

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/344137242>

# Développement d'un algorithme de détection des fake news en français – Une approche transverse utilisant la base de données du Décodex

Thesis · July 2020

DOI: 10.13140/RG.2.2.11038.36167

---

CITATIONS

0

READS

1,207

1 author:

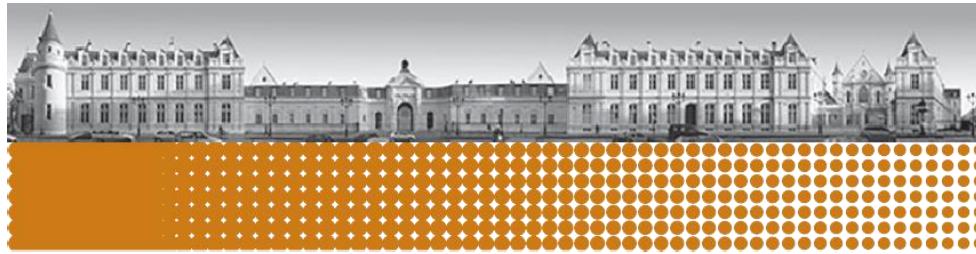


[Emmanuel Daveau](#)

Service hydrographique et océanographique de la marine

5 PUBLICATIONS 13 CITATIONS

[SEE PROFILE](#)



Mémoire pour l'obtention du  
**Master Sciences humaines et sociales mention humanités numériques**  
**Parcours Mégadonnées et analyse sociale (MEDAS)**

**Développement d'un algorithme de détection  
des fake-news en français : *Une approche transverse  
utilisant la base de données du Décodex***

Emmanuel DAVEAU

**Date et lieu de la soutenance**

- 10 Juillet 2020
- Nantes

**Membres du jury**

- Tuteur entreprise : Thomas HAUGOU
- Tuteur pédagogique : Olivier PASQUIER
- Président du jury : Béatrice ARRUABARRENA

**Promotion (2018-2020)**



Paternité Pas d'Utilisation Commerciale - Pas de Modification

# Sommaire

<b>1 – Cadre théorique .....</b>	<b>6</b>
<b>1.1 – Contexte Général .....</b>	<b>6</b>
<i>1.1.1 - L’ère de la post-vérité .....</i>	<i>6</i>
<i>1.1.2 - Etablir la vérité .....</i>	<i>8</i>
<i>1.1.3 - Les théories de la vérité, la vérité scientifique.....</i>	<i>9</i>
<i>1.1.4 - Définir une fake-news .....</i>	<i>12</i>
<i>1.1.5 - La vérité à l’ère de la post-vérité .....</i>	<i>15</i>
<i>1.1.6 - L’Amplification Sociale du Risque : Un background théorique de la propagation d’une information .....</i>	<i>18</i>
<b>1.2 – Les mécanismes psychosociaux au cœur de la recherche d’informations .....</b>	<b>20</b>
<i>1.2.1 - Je reconnaiss donc je sais : Le biais d’exposition.....</i>	<i>20</i>
<i>1.2.2 - Les tenants psychologiques de la croyance en une information.....</i>	<i>21</i>
<i>1.2.3 - La recherche d’information à l’ère d’internet .....</i>	<i>25</i>
<i>1.2.4 - La robustesse et le danger des fake-news .....</i>	<i>26</i>
<i>1.2.5 - Pourquoi préfère-t-on les histoires fausses ?.....</i>	<i>28</i>
<i>1.2.7 - Réduire l’impact des fake-news .....</i>	<i>29</i>
<b>1.3 – L’état de l’art des algorithmes de détection des fake news .....</b>	<b>30</b>
<i>1.3.1 - Aux origines : La détection de spam .....</i>	<i>31</i>
<i>1.3.2 - Déetecter les fake-news à l’ère des réseaux sociaux .....</i>	<i>32</i>
<i>1.3.3 - Les stratégies actuelles pour combattre les fake-news .....</i>	<i>33</i>
<i>1.3.4 - L’extraction de features .....</i>	<i>34</i>
<b>2 - Présentation et acquisition des données .....</b>	<b>36</b>
<b>2.1 – La base de données du Décodex.....</b>	<b>36</b>
<i>2.1.1 - « Un premier pas vers la vérification de masse de l’information ».....</i>	<i>36</i>

