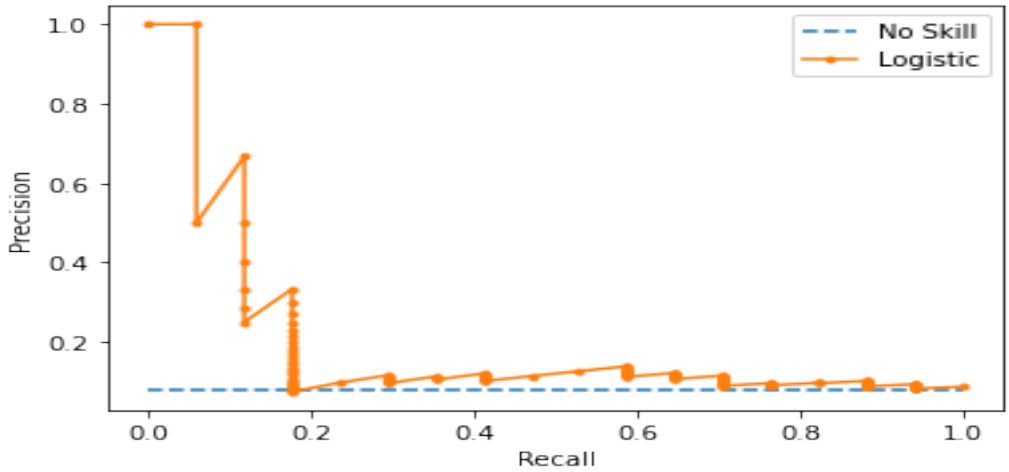

Figure 43 - La courbe Précision/Rappel du 1er sous corpus

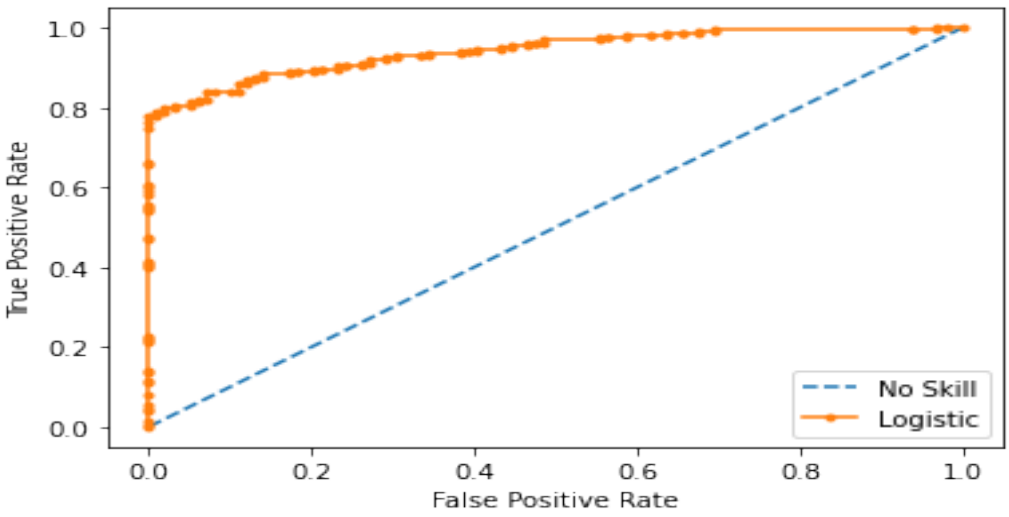

Figure 44 - La courbe ROC du corpus global

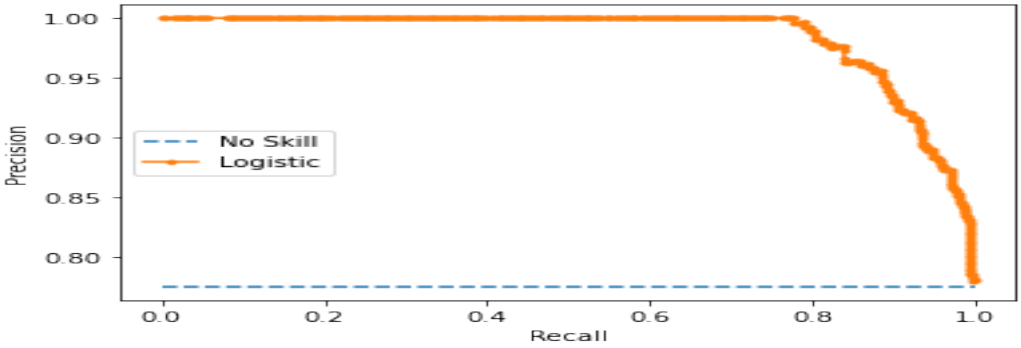

Figure 45 - La courbe Précision/Rappel du corpus global

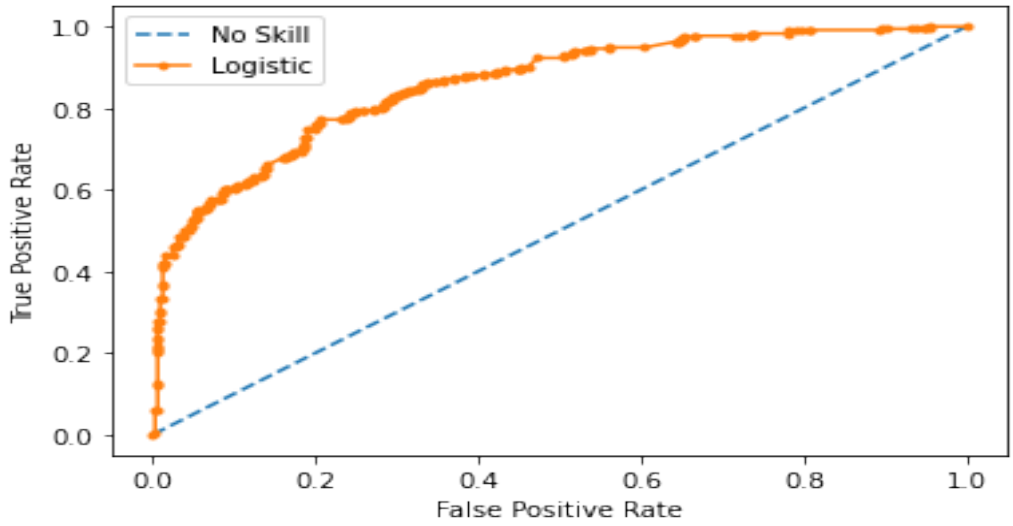

Figure 46 - La courbe ROC du corpus global

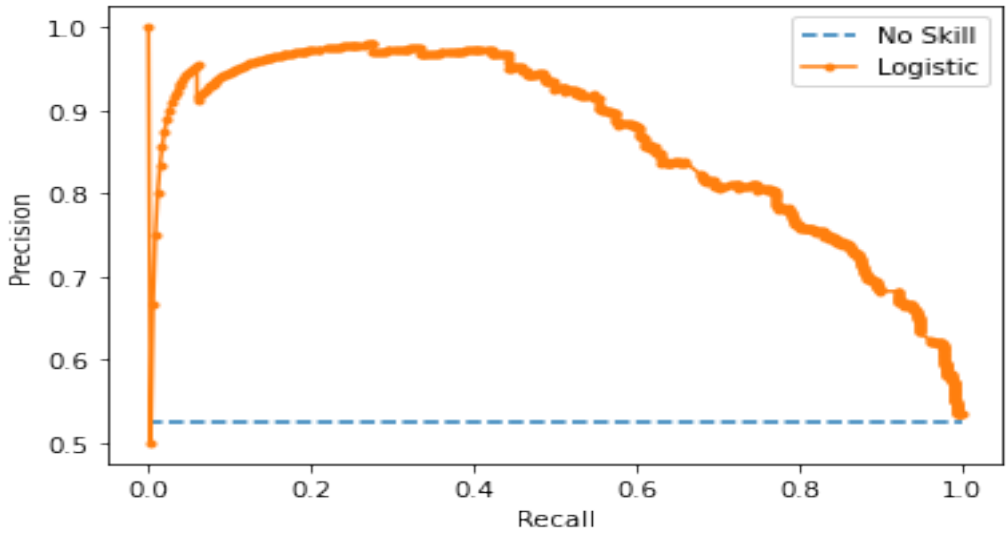

Figure 47 - La courbe Précision/Rappel du corpus global

Annexe 2

	Précision	Rappel	F1-score	Accuracy
Naive Bayes	0.69	0.53	0.5995	0.84
SVM	0.625	0.57	0.5962	0.82
RFC	0.75	0.5	0.6	0.85

Tableau 13 - Deuxième expérimentation 1er sous corpus

	Précision	Rappel	F1-score	Accuracy
Naive Bayes	0.895	0.7	0.7855	0.91
SVM	0.87	0.715	0.7849	0.91
RFC	0.905	0.705	0.7925	0.91

Tableau 14 - Deuxième expérimentation 2ème sous corpus

	Précision	Rappel	F1-score	Accuracy
Naive Bayes	0.7	0.545	0.6128	0.89
SVM	0.59	0.525	0.5556	0.91
RFC	0.965	0.535	0.6883	0.93

Tableau 15 - Deuxième expérimentation 3éme sous corpus

	Précision	Rappel	F1-score	Accuracy
Naive Bayes	0.83	0.89	0.8589	0.89
SVM	0.86	0.8	0.8289	0.88
RFC	0.83	0.775	0.8015	0.86

Tableau 16 - Troisième expérimentation

	Précision	Rappel	F1-score	Accuracy
Naive Bayes	0.775	0.775	0.775	0.77
SVM	0.76	0.75	0.7549	0.75
RFC	0.735	0.715	0.7248	0.72

