Table des matières

INTRODUCTION 1
I. CONTEXTE GENERAL 2
II. OJECTIFS : 5
III. ANALYSE DE FAISABILITE : 6
1. Forces et faiblesses de l'IA dans la détection de fake-news 6
2. Menaces et Opportunités 10
IV. PLANIFICATION : 12
1. Définition des challenges initiaux : 13
2. Le cadrage stratégique : 16
3. La création d'algorithmes modulaires : 16
CONCLUSION 17

INTRODUCTION

Aujourd'hui, nous vivons dans un monde où l'information recèle une importance décisive dans les plus hautes sphères décisionnelles. Elle circule à une vitesse insoupçonnée grâce aux technologies de l'information et de la communication : un accident survenu aux USA peut faire le tour du monde en moins de cinq minutes car, rien qu'avec un smartphone, tout le monde peut être consommateur et/ou producteur d'informations et donc chacun est susceptible de manipuler l'information et de la remodeler à sa guise.
dans un tel monde, où le développement des TIC[1] connaît un succès fulgurant, s'inscrit l'interrogation sur la fiabilité et la véracité de ces informations d'où l'entrée en scène du terme « fake news »

Le terme « fake news » désigne toute information volontairement fausse et diffusée avec un but précis, pour provoquer une réaction qu'on a cherché à anticiper. Ce phénomène a toujours été présent dans la société et s'est manifesté sous diverses formes à travers les époques jusqu'à son nom actuel : « fake news ». Elle est la source de multiples problèmes à travers le monde.

Pour identifier les fausses informations considérées comme vraies, bon nombre de monde s'emploie à des techniques diverses ; d'où notre projet de détection de fake news par l'intelligence artificielle.
Le terme « Intelligence artificielle » désigne l'ensemble des théories et des techniques développant des programmes informatiques complexes capables de simuler certains traits de l'intelligence humaine (raisonnement, apprentissage…).

Dans quel contexte les fake news constituent-ils un danger pour le monde actuel ?

Est-il possible de les discerner ?

Comment ?
pour mieux satisfaire aux exigences de cette problématique, nous allons d'abord présenter le contexte général de la situation. Ensuite, nous montrerons, si possible, comment on compte les fake news dans la mêlée.

I. CONTEXTE GENERAL

[1] Technologies de l'information et de la communication

Les fake news ont pris une très grande ampleur au sein de notre société. La diffusion de telles informations prend un effet boule neige, c'est-à-dire qu'elle grossit à mesure qu'elle circule au point où on ne parvient même plus à identifier la source. Elles ont une si grande ampleur au point d'occasionner des conséquences néfastes sur les points économiques, politiques, sociales et sanitaires.

Tout d'abord, l'impact des fake news sur l'économie mondiale est estimé à 78 milliards de dollar par an au point que le forum économique mondiale classe la diffusion de la désinformation et des fakes news parmi les principales causes d'instabilité de l'économie mondial. Cette somme est répartie comme suit :

- 39 milliards de dollars en valeurs boursières
- 17 milliards de dollars à cause de la désinformation financière rien qu'aux États-Unis
- 9,54 milliards de dollars en protection de la réputation sur Internet
- 9 milliards de dollars à cause de la désinformation en matière de santé
- 3 milliards pour la sécurité des plateformes
- 400 millions de dollars en dépense "politique"
- 250 millions de dollars en protection des marques

Il existe aussi des gens qui diffusent des « fake news » dans une optique purement mercantile, pour faire de l'argent. Une « fake news » peut ainsi être conçue comme « appeau à clics » pour attirer les consultations des internautes et accroître les revenus publicitaires d'une page web.

- Elles sont parfois utilisées dans l'hameçonnage par courriel, en présentant du contenu très attractif ou sensationnaliste pour inciter les utilisateurs à cliquer sur un lien, ce qui permet ensuite à l'envoyeur d'infecter leur ordinateur.
- Elles peuvent être le fait d'un site humoristique qui lance un canular. Exemple : le « projet » de la dirigeante de l'extrême droite Marine Le Pen « d'entourer la France d'un mur payé par l'Algérie » inventé par le site parodique *Le Gorafi*, repris par erreur dans un journal algérien.