2.2.2 - Méthodologie : Téléchargement et mise en forme des données.....	37
<b>2.2 – Le web scraping et l’acquisition de données sur le web.....</b>	<b>40</b>
2.2.1 - Le cadre légal du web scraping .....	40
2.2.2 - Le cadre éthique du web scraping .....	43
2.2.3 - Acquisition des données : Développement du Web Scraper .....	44
2.2.4 - Nettoyage des données issues du web scraper.....	44
<b>3 – L’algorithme de classification de fake-news .....</b>	<b>46</b>
<b>3.1 – Nettoyer les données acquises .....</b>	<b>46</b>
<b>3.2 – L’algorithme de détection des fake-news.....</b>	<b>48</b>
3.2.1 - Le problème du déséquilibre de classes.....	48
3.2.2 - Préparation du pipeline de l’algorithme.....	51
3.2.3 - Exécution du pipeline de l’algorithme .....	52
<b>4 – Conclusions et perspectives.....</b>	<b>56</b>
<b>4.1 – Classifier les fake-news .....</b>	<b>56</b>
<b>4.2 – Limites méthodologiques.....</b>	<b>59</b>
<b>4.3 – Perspectives futures .....</b>	<b>61</b>
<b>4.4 – Conclusion .....</b>	<b>61</b>
<b>Webographie .....</b>	<b>63</b>
<b>Bibliographie .....</b>	<b>64</b>
<b>Annexes .....</b>	<b>70</b>
<b>Annexe 1 – Transformation des données du Décodex en dataframe .....</b>	<b>70</b>
<b>Annexe 2 – Web Scraping .....</b>	<b>79</b>
<b>Annexe 3 – Traitement des données scrapées .....</b>	<b>82</b>
<b>Annexe 4 – PreProcessing.....</b>	<b>88</b>
<b>Annexe 5 – Application de l’algorithme de Machine Learning .....</b>	<b>96</b>

Je tenais à remercier l'ensemble de l'équipe pédagogique du CNAM pour la qualité de ses enseignements, ainsi que pour la volonté de chacun de nous transmettre des connaissances qui auront été très précieuses dans la rédaction de ce mémoire. Merci aussi à Xavier Aimé pour l'écoute, la patience et la gentillesse dont il aura su faire preuve au cours de ces deux années de Master. Les qualités et valeurs qu'il aura portées auront certainement joué pour beaucoup dans la réussite de ses étudiants. Merci aussi à mes collègues qui auront toujours été présents pour aider, reformuler et préciser à ceux qui avaient des difficultés à comprendre certaines parties du cours. Enfin, je me dois de remercier mon épouse, qui aura été d'un support et d'un soutien sans faille tout du long de ces deux années de Master, et sans qui je n'en serais certainement pas là à l'heure actuelle.

*« Eh bien, il disait souvent que la réussite sourit à ceux qui n'ont pas honte de leurs propres mensonges. Et ça aussi : ‘Les choses n'ont pas besoin d'être vraies, du moment qu'elles en ont l'air’.* » Seconde Fondation (Asimov, 1953)

**DAVEAU Emmanuel.** Développement d'un algorithme de détection des fake-news en français : Une approche transverse utilisant la base de données du Décodex. Mémoire professionnel MEDAS, Titre I, Data Scientist. Conservatoire national des arts et métiers des Pays de la Loire – Promotion 1.

Les fake-news sont désormais un enjeu bien connu, et plusieurs initiatives sont actuellement en cours de développement afin d'y répondre. Il apparaît cependant que la quasi intégralité des données et réponses disponibles sont en anglais. Des journalistes de Le Monde ont cependant réalisé un travail d'investigation afin de générer une base de données pointant vers des articles propageant de fausses informations en français, le Décodex. Notre objectif était d'utiliser cette base de données afin de collecter des textes déjà classifiés comme propageant de fausses informations. L'un des principes fondateurs de ce travail est l'approche pluridisciplinaire adoptée. Nous avons traité ces textes sur plusieurs niveaux (titres, corps de texte et sources) afin de créer un modèle de machine learning de classification et de reconnaissance des fake-news. Plusieurs problèmes méthodologiques ont rendu l'interprétabilité de notre modèle nulle en dépit de bons résultats. Nous discutons de ces limites et proposons un ensemble de pistes ouvertes pour des recherches futures par ce travail.

#### Mots-clés

Fake-news  
Vérité  
Véritudes  
Information  
Décodex  
Machine-learning  
Traitement du langage  
Approche transverse  
Classification

Fake-news are now a regular issue discussed in the media, and several initiatives are currently being developed in order to tackle this problem. However, most of these are based on English data. Resources are however available, as journalists from the journal Le Monde created a labeled database called Décodex, where they classified several articles written in French as “true” or “false”. Our main objective was to use this database to collect data from these sources, as the labelling had already been done. One of this work’s core principles was to adopt a transversal approach. After collecting them, we processed these data on different levels (headers, body and sources) in order to create a classification machine learning algorithm. Many methodological issues made the interpretation of our model complex and unusable despite good results. We discuss these issues and how this preliminary work might help in developing future research.

#### Keywords

Fake-news  
Truth  
Truthiness  
Information  
Décodex  
Machine-learning  
Natural Language Processing  
Pluridisciplinarity  
Classification

# 1 – Cadre théorique

## 1.1 – Contexte Général

### 1.1.1 - L'ère de la post-vérité

Dans une interview pour le Gentlemen's Quarterly réalisée en 1995, Terry Pratchett avait été une des premières personnes à questionner Bill Gates sur le potentiel de désinformation auquel pouvait conduire internet. Pratchett citait l'exemple d'une théorie de la conspiration qui pourrait être rédigée par un faux institut, et qui serait aussi accessible, et à titre égal, à toute autre information historique académique, documentée et vérifiée. Bill Gates lui avait alors répondu que des autorités réguleraient les informations, que des systèmes de référencement et de suggestion pointeraient les internautes dans la bonne direction. Un quart de siècle plus tard, la prédiction réalisée par l'auteur de littérature fantastique n'a jamais été plus tangible et concrète qu'aujourd'hui.