Tableau 17 - Quatrième expérimentation

	Précision	Rappel	F1-score	Accuracy
K-means	0.35	0.30	0.38	0.40

Tableau 18 - Deuxième expérimentation 1er sous corpus

	Précision	Rappel	F1-score	Accuracy
K-means	0.39	0.28	0.34	0.36

Tableau 19 - Deuxième expérimentation 2ème sous corpus

	Précision	Rappel	F1-score	Accuracy
K-means	0.43	0.5	0.47	0.46

Tableau 20 - Deuxième expérimentation 3ème sous corpus

	Précision	Rappel	F1-score	Accuracy
K-means	0.49	0.42	0.45	0.48

Tableau 21 - Troisième expérimentation

	Précision	Rappel	F1-score	Accuracy
K-means	0.52	0.48	0.50	0.53s

Tableau 22 - Quatrième expérimentation

	Précision	Rappel	F1-score	Accuracy
SVM	0.625	0.57	0.59	0. 82445

Tableau 23 - Deuxième expérimentation 1er sous corpus

	Précision	Rappel	F1-score	Accuracy
SVM	0.87	0.715	0.67	0. 91481

Tableau 24 - Deuxième expérimentation 2ème sous corpus

	Précision	Rappel	F1-score	Accuracy
SVM	0.59	0.525	0.53	0. 91165

Tableau 25 - Deuxième expérimentation 3ème sous corpus

	Précision	Rappel	F1-score	Accuracy
SVM	0.86	0.8	0.825	0.88

Tableau 26 - Troisième expérimentation

	Précision	Rappel	F1-score	Accuracy
SVM	0.76	0.75	0.75	0. 7791

Tableau 27 - Quatrième expérimentation

ملخص

يعد تحليل البيانات على وسائل التواصل الاجتماعي مجالًا مثيرًا للاهتمام للبحث. إن أهمية هذه البيانات كبيرة بالنسبة لمؤسسات الدولة والمؤسسات التجارية، التي ترغب على التوالي في الحصول على رأي من المواطنين حول حدث سياسي معين أو ملاحظات العملاء على منتجات محددة للغاية. في إطار أطروحة الماستير، نركز عملنا على تحديد السخرية في نوع معين من البيانات، وهي التغريدات. في هذا السياق، نحن مهتمون ببناء مجموعة تتكون من 5874 تغريدة جزائرية، 70٪ منها ساخرة والباقي غير ساخرة. للقيام بذلك، ثلاث مراحل. في البداية، كنا مهتمين بجمع التغريدات الجزائرية وشروحها. في الخطوة الثانية، قمنا بعملية معالجة البيانات مثل الترميز و أخيرًا lemmatisation ، في الخطوة الثالثة، طبقنا المصنفات على أساس التعلم الآلي للكشف التلقائي عن التغريدات الجزائرية الساخرة. النتائج التي تم الحصول عليها من خلال هذه المذكرة مشجعة وواعدة للغاية وفتحت لنا خطوط بحث مستقبلية خاصة في مجال تحليل الشعور.

كلمات مفتاحية: شبكات اجتماعية، تغريدات جزائرية، مدونة، معالجة مسبقة، تصنيف، تعلم اصطناعي.

Abstract

Analyzing data on social media is a very interesting area of research. The importance of this data is considerable for state institutions and commercial enterprises, which respectively wish to obtain an opinion from citizens on a particular political event or customer feedback on very specific products. Within the framework of this master's thesis, we focus our work on identifying irony in a particular type of data, namely tweets. In this context, we are interested in building a corpus made up of 8178 Algerian tweets, 70% of them are ironic and the rest are non-ironic. To do this, we followed a three-phase approachs. At first, we were interested in the collection of Algerian tweets and their annotations. In a second step, we apply a data preprocessing operation such as tokenization and limmetasation and stop words elimination. Finally, in the third step, we implemented classifiers based on machine learning for automatic irony detection of Algerian tweets. The results obtained for this work are very promising and open future lines of research especially in the domain of feeling analysis.

Key words: social networks, Algerian tweets, corpus, preprocessing, classification, artificial learning.

Resumé

L’analyse des données sur les réseaux sociaux est un domaine de recherche en plein ébullition. L’importance de ces données est considérable pour les institutions étatiques et les entreprises commerciales qui souhaitent obtenir respectivement un avis des citoyens sur un événement politique particulier ou un retour client sur des produits bien spécifiques. Dans le cadre de ce mémoire de master, nous nous focalisons sur l’identification de l’ironie dans un type particulier de données à savoir les tweets. Dans ce cadre, nous nous sommes intéressés à la construction d’un corpus formé de 8178 tweets algériens dont **70%** sont ironiques et le reste sont non ironiques. Pour ce faire, nous avons suivi une démarche en trois phases. Dans un premier temps, nous nous sommes intéressés à la collecte des tweets algériens et leurs annotations. Dans une deuxième étape, nous avons mené une opération de prétraitement de données comme la tokenisation, la limmetasation et l’élimination des mots vides. Enfin, dans la troisième étape, nous avons implémenté des classifieurs à base d’apprentissage artificiel pour la détection automatique de l’ironie pour les tweets algériens. Les résultats obtenus pour cette tâche sont très prometteurs et ouvrent des pistes de recherche lors de l’analyse future de sentiments.

Mots clés : réseaux sociaux, tweets algériens, corpus, prétraitement, classification, fouille de données, apprentissage artificiel.