Ensuite vient le domaine politique qui n'est pas en reste car selon la même étude , 400 millions de dollars sont associés à la désinformation politique dans le monde. Selon l'étude donc, au moins 400 millions de dollars ont été dépenses pour la diffusion de fake news dans les

campagnes politiques lors des dernières élections en Inde (140 millions de dollars en 2019), au Brésil (34 millions de dollars en 2018), au Royaume-Uni (1 million de dollars), en France (586.000 euros), en Australie (828.000 dollars), au Kenya (20 millions de dollars), en Afrique du Sud (2,7 millions de dollars) ou au Mexique (642.000 dollars) par exemple.
Et les États-Unis ne sont pas en reste puisqu'une étude menée par l'université de Princeton sur la consommation de fake news pendant la campagne américaine de 2016 a révélé que les fausses informations représentaient 2,6% de tous les articles d'actualité publiés lors de la fin de la dernière course à la présidence en 2016 remportée par Donald Trump.
Par conséquent, sur la base des volumes actuels et estimés de fausses nouvelles, l'étude révèle que 200 millions de dollars devraient être dépensés pour produire, diffuser et rendre virales des fake news lors de l'élection présidentielle américaine qui devait suivre en 2020.

De plus les fake news ont aussi des impacts majeurs sur la société. Les gouvernements s'en servent pour maintenir le peuple sous contrôle à travers la diffusion de fausses nouvelles visant soit à intimider, soit à rassurer la population.
A titre d'exemple, un des plus célèbres exemples de « fake news » reste le canular radiophonique d'Orson Wells, *La guerre des mondes*, diffusé le 30 octobre 1938. Pour des dizaines de milliers d'auditeurs, cette émission annonçait réellement le débarquement sur Terre d'une horde de Martiens agressifs. Cette émission a entraîné un mouvement de panique à travers les États-Unis même si elle a été très exagérée par rapport à ce qui s'est vraiment passé

Enfin, l'incidence des fake news sur le domaine de la santé est encore pire que ce qu'on a eu à voir jusque-là. En effet, dans le cadre d'une étude publiée dans *The American Journal of Tropical and Hygiene*, des chercheurs parlent même d'une « Infodémie liée au COVID-19 ». « Nous avons suivi et examiné les rumeurs, la stigmatisation et les théories du complot liées au COVID-19 circulant sur les plateformes en ligne, y compris les sites web des agences de vérification des faits, Facebook, Twitter et les journaux en ligne, et leurs impacts sur la santé publique », expliquent-ils.

À partir de 2 311 articles sur des rumeurs, théories conspirationnistes et stigmatisations, en 25 langues et issus de 87 pays, ils ont établi un premier bilan de l'influence des fausses informations sur les comportements. Ils se sont concentrés sur la première période de la crise, du 31 décembre 2019 au 5 avril dernier. **Concernant les «** *fake news* **» analysées, trois pays se démarquent pour leur caractère prolifique : l'Inde, les États-Unis et la Chine.** Thématiquement, elles se répartissent de la façon suivante : la maladie, sa transmission et la

mortalité (24 %), les mesures de contrôle (21 %), le traitement et la guérison (19 %), la cause de la maladie, y compris l'origine (15 %), la violence (1 %) et divers (20 %).

Des centaines de morts, des milliers de personnes hospitalisées

Pour s'auto-diagnostiquer, des rumeurs invitaient par exemple à retenir sa respiration pendant 10 secondes. Des « traitements » plus farfelus les uns que les autres ont émergé partout sur le globe. Le mythe selon lequel la consommation d'alcool concentré permettrait d'éradiquer le virus a fait, selon la recherche, plus de 800 morts et 5 876 personnes ont été hospitalisées. **Une soixantaine ont perdu complètement la vue en ingurgitant du méthanol.** En Arabie-Saoudite, certains ont recommandé la consommation d'urine de chameau mélangée à de la chaux. En Inde, le thé agrémenté d'urine ou de bouse de vache a été vu comme le traitement miracle.

Les fausses nouvelles ont également été le terreau de stigmatisations de certaines populations et de racisme à travers les personnes d'origine asiatique. « Au cours de la pandémie, il y a eu des récits répétés d'abus verbaux et physiques contre des personnes d'origine asiatique et de la part de personnes impliquées dans des activités de santé », exposent les scientifiques. Ils soulignent d'autre part le refus de certains professionnels de santé d'accueillir les patients atteints du Covid-19, notamment en Ouganda.

À la vue de tous ces problèmes liés aux infox, naquit notre idée de projet à savoir : « **la détection des fake news par l'intelligence artificielle »**

II. OJECTIFS :

La désinformation présente un danger pour différentes sphères de l'activité humaine et de la société en général. Les fake news constituent un frein pour la démocratie, la santé publique et l'économie.