En Novembre 2016, le groupe Oxford Dictionaries a annoncé que le mot de l'année serait ‘post-vérité’<sup>1</sup>. Ce terme fut créé dans l'ouvrage ‘*The Post-Truth Era : Dishonesty and Deception in Contemporary Life*’ (Keyes, 2004) et renvoie à un système éthique dans lequel la présentation de faits objectifs sont moins influents aux yeux de l'opinion publique que l'appel aux croyances et aux émotions. Bien qu'apparu et débattu il y a plus d'une décennie, le concept de post-vérité est récemment devenu accepté et largement utilisé par les médias, suite à la coïncidence de deux événements majeurs entre 2016 et 2017 : L'élection du candidat à la présidentielle américaine Donald J. Trump et le vote des anglais en faveur du Brexit.

Au cours de la campagne présidentielle, le site internet Politifact – un site internet spécialisé dans la vérification de faits – a évalué que 76% des affirmations réalisées par Trump étaient fausses, quasiment fausses voire tout simplement ridicules<sup>2</sup>. Son opposante Hillary Clinton avait quant à elle un score de 26% sur les mêmes métriques<sup>3</sup>. Un des exemples fréquemment cités est l'affirmation de Trump selon laquelle 81% des américains blancs sont tués par des noirs, lorsque les statistiques ethniques du FBI montrent que 84% des blancs sont tués par d'autres personnes blanches<sup>4</sup>. En dépit des preuves avancées montrant la fausseté de plusieurs affirmations impactantes, Newt Gingrich, président de la Chambre des représentants des Etats-Unis, défendit ce genre d'affirmation en disant que les américains ne se sentaient plus en sécurité et que ce seul sentiment valait pour preuve.

Ce sentiment est ici un concept crucial de l'époque actuelle, auquel les anglo-saxons ont donné le nom de « *truthiness* », que les médias francophones ont traduits par le néologisme

*véritudes* : une assertion inexacte mais qui est ressentie vraie et réelle aux yeux du public, bien qu'elle ne soit pas appuyée par des faits. Un des points les plus intéressants liés à ce concept de véritude est qu'il apparaisse lorsqu'internet est aussi développé à l'heure actuelle, permettant théoriquement à n'importe qui d'accéder instantanément depuis son smartphone ou ordinateur à toute l'information disponible dans le monde entier. Une dernière illustration permettant d'appuyer l'importance des véritudes à l'heure actuelle dans le débat public nous vient des Royaume-Unis, où interrogé pendant sa campagne pour le Brexit, le président de campagne en faveur du Brexit avait proclamé que « *l'exactitude [des informations ; ndt] est faite pour les vendeuses de poudre de perlimpinpin* »<sup>5</sup>.

Ainsi, chaque année, une véritude fait son apparition dans le débat public, et chacune instille différentes notions plus ou moins problématiques : Terre plate, vaccinations toxiques, majorité des viols à Paris commis par des étrangers, glyphosates responsables de cancers... L'un des principaux problèmes de ces véritudes est leur persistance dans le temps. La défiance envers les vaccins avait par exemple commencé à la fin des années 1990.

L'accès facilité aux fausses informations et la distinction entre un fait et une véritude sont désormais devenus des problèmes majeurs à l'ère de l'information. C'est pourquoi les questions auxquelles nous tenterons de répondre dans ce mémoire seront les suivantes : Peut-on extraire des informations permettant de caractériser l'article d'une page web comme présentant un risque de désinformation ? Y a-t-il des marqueurs sémantiques ou syntaxiques présents dans le contenu d'un article ou dans la façon dont l'article est agencé sur une page web (eg. Présence de liens URL vers des sites fiables) permettant de distinguer les sites fiables des autres ? Enfin peut-on développer une preuve de concept d'un algorithme permettant de distinguer automatiquement une page comme présentant une information vérifiée d'une page véhiculant des *fake-news* ?

Pour répondre à ces questions, nous veillerons tout d'abord à présenter en détails les différents enjeux qui entourent la question des fake-news. C'est-à-dire définir ce qui distingue une information vraie d'une fausse, les déterminants sociétaux et psychologiques qui entourent ce phénomène, avec les risques associés. Pour finir, nous ferons un état de l'art sur la recherche en détection automatique de fake news. Nous présenterons ensuite la base de données du Décodex et discuterons après l'avoir présentée de la méthode de récupération des données. Enfin, nous développerons l'algorithme avant de présenter les forces et faiblesses d'une approche liée à la détection automatique des fake-news.

### **1.1.2 - Etablir la vérité**

L'un des exercices les plus complexes lorsque l'on tente de trier une information « *fausse* » d'une information dite « *vraie* » est de savoir où positionner le curseur de la vérité. En effet, de la même manière que l'art ou la justice, la vérité est un des concepts les plus facilement contestés de la philosophie (Gallie, 1956). Cela signifie qu'avant d'essayer de distinguer quelles affirmations peuvent-être vraies ou fausses, les philosophes continuent encore de questionner la signification même de ce qui distingue une affirmation vraie d'une affirmation fausse. Ainsi, ce n'est que récemment dans le développement de la philosophie et de l'épistémologie que l'on a commencé à distinguer les informations appuyées sur des faits, des éiphanies et des témoignages faisant autorité.

Bien que ce ne soit pas le but de cet essai de réaliser une analyse philosophique sur la notion de vérité, il reste important de distinguer les différentes notions qui seront développées ici afin de laisser aussi peu de place à l'ambiguïté des termes que possible. De plus, cette analyse nous permettra de réduire notre champ d'intervention afin de faire la distinction entre ce qui relèverait de notions comme la propagande politique ou économique, des rumeurs ou encore de la notion plus contemporaine de « *buzz* ». Cette analyse est d'autant plus essentielle à l'époque actuelle, ou les médiums de diffusion de l'information peuvent rendre les barrières entre ces différents termes plus floues. Une information établie sur un journal d'actualité pourra en effet trouver une critique pertinente sur un réseau social, une réinterprétation fallacieuse sur un blog ou une critique virulente par un politique qui sera reprise par la suite dans la presse. Pourtant, toutes ces informations ne peuvent pas être appréhendées sous le même prisme, ce pourquoi plusieurs angles d'approche sont nécessaires pour étudier la question des fausses informations. On pourra citer pour exemple l'émergence très récente de l'étude du comportement et de la propagation des informations sur les réseaux sociaux.

Cette approche philosophique est aussi bénéfique dans l'étude de la difficulté historique à trouver une définition unique à ce que l'on peut considérer comme « *vrai* ». Dans la Grèce antique déjà, plusieurs formes de vrai existaient : de la notion « *Alethes* » signifiant « *non caché* » aux concepts de « *atrekes* » pour « *non-déformé* », « *nemertes* » pour « *sans faute* », « *adolos* » pour « *non mensonger* », « *ortos* » pour « *sans faux semblants* », « *apseudos* » pour « *sincère* » et « *etymos* » pour « *authentique* » (Woleński, 2019, p.5). L'étymologie même du concept de vrai montre la complexité de cette notion et la précaution avec laquelle chaque terme doit être employé lorsqu'on essaie de distinguer une information que l'on pourrait qualifier de fausse ou de vraie. L'une des étymologies les plus parlantes toutefois, est celle de la notion

anglaise de « *truth* », originaire de « *troth* », mot désignant la fidélité à la source. Ce terme contraste en effet avec les questions de fidélité à la cible (questionner l’observation en elle-même, et non son origine), et rappelle la notion de déformation du message original. Cela signifie qu’originellement, celui qui est *vrai* est celui qui est fidèle à la source, qui respecte le message d’une autorité qui lui est supérieure. Que cette dernière soit citoyenne, militaire ou divine.

Au cours du Moyen-Âge, la notion de vérité a commencé à évoluer avec le théologiste Thomas d’Aquin, dont la philosophie a empreint l’église catholique au cours des siècles suivants. Pour d’Aquin, la vérité n’était plus en lien avec la source, qui était pour lui indubitablement Dieu, mais avec l’aspect empirique du monde. Ainsi, il n’était plus question pour lui de remettre en question la nature divine des choses du monde. Les choses existaient par Dieu, et l’Eglise représentant sa parole, en était l’autorité principale. Ce qui comptait était alors d’étudier la réalité du monde pour en percevoir le caractère divin. Ce changement est crucial car la vérité sur l’existence du monde n’est alors plus questionnable mais divine, et l’étude des objets physiques ne pouvait alors qu’inévitablement corroborer l’hypothèse divine (Fuller, 2018). En somme, la fidélité vis-à-vis de la cible du message prévaut alors sur la fidélité de la source.

Ce changement d’objet de fidélité, ainsi que la non remise en cause de la divinité des faits du monde marqua ainsi l’histoire des sciences. Par exemple, dans son *Principia Mathematica*, Newton déclara ne pas feindre d’hypothèses sur la raison des propriétés de la gravité (*hypotheses non fingo*). Par-là, il témoignait son souhait de ne pas remettre en cause l’explication théologique au phénomène de la gravitation. En d’autres mots, il voulait prouver être de bonne foi en ne cherchant pas à remettre en question l’explication divine des relations causales observables sur Terre. L’avancée des progrès scientifiques à l’époque de la Renaissance, combinée à plusieurs autres facteurs historiques, ont toutefois permis d’instaurer de nouveaux standards d’évaluation de ce que l’on peut qualifier de vrai.