Les internautes propagent les fake-news, sans vérifier, au préalable, la fiabilité de celles-ci ; c'est de là que notre projet prend tout son sens. Notre but est de mettre à la disposition des internautes sénégalais un outil accessible est facile d'utilisation et leurs permettre ainsi de vérifier la véracité de toutes les nouvelles susceptibles d'avoir un impact social, économique et géopolitique au sein du territoire national.

Notre modèle[2] de détection de fake news aura comme objectifs principaux :

- L'analyse automatique du contenu des réseaux sociaux et de la presse électronique ;
- La détection des infractions commises par les médias audiovisuels et par les acteurs politiques sur les réseaux sociaux et l'espace web ;
- La détection de la désinformation sur le Web de surface ;
- La disposition d'une base de données nationale sur la désinformation et sur le contenu des réseaux sociaux et de la presse électronique.

L'objectif premier de ce projet sera donc, à travers l'IA, de stimuler l'esprit critique des internautes, devenus léthargiques face à l'importante quantité de flux d'informations circulant sur le net.

Puis, s'ajoutera à cela la mise en place d'un centre de recherche spécialisé dans la détection de fake news. Notre centre regroupera des outils d'intelligence artificielle ainsi que des experts dans différents domaines à savoir : l'informatique, les médias, le domaine psychosocial, la politique, l'écologie et l'économie.

Cela permettra d'avoir un contrôle du flux d'informations circulant sur la toile. Ainsi ces informations devront dès lors être classées comme vraies ou fausses.

III. ANALYSE DE FAISABILITE :

Pour un meilleur aperçu de notre projet, nous proposons une analyse SWOT, l'analyse SWOT (Strength, Weaknesses, Opportunities, Threats) est un outil d'analyse stratégique. Il combine l'étude des forces et des faiblesses du projet avec celles des opportunités et des menaces de son environnement, afin d'aider à la définition d'une stratégie de développement.

1. Forces et faiblesses de l'IA dans la détection de fake-news

Forces

- De nos jours, des techniques d'intelligence artificielle peuvent aider à repérer les textes similaires, en tenant compte du contenu, mais aussi du canal de diffusion, de la personne qui relaie le message, et d'autres éléments contextuels, par exemple les images et illustrations. On est alors proche du mode de fonctionnement des moteurs de recherche :

[2] Un modèle de Machine Learning est le résultat final qui nous permettra d'obtenir des prédictions à partir des données en entrées. On lui fournit un problème et il trouve des solutions.

les modèles de recherche d'information actuels sont plutôt efficaces pour retrouver des textes similaires, même s'ils n'emploient pas exactement les mêmes mots ou les mêmes tournures :

Mais le but ultime serait évidemment de repérer directement les fake news par des moyens automatiques. Ceci semble en fait extrêmement difficile en l'état des choses et, si on écarte les faux grossiers, même un humain aura du mal à caractériser certains textes. Plusieurs techniques sont explorées en intelligence artificielle.

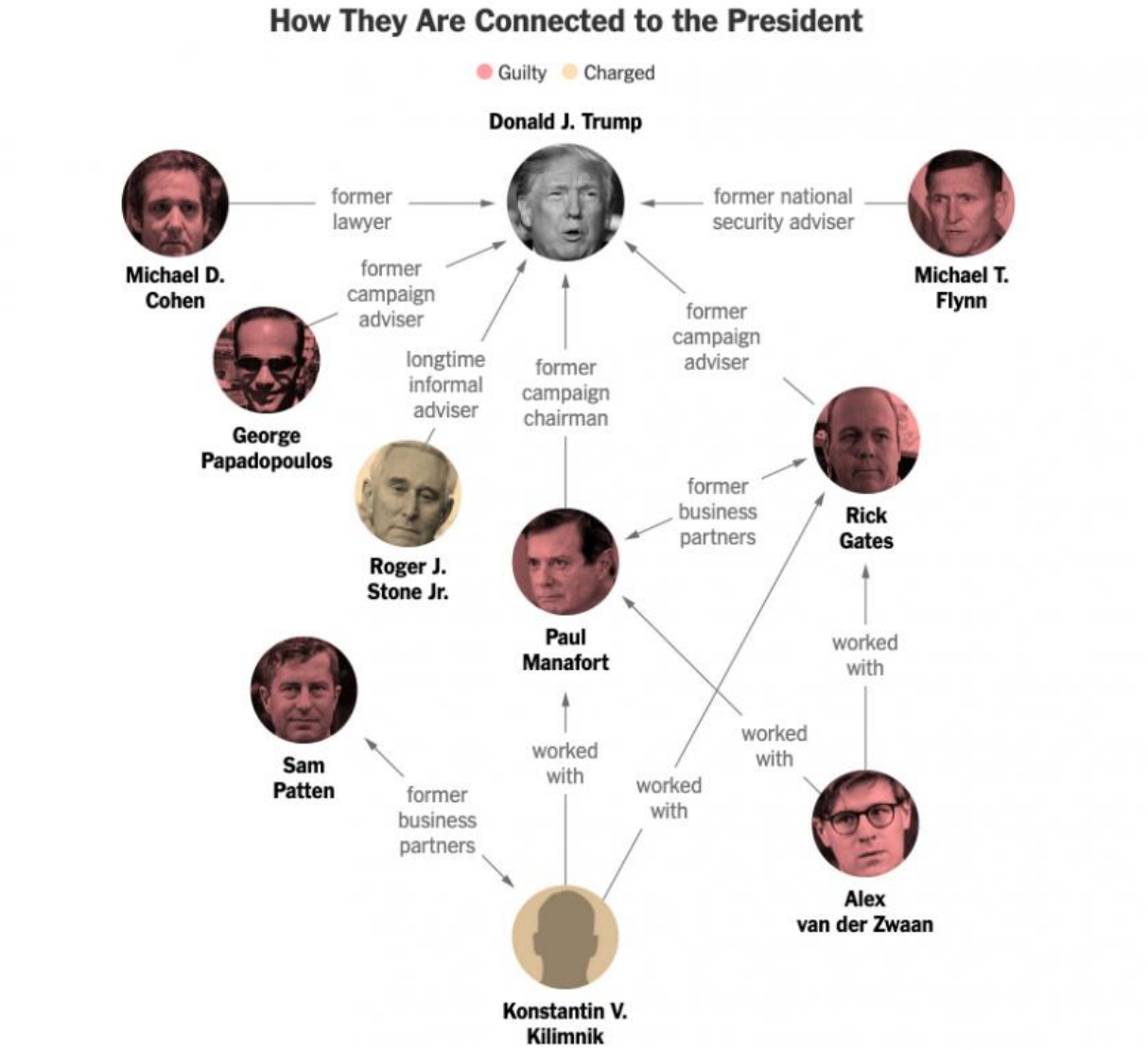
- La première technique consiste à repérer des informations factuellement fausses en comparant un texte donné avec les informations contenues dans une base de données. Ceci peut fonctionner en théorie (un jeu de donné appelé FEVER, pour *fact extraction and verification*, a même été développé pour cela), mais on dispose rarement de bases de connaissances adaptées au problème. En gros, l'actualité ne se réduit pas à une base de données et les fake news ne portent pas tellement sur des informations factuelles isolées.

- Une seconde technique est de repérer des documents types, grâce à leur titre, leur mise en page, les illustrations qui les accompagnent, entre autres. Ceci peut aussi fonctionner jusqu'à un certain point, mais ce n'est pas très précis. Par exemple, de nombreux titres racoleurs utilisent une mise en page tape-à-l'œil sans pour autant être des fakes news.

- L'IA en faveur de l'investigation ?
  Il est important de sortir d'un prisme déformant qui présente l'IA comme source de troubles et qui vole les compétences des journalistes. Souvenons-nous du film Spotlight (Tom McCarthy, 2015). Une poignée de journalistes enquête alors sur la culpabilité de l'Église catholique dans des scandales de pédophilie. Le journaliste Matt Carroll réussit, précisément grâce à la data, à croiser les différentes affaires. Mais depuis 2002, les outils technologiques se sont transformés : « L'augmentation toujours grandissante de la puissance de calcul et de la performance des algorithmes va permettre de collecter et de traiter encore plus rapidement des données de plus grande taille », synthétise le rapport.

  Et nous en sommes déjà là, avec l'émergence du data journalisme que le rapport considère comme étant « à l'origine d'un nouveau journalisme d'investigation ». Une approche où l'homme et la machine collaborent au lieu de s'opposer et qui permet au

journaliste de se concentrer « sur des interviews et des échanges privilégiés avec des individus mieux identifiés ». Ce qui pourrait d'ailleurs être une réponse intelligente à la défiance envers les médias, qu'on accuse souvent de tomber dans le piège de l'immédiateté de l'information et de la course aux clics.



Faiblesses

La première limite identifiée est structurelle : la solution mise en avant est précisément la cause du problème. Le *deepfake* est un parfait exemple de la menace informationnelle permise par l'intelligence artificielle. Cette technique permet la création et la modification de contenu audio-visuel d'une qualité telle qu'il est très complexe d'identifier la manipulation. Les images et vidéos, auparavant mise en avant comme preuves irréfutables ne pourront plus être

considérées comme telles. Par ailleurs, les cybers criminels ont d'ores et déjà intégré l'intelligence artificielle à l'attirail d'outils qu'ils utilisent. Déjà en 2017, 62% des conférenciers, experts de la cybersécurité, du *Black Hat* de Las Vegas estimait déjà que les cybers criminels utiliseraient l'intelligence artificielle pour passer à l'offensive.