### **1.1.3 - Les théories de la vérité, la vérité scientifique**

Avec le temps, plusieurs évaluations de la vérité ont été faites. On peut noter la plus simple à définir, c’est-à-dire la théorie « *conformiste* » de la vérité. Dans cette théorie, est vrai ce qui est unanime ou ce qui fait la majorité. Cette vision de la vérité est la plus importante en démocratie, mais peut s’avérer peu fiable lorsqu’il existe un rapport de force entre les individus, auquel cas les influences des plus forts peuvent biaiser les notions de vrai pour leur permettre de répondre à leurs besoins particuliers.

En rupture avec cette définition fut développée la notion de théorie rationaliste de la vérité, illustrée par les « *idées claires et distinctes* » de Descartes dans le premier principe de son *Discours de la Méthode*. Cette vérité correspond à ce que l'on pourrait qualifier d'évidence intellectuelle. Ces évidences concernent toutes les conclusions logiques que l'on peut tirer d'un ensemble de connaissances, sans avoir besoin de démonstrations empiriques, ce qui peut concerner entre autres l'ensemble des démonstrations mathématiques. C'est ce principe que le philosophe allemand Emmanuel Kant critique dans sa *Critique de la Raison Pure*, en soulignant l'importance de notre sensibilité et de notre rapport au réel dans le développement de telles conclusions.

Dans la suite logique est apparue la notion de théorie de la vérité-cohérence. Dans une correspondance, le mathématicien allemand David Hilbert écrit « *Si des axiomes arbitrairement posés ne se contredisent pas l'un l'autre ou bien avec une de ses conséquences, ils sont vrais et les choses ainsi définies existent. Voilà pour moi le critère de la vérité et de l'existence* » (Rivenc, 1992, p. 227). Cette notion étend la théorie rationaliste de Descartes en appliquant les critiques de Kant pour unifier le tout dans une théorie cohérente où seule la non-contradiction permet d'accepter une proposition. On retrouve cette philosophie dans la recherche de la « *Théorie du Tout* » d'Einstein, qui cherchait une théorie permettant d'unifier la loi gravitationnelle de Newton avec la relativité restreinte, et donc par là-même d'obtenir des lois dénuées de toute incohérence dans leurs propositions. Cette conception a permis aux sciences logiques de se développer, mais reste toutefois complexe à intégrer dans les sciences naturelles. Dans un référentiel statique, il est en effet simple de contrôler la cohérence de deux énoncés. Toutefois, dans un système dynamique, et par conséquent, chaotique - celui d'un être humain évoluant dans un environnement en constante mutation par exemple - le contrôle de ces facteurs et la vérification des incohérences sont eux plus complexes.

Enfin, la vérité-correspondance est une dernière forme de vérité héritée de la philosophie aristotélicienne, et employée dans la science contemporaine. Dans sa *Métaphysique*, Aristote dit que « *Dire de ce qui est qu'il n'est pas, et de ce qui n'est pas dire qu'il est, voilà le faux ; dire de ce qui est qu'il est, et de ce qu'il n'est pas dire qu'il n'est pas, voilà le vrai* ». Pour simplifier, cette théorie dit que « *la vérité est la propriété des phrases, des assertions, des croyances, des pensées ou des propositions qui, dans un discours ordinaire, disent être en accord avec les faits ou indiquer ce qui se passe* » (Dufour, 2018). Dans cette idée, la vérité consiste en une relation à la réalité. On peut dire d'un principe ou d'une croyance qu'il est vrai si et seulement si ce principe ou cette croyance est étayée par quelque chose d'extérieur à la croyance même. On n'est alors plus dans une relation cohérente où « Si A provoque B, et B,

lorsque provoqué par A provoque C, alors A provoquera C », mais plutôt du type « A est vrai si A est étayé par un fait » ou « B est faux si aucun fait ne vient étayer B ».

Ces courants furent et continuent de faire partie des principaux courants philosophiques liés à la notion de vérité, bien que d'autres n'aient pas été développés ici. Toutefois, se pose alors la question suivante : Si des différences philosophiques existent autour de la notion de vérité au sein des différentes pratiques scientifiques, que peut-on qualifier de vérité scientifique ? Dans une interview donnée en 2012<sup>6</sup>, le philosophe australien Ellerton explique que si l'on souhaite rester pragmatique, des considérations plus simples de la vérité doivent être énoncées. Il développe ainsi trois notions de vérité auxquelles chaque scientifique doit-être sensibilisé afin d'éviter de mal communiquer diverses idées au public. Il définit ainsi dans un premier temps la vérité subjective, simple énoncé de croyances, faits ou opinions sujettes aux biais et limitations de chacun. Pour citer un exemple simple, la couleur d'un objet sera dépendante de la quantité et de la répartition des cônes au centre de l'iris, qui en dehors de facteurs héréditaires comme le daltonisme, dépend aussi de facteurs biologiques comme le genre (Mancuso et al., 2009).