On dit souvent de l'IA (et des nouvelles technologies en général) qu'elle est une épée à double tranchant. D'une part, elle permet l'émergence de menaces informationnelles en ligne de plus en plus sophistiquées et l'abaissement des remparts contre les acteurs malveillants. L'IA serait notamment utilisée pour cadrer plus précisément les paramètres d'une attaque informationnelle : quoi, qui et quand attaquer. D'autre part, l'IA présente de nombreuses opportunités pour la lutte contre ces mêmes menaces informationnelles.

- La seconde limite est la capacité de l'IA à se tromper. L'intelligence artificielle est aujourd'hui très au point sur la détection sémantique. Seulement certains aspects de langage lui sont encore complexes à appréhender. Il s'agit par exemple du sarcasme, de la persuasion ou de l'ambivalence. C'est une des raisons qui explique l'existence de *faux positifs,* limites majeures de l'IA. Le fait de désigner un post comme étant de la désinformation, un abus ou un troll alors qu'il ne le serait pas, minimise la propagation des outils de détection et leur efficacité.

Aussi, pour une analyse efficace, l'IA a besoin de jeux de données mais n'en dispose pas toujours pour chaque problème qui se pose (problèmes de Copyright notamment).

- Il faudrait ajouter que les plates-formes ont alors beau jeu d'en appeler à la régulation de la part des États. Mark Zuckerberg a ainsi dit : « nous ne souhaitons pas que les entreprises privées prennent autant de décisions importantes qui touchent aux valeurs fondamentales sans contrôle démocratique » (« We don't want private companies making so many decision – balancing social equities without democratic processes »), tout en protestant quand une nouvelle législation se met en place.

- La dernière limite réside dans l'action humaine en amont de l'utilisation de la machine. Les outils utilisant le Machine-Learning sont paramétrés par des Hommes, qui ont eux-mêmes leurs propres biais cognitifs. Se pose ainsi la question de l'objectivité du paramétrage de la machine pour déterminer ce qu'est une fausse information et ce qui n'en est pas. Certaines plateformes utilisent même l'IA pour identifier des contenus légaux mais définis par ces mêmes plateformes comme étant nuisibles :

« comportements inauthentiques » et « post insensibles » sont par exemple visés. Au-delà de la question de l'objectivité, comment admettre qu'une plateforme définisse la limite entre le sensible et l'insensible, l'authentique et l'inauthentique.

2. Menaces et Opportunités

➢ Menaces

Les menaces déterminent les facteurs externes pourraient avoir un impact négatif sur le lancement du projet. On a ainsi listé quelques éléments qui s'avèrent être néfastes au bon développement de l'intelligence artificielle en Afrique, plus particulièrement dans les pays d'Afrique de l'ouest :

- Les Etats Africains n'investissent pas considérablement dans la recherche en IA ;
- Il y a actuellement très peu de formations en IA dans le continent, ce qui ne favorise pas une bonne appropriation du domaine et sa mise en pratique avec des solutions concrètes ;
- Une dimension importante à considérer est le manque d'infrastructures technologiques en Afrique ;
- Un défi majeur reste l'encadrement réglementaire et légal de l'IA ;
- La plupart des experts africains du domaine se trouvent hors du continent ;

➢ Opportunités

On note toutefois certains facteurs externes positifs non négligeables capables de marquer un grand tournant dans l'évolution de l'intelligence artificielle au Sénégal. Les éléments suivants seront favorables au succès de notre projet :

- Plusieurs pays à travers l'Afrique ont pris conscience de l'importance de l'intelligence artificielle et de la nécessité absolue de la maîtriser, voire d'y être un leader au niveau mondial. Le Sénégal s'est également engagé dans cette réflexion ;
- Plusieurs pays se sont également engagés dans la définition d'une stratégie nationale de l'IA ;
- Les gouvernements africains manifestent sur le sujet un intérêt croissant, d'où la multiplication des initiatives autour des « stratégies nationales de l'IA » destinées à contribuer à l'émergence et à la bonne gouvernance de l'IA à travers le continent ;

- Au Sénégal, l'Académie nationale des Sciences et Techniques a initié une étude sur l'Intelligence artificielle, en mars 2019, dont le rapport devait être soumis au Président de la République lors de la prochaine séance académique solennelle ;

- Le Sénégal pourrait éventuellement se positionner aussi dans la course à la maîtrise de l'IA ;
- Malgré le manque général d'infrastructures technologiques en Afrique, le Sénégal s'est investi dans l'acquisition d'un Data-center (centre de données pour stocker et sécuriser les données des usagers du secteur publiques et privées) et du plus grand supercalculateur de l'Afrique subsaharienne.

IV. PLANIFICATION :

Maintenant que nous avons étudier, de fond en comble, notre projet (définition des termes, faisabilité, objectif), il nous reste une question fondamentale à répondre avant de conclure notre présentation : Quelles sont les différentes étapes qui vont nous permettre de valider le projet ?
Pour cela on s'est basé sur les études effectuées par Ekimetrics, un grand leader en Data Science qui a pour objectif d'ouvrir un nouveau champ dans l'efficacité de la stratégie grâce à des méthodes statistiques innovantes. On peut se baser sur cette citation dans son article « C'est une nécessité pour les entreprises de disposer de données de qualité pour la mise en place de parcours client permettant une expérience sans couture et personnalisée » et sur celle-ci « atteindre une proximité avec le client peut s'avérer plus complexe que prévu. En effet, les seuls investissements technologiques ne suffisent pas, et ces derniers doivent également être accompagnés de changements structurels et organisationnels. » pour dire qu'il ne suffit pas seulement de disposer d'un investissement en grande quantité pour réussir un projet de data science mais il faut également avoir les outils et compétences nécessaires pour mettre en relation les structures de l'entreprise avec le projet qui sera accompli. Ainsi nous avons résumé les 6 parties de l'article réalisé par Ekimetrics en 3 parties que voici :

- La définition des challenges initiaux
- Le cadrage stratégique
- La création d'algorithmes modulaires

Il faut préciser que la partie technique du projet sera assurée par une équipe de Data Scientist. Ces derniers ont les compétences nécessaires à la collecte des données, à son traitement et aux

entrainements et test qui seront réalisés afin de créer un modèle de machine learning capable de prédire les fakes news.

1. Définition des challenges initiaux :

La première étape consiste à un élément clé qui est de s'assurer d'avoir les outils nécessaires à la mise en place du projet. Elle peut être considérée comme l'étape la plus longue et la plus importante du projet. Elle demande une grande précision en matière d'analyse car chaque variable à un impact sur la mise en place du reste du projet et ainsi, elle ne doit pas être négligée :

a. La première chose à faire est d'identifier les plateformes et outils de Système d'information déjà en places et ceux qui seront implantés plus tard. Ces outils sont essentiels pour la coordination des différents membres de l'équipe et la fluidité et le rythme du projet de Data science. Nous en noterons trois en particuliers :

- ERP (Entreprise Resource Planning) traduit par Progiciel de Gestion Intégré : C'est un progiciel[3] qui permet de gérer l'ensemble des processus d'une entreprise en intégrant l'ensemble de ses fonctions, dont la gestion des ressources humaines, la gestion comptable et financière, l'aide à la décision, mais aussi la vente, la distribution, l'approvisionnement et le commerce électronique.
- CRM (Customer Relashionship Management) traduit par Gestion de la Relation Client : Le CRM regroupe l'ensemble des dispositifs ou opérations de marketing ou de support ayant pour but d'optimiser la qualité de la relation client, de fidéliser et de maximiser le chiffre d'affaires ou la marge par client.
- DMP (Data Management Platform) ou Plateforme de Gestion de Données : Il s'agit d'une plateforme proposée généralement en mode SaaS[4] et permettant de récupérer, centraliser, gérer et utiliser les données relatives aux prospects et clients dans l'environnement digital.

b. La deuxième étape consiste d'abord à accéder aux données qui seront utilisées pour entrainer notre modèle et ensuite à les traiter. Dans ce projet, la collecte de données se fera à l'aide du web scrapping.

- Définition : le web scraping est définit comme le processus d'utilisation d'outils technologiques à but d'extraire automatiquement et d'organiser les données du web,

[3] Un terme commercial qui désigne un logiciel applicatif généraliste aux multiples fonctions, composé d'un ensemble de programmes paramétrables et destiné à être utilisé par une large clientèle.