La seconde forme de vérité à distinguer pour en conserver une vision simple est la vérité déductive. Ce sont les plus simples à reconnaître car elles sont toujours du type « Si tout B est C, et que A est B, alors A est C ». Toutefois, elles se distinguent des vérités de type vérité-cohérentes, car le raisonnement déductif se base toujours sur des axiomes sensés aller de soi, mais que l'on accepte au conditionnel. Dans la théorie de la cohérence, les propositions initiales sont considérées comme des lois ou des absous.

Enfin, les vérités inductives sont le dernier type de vérité qu'Ellerton distingue. À l'inverse des déductions, les vérités inductives nous permettent d'extraire des conclusions suite à l'observation empirique d'une notion de causalité, comme l'apparition d'un champ magnétique suite à la génération d'un courant électrique par exemple. Ces inférences, une fois combinées, permettent ensuite d'être généralisées pour fournir une théorie sinon unifiée et absolue, du moins suffisamment solide pour être prédictive avec une faible marge d'erreur. Dans cette vision, le terme vrai n'est pas utile car il ne représente aucune réalité en lien avec le domaine de l'induction. Le vrai représentera plutôt ici un fort degré de confiance dans une prédition. Ce mode de pensée, bien que courant et répandu dans le milieu académique, n'est pas sans faille non plus, car il suppose deux problèmes :

- Le problème de généralisation : Dans l'induction, des conclusions seront portées sur l'ensemble de la population à partir d'un échantillon. Ainsi, si l'on n'a observé que des cygnes blancs, on ne peut pas conclure à la non existence des cygnes noirs.

- Le problème de causalité : Développé sous le principe « d'uniformité de la nature » (*An Enquiry Concerning Human Understanding*. David Hume., 1748), ce problème consiste à dire que les événements prédis ne se déroulent pas systématiquement de la même manière qu'ils l'ont fait dans le passé. Ainsi, une observation à un temps  $t$  ne tiendra pas toujours à un instant  $t+1$ .

Ce problème fut un de ceux qu'essaya de développer Sir Karl Popper, l'un des philosophes des sciences contemporains les plus influents du XX<sup>ème</sup> siècle. Dans son œuvre *La Logique de la découverte scientifique* (Devaux et al., 1973), Popper constate que les questions et problèmes du raisonnement inductif ne sont pas forcément les bons, car l'induction n'est pas là pour justifier une théorie. Elle est seulement là pour permettre de corriger les erreurs des conclusions réalisées à la suite d'observations, si erreur ou incohérence il y a.

Pour Popper, la science n'est pas là pour déclarer une théorie comme vraie ou fausse, car aucune expérience ne peut prouver une théorie scientifique. La qualité d'une théorie scientifique est au contraire de pouvoir être critiquée et questionnée par une seule observation ou expérience. C'est ce qu'il développa sous la notion de falsification. Pour Popper, une théorie scientifique ne peut donc être qualifiée de vraie et absolue, mais y préfèrera la notion de certitude, qui est quant à elle subjective. Dans ce contexte, une vérité peut donc être décrite comme une certitude consensuelle appuyée par des observations, mais que l'on ne peut considérer comme absolue.

#### 1.1.4 - Définir une fake-news

Comme nous l'avons vu, plusieurs approches conceptuelles peuvent être reliées à la notion de fake-news et d'information vraie. Entre la désinformation, la mésinformation, l'information déformée volontairement ou involontairement, on peut alors se retrouver avec une constellation de termes pourtant bien distincts les uns des autres. Les fake news ont pourtant reçu une définition permettant de dissiper toute ambiguïté. Ces dernières sont définies comme « *des informations fabriquées imitant le contenu médiatique dans la forme, mais ignorant leur intention ou processus organisationnel caractéristique* » (traduit de Lazer et al., 2018). Les fake news sont en effet dénuées de toute norme éditoriale et s'affranchissent de toute vérification de la fiabilité, de la précision ou de la crédibilité de l'information. A l'inverse, la mésinformation peut-être, elle, définie comme une information manipulée pour être dirigée vers un but précis, lorsque la désinformation vise quant à elle uniquement à tromper ses lecteurs.

Les normes entourant le contenu médiatique sont apparues dans les années 1920, à la suite des campagnes de propagande médiatique ayant eu lieu pendant la Première Guerre

Mondiale. C'est aussi l'époque des premiers canulars radio où le Père Ronald Arbuthnott Knox réalisa un faux bulletin d'information informant que Londres était attaquée par des communistes, que le Parlement était en état de siège, et que Big Ben et l'Hôtel Savoy avaient été détruits dans des explosions, causant une panique heureusement relativement faible dans le pays. S'en suivirent plusieurs dont le plus célèbre était celui d'Orson Welles, qui était de base la simple lecture d'une œuvre de science-fiction et non un canular volontaire (Burkhardt, 2017).