[4] un mode de mise à disposition de solutions logicielles, notamment marketing, qui permet au client d'y accéder à distance et d'être en grande partie "débarrassé" des contraintes d'installation / intégration et de maintenance.

dans le but de les analyser ultérieurement. Le principe est simple : Un script ou un programme accède à un page web via une URL donnée en entrée (web crawler), en utilisant le protocole HTTP et/ou via un navigateur web comme le ferait un être humain. Une fois qu'il y a accès, le web scraper peut copier le site web et le stocker dans une base de données locale dans le but d'une récupération et d'une analyse ultérieure de celui-ci. C'est un processus qui comprend donc 3 phases principales : l'accès au site web, l'extraction du site web et enfin la décomposition du code de ce dernier. C'est un procédé couramment utilisé dans les domaines comme le web mining, le data mining, la comparaison de prix, la veille concurrentielle ou encore la détection de changement de sites webs. C'est aussi le principe qu'utilise un navigateur pour se connecter à un site web et afficher les données qu'il contient.

- Avant de faire du web scraping il faut disposer d'une base de données de toutes les url de sites de médias comportant de douteuses informations. On aura essentiellement besoin d'analyser au préalable les données du Decodex :

  Au début de l'année 2017, A. Sénécat sortait un article sur LeMonde.fr pour annoncer la sortie du Décodex, un outil visant à repérer les sites réputés peu fiables. Cet outil fait suite à un blog créé par le journal intitulé « Les Décodeurs » et qui visait alors à vérifier les différentes rumeurs propagées sur le web. L'offre du Décodex se décline en 3 outils : Un moteur de recherche, une extension de navigateur permettant de notifier la fiabilité d'un site et enfin un chatbot Facebook pouvant indiquer la fiabilité d'un site web. On pourrait en citer un quatrième qui n'est pas explicitement nommé, mais qui est pourtant à la base de ces outils et du développement de ce mémoire qui est la base de données librement accessible ayant permis de construire ces outils. Pour les développer, l'équipe des Décodeurs a référencé plus de 600 sites d'information qu'ils ont analysé et classé à la main. Cette base fut développée par la suite et un an plus tard, le Décodex recensait « près de 200 fausses informations et plus de 5400 liens qui les diffusent ».

- Quel est le cadre éthique du web scraping :

  Le cadre éthique du web scraping est très récent, et a été développé par Krotov et Silva (2018). Ils y définissent les conséquences du web scraping qui pourraient s'avérer dangereuses pour toute créature sensible et développent 3 champs sur lesquels il faut être particulièrement attentif dans cette pratique : Le respect de la vie privée, le respect de la confidentialité des organisations et du secret des affaires, diminuer la valeur produite par une orgarnisation.

A partir de là, plusieurs pratiques peuvent être mises en place pour éviter de tomber dans l'infraction à ces principes. Tout d'abord, la parcimonie[5] dans la collecte de données est de rigueur afin de ne pas surcharger le serveur hébergeant un site web et pour ne pas recueillir trop d'informations pouvant amener directement ou indirectement à la violation de la vie privée d'un individu ou d'une organisation. Ensuite, la non appropriation des données récoltées afin de ne pas violer le droit d'auteur du créateur d'un site web. Enfin, la récolte des données doit suivre un but de création de valeur ajoutée servant au bien commun et ne visant pas l'enrichissement financier personnel en premier lieu.

- En dernier lieu, nous devons réaliser l'acquisition de données des sites dont les url sont précisés dans de le dataframe issu du nettoyage du Decodex, à l'aide du web scrapping, et faire un nettoyage préalable de ces dernières avant de passer à l'étape suivante (entrainement des modèles de machine learning[6]) :

  Nous devons tout d'abord importer les données précédemment obtenues en nettoyant la base de données du Décodex pour avoir l'identifiant de la fake news, l'URL et la catégorisation « VRAI » « FAUX ». Trois colonnes doivent ensuite être créées, dans lesquelles seront insérées les données récoltées. L'algorithme devra suivre la logique suivante :

  - Importer les données du Décodex pré traitées dans une dataframe.
  - Déclarer le moteur utilisé pour le web scraping.
  - Déclarer une variable égale à l'index de la dataframe le plus petit ne contenant pas de données web scrapées. Cette variable nous permet de reprendre le traitement là où il s'était arrêté s'il tombe en erreur.
  - Déclarer une variable « erreurs » qui nous permettra de compter le nombre d'erreurs qui sont apparues pendant l'acquisition de données.
  - Pour chaque URL dans la colonne de la dataframe les contenant, ouvrir le navigateur, aller à l'URL correspondante et web scraper les données puis quitter le navigateur.
  - Trouver le titre, représenté par la balise HTML <h1/> et le sous-titre de l'article par la balise <h2/>, puis concaténer le titre et le sous-titre dans une seule chaîne de caractères.
  - Faire la même chose pour le texte contenu dans l'article (balise HTML <p/>).
  - Trouver tous les URLs (balise <a/>) dont le contenu commence par http.

[5] Principe consistant à n'utiliser que le minimum de causes élémentaires pour expliquer un phénomène.

[6] **Le Machine Learning est une technologie d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet.**

- Insérer ces données dans les colonnes correspondantes de la dataframe pour le numéro de ligne égal à row_number.
- Afficher le nombre d'erreurs à la fin de la boucle.
- Ecrire ces données dans un nouveau fichier JSON

Suite au webscraping, des erreurs peuvent apparaitre notamment liées au fait que des sites web ne soient désormais plus disponibles, ou parce que les serveurs en ont refusé l'accès au web scraper. Toutes les observations contenant une liste de mots couramment rencontrés face à des erreurs de serveur (erreur 404, erreur 503) doivent être supprimés. Ensuite, les articles en langues autres que le français vont être filtrés. Afin d'éviter toute perte non nécessaire d'information, ce filtrage doit cependant être réalisé à la main.

2. Le cadrage stratégique :

Selon les analyses réalisées par Ekimetrics « La deuxième étape d'un projet de science des données réussi réside dans un cadrage stratégique impliquant un ou plusieurs membres du comité exécutif[7] ». Cela permet, en effet, de maintenir une perspective adéquate en tenant compte très tôt des contraintes et des ambitions de l'entreprise. Par exemple, l'optimisation de la marge dans l'ensemble de l'entreprise peut entrer en contradiction avec la nécessité d'investir dans les données et la technologie. **Il s'agit également de la phase où se posent les questions d'attribution des ressources, toujours avec le ROI en tête.**

Cette phase initiale permet, en fin de compte, de mesurer tous les risques, enjeux et bénéfices tout en définissant dans quelle mesure ce nouveau projet peut avoir un impact positif à tous les niveaux de l'entreprise. Plus ces variables sont connues et prises en compte, plus le projet aura de chance d'aboutir.

3. La création d'algorithmes modulaires :

Plusieurs méthodes et challenges apparaissent pour classifier les données issues du web scraping. Dans l'ensemble, le workflow suivant sera suivi :

- Nettoyer les données capturées
- Séparer les données en jeux d'apprentissage et de test

[7] Un groupe généralement restreint de personnes, formant un ensemble constitué, investi d'un pouvoir de surveillance et de décision, assurant la direction d'une entreprise sous les ordres du directeur général de celle-ci

- Gérer le déséquilibre entre les classes
- Choisir l'algorithme le mieux adapté
- Comparer les performances des différents modèles obtenus

Toutefois, plusieurs difficultés apparaissent pour ces différentes parties. Le plus important est la gestion des différentes colonnes qui ont été sélectionnées pour développer cet algorithme (le titre, le corps de texte et la ou les URLs présentes dans le texte). En effet, chaque algorithme prend en entrée une matrice représentant les mots dans le texte. Il y a ici 3 niveaux de textes que l'on peut gérer de deux façons différentes. Tout d'abord, on peut concaténer les 3 niveaux en un seul que l'on prendra comme matrice d'entrée, ce qui revient à mettre dans une même chaîne de caractère le titre, le texte et l'URL. Le problème principal lié à cette solution est qu'il sera plus compliqué d'assigner un poids différent à chaque type de texte (rendre le poids du titre plus important par exemple). L'autre solution est d'utiliser un algorithme d'apprentissage ensembliste (bagging / boosting) ou des méthodes hybrides. Ces algorithmes cherchent dans un espace d'hypothèse celle qui permet de réaliser la meilleure prédiction en prenant en entrée plusieurs algorithmes. La mise en place de ces derniers nécessite cependant des compétences techniques très avancées que nous ne pouvons mettre en place, et la première méthode sera donc préférée.

CONCLUSION

Tout type de projet comporte des erreurs. En particulier un projet Data Science. Les algorithmes qui seront utilisés devront forcément comporter des erreurs et le résultat final ne sera pas forcément celui qu'on espère obtenir. Toutefois, tout dépend de l'envergure des efforts déployés lors de son déploiement et de la fluidité et du rythme de l'équipe de projet en collaboration avec l'ensemble des acteurs de l'entreprise. Notre projet prendra tout son sens que lorsqu'on fera prendre conscience aux internautes du danger que constitue les Fakes News. On espère que, malgré les grandes quantités de fausses nouvelles circulant sur la toile, on réussira à limiter à temps la propagation de ce 'virus' numérique au sein de notre pays.