Avec internet sont apparus les premiers sites publient spécifiquement des fausses informations parodiques (eg. Dhmo.org parlant des « *dangers du monoxide de dihydrogène* ») ou non (eg. MartinLutherKing.org, site web créé par un groupe de suprémacistes blancs). De la même manière, le respect des normes de publication de l'information a pu être mis en compétition avec du contenu favorisant l'acceptabilité par le grand public. Par exemple, l'arrivée des réseaux sociaux a changé l'impact du contenu d'un article par l'impact de son titre. Les personnes se rendant sur les réseaux sociaux étant sélectives sur le contenu qui retiendra leur attention et les amènera à cliquer, la pertinence du titre n'a désormais plus à être en rapport avec le contenu de l'article, car le principal intérêt des diffuseurs de contenu sera d'attirer l'attention des utilisateurs pour que ces derniers passent le plus de temps possible à consommer leur contenu. Le développement du financement de plateformes de diffusion de contenu par les agences publicitaires a ainsi renforcé certaines pratiques visant à rendre l'information attractive au plus grand nombre le plus vite possible. Celui de l'analyse de données a aussi permis de mesurer l'étendue de l'influence de chaque plateforme, et de chaque type de contenu.

Comme le dirent Matthew A. Baum et David Lazer, deux chercheurs spécialisés dans le domaine des fake news, dans une interview au Los Angeles Times<sup>7</sup> : « *Ce que l'on sait, c'est que les phrases choquantes restent en mémoire. Une part considérable des articles de recherche montrent que les individus auront davantage tendance à être attentifs à, et à se rappeler plus tard, des titres à sensation ou négatifs, même si ce dernier est indiqué comme étant trompeur* ». Ainsi, nous sommes cognitivement amenés à favoriser certains types de contenus plutôt que d'autres, et ce sont sur ces biais que sont capitalisés les investissements sur la diffusion de contenu. C'est ainsi que sont financées et c'est ainsi que se développent les fake-news, en se concentrant sur le sensationnel davantage que sur la vérité et l'authenticité de l'information.

Il nous reste enfin à mettre les fake-news en perspective avec les autres termes couramment employés, mais parfois sans en saisir les nuances, pour désigner des fake-news. Verstraete et al. (2017) ont ainsi proposé dans un article une matrice permettant de distinguer les notions de satire, de hoax, de propagande, de *trolling* et de parodie. Une version traduite de

cette dernière est présentée en figure 1. Cette typologie n'est pas unique (eg. Waldrop, 2017), mais elle permet une lecture sur deux axes clairs et simples à comprendre. Elle scinde les fake-news sur un premier continuum, les enjeux financiers et culturels. En effet, certaines fake-news peuvent avoir des buts corporatistes, politiques ou économiques, lorsque d'autres peuvent en avoir les retombées, mais pas les intentions. Le deuxième axe sur lequel ces fake-news sont distinguées sont les intentions de duperie. Le but d'un auteur peut en effet d'être de tromper le maximum de personnes pour atteindre son but, ou bien d'agir en étant conscient de ne tromper personne.

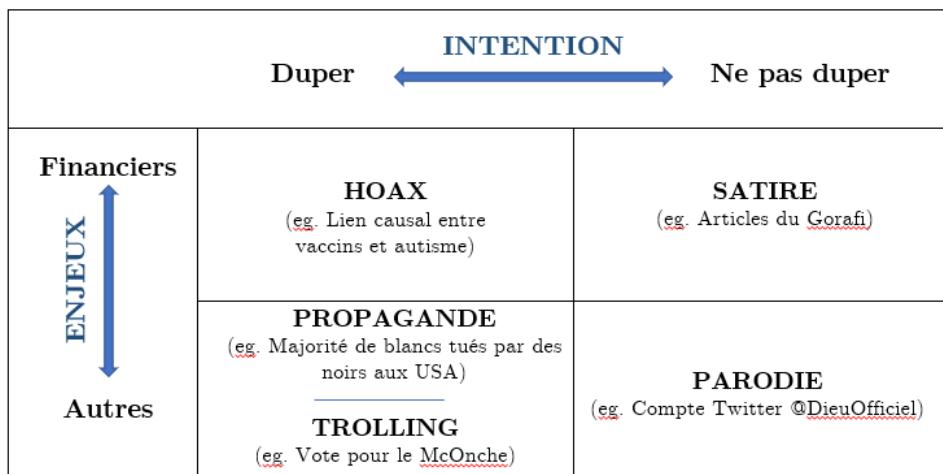


Figure 1 : Types de fake-news selon Verstraete et al. (2017)

Les satires et les parodies font ainsi partie des fake-news dont l'intention n'est pas de dupler. Toutefois, le but de sites web satiriques comme le Gorafi ou TheOnion sont bien d'avoir un impact sociétal, en proposant des critiques plus ou moins pertinentes de faits de société. Par exemple, un de leurs articles, au sein d'une controverse sur la qualité de vie des étudiants titrait : « *Emmanuel Macron aux étudiants : ‘Je n’adore pas le mot de précarité car cela donne l’impression que vivre dans la pauvreté est pénible’* ». Les parodies n'ont quant à elles aucune ambition financière ou sociétale, comme l'illustrent plusieurs comptes parodiques sur Twitter (@dieuoff, @ComplotsFaciles, @JonyIveParody...).

Du côté des fake-news dont l'intention est de tromper les gens, on peut à nouveau distinguer celles ayant des enjeux sociaux et financiers, auquel cas ces informations sont qualifiées de hoaxes. Un des hoaxes qui s'est le mieux répandu en France et dont l'empreinte plus de 15 ans plus tard reste encore complexe à dissiper (Larson et al., 2016) concerne le lien causal entre vaccin ROR et autisme. Cette étude, publiée à l'origine dans The Lancet, un journal médical de renommée, fut retirée peu de temps après sa publication. Son auteur, Andrew Wakefield, fut reconnu de comportements contraires à l'éthique, d'avoir faussé ses résultats et

était au centre d'un conflit d'intérêt dans le cadre du financement de produits qualifiés « d'alternatives à la vaccination ». La raison de son hoax était ici clairement d'induire les gens en erreur afin d'amener à des conséquences financières qui lui seraient favorables.

Enfin, lorsqu'on s'écarte des fake-news ayant des ambitions financières ou sociétales, on rencontre alors la propagande. Cette information volontairement biaisée vise à promouvoir une cause politique ou un point de vue, sans que la cible ne se doute de l'intention de départ. On peut y ranger l'affirmation de Trump sur les personnes noires responsables du meurtre des blancs mentionnée en introduction. Enfin, ceux cherchant à duper le plus grand nombre sans en tirer un quelconque bénéfice sont ceux appelés les trolls. Un des exemples les plus flagrants fut lors d'une campagne promotionnelle de MacDonald's en Suisse pour inciter les gens à voter pour leur burger favori parmi des créations en ligne. Des participants du forum 15-18 de jeuxvideo.com permirent aux deux burgers finalistes d'être sélectionnés, en faisant apparaître en finalistes les « McOnche » et le « Classic1518 », dont la plupart des consommateurs ne comprendraient pas le sens, ces mots étant des références internes explicites uniquement pour les utilisateurs du forum. Ces références sont considérées comme des « *private jokes* », des blagues originaires d'une communauté que seule cette dernière est en mesure de comprendre. La plupart du temps, la seule finalité du trolling reste l'humour, propre à son créateur.

Ces dimensions ne sont toutefois pas fixes, et elles ne sont pas non plus immunisées face à la critique. Toutefois, l'intention et la finalité restent deux axes à prendre en compte dans la qualification d'un type de fake-news.

### **1.1.5 - La vérité à l'ère de la post-vérité**

A l'ère de la communication, le besoin de vérité dans le processus de prise de décision se fait de plus en plus omniprésent. Des questions cruciales pour l'humanité contemporaine telles que « *Y a-t-il un dérèglement climatique ? Si oui, quelles en sont les conséquences ? Peut-on le prévenir, et comment ?* » commencent à émerger à l'échelle mondiale (Lee et al., 2015), et les résultats scientifiques sont souvent décisifs dans la prise de décision pour répondre à ces nouvelles interrogations. En effet, de par sa méthodologie, sa recherche de reproductibilité et sa faillibilité, les résultats peuvent par la suite être traduits en lignes directrices pour l'implémentation de politiques locales ou globales (mise sur le marché d'un nouveau médicament par exemple). Bien évidemment, la prise de décision n'est pas systématiquement rationnelle, car elle peut se baser sur d'autres réalités, ou vérités, mais l'expertise scientifique demeure une des seules formes d'expertise communément admise dans les instances décisionnelles de haut niveau.

Cependant, c'est au moment de la communication des résultats scientifiques que le principe même de vérité en tant que fidélité à la source pose problème. En effet, les enjeux des institutions non-académiques (politiques, médias, finances...) ne sont pas ceux des universitaires, dans un référentiel occidental capitaliste. Par exemple, le profit à court terme, la recherche d'immédiateté et la croissance feront partie des premiers lorsque les besoins de visibilité, de compétitivité et de contribution à long terme définiront, parmi d'autres, les enjeux académiques. Ainsi, leur notion même de *vérité* n'aura pas la même orientation philosophique que celle des institutions académiques.

Historiquement d'ailleurs, la connaissance scientifique a toujours été opposée au « *bon sens commun* ». La science est dite de produire une « *vraie* » connaissance, plus exacte que le bon sens, qui nécessite d'être vulgarisée pour être transmise au plus grand nombre (Weingart, 2002). En cela, cette complexité propre à la science permet sa production de résultats moins ambigus que ceux que l'on peut acquérir par le sens commun. Toutefois, elle ne se place pas dans le même spectre de véridicité des résultats que le public ou les médias ne pourront le faire. Pour vérifier si un résultat peut être qualifié de « *vrai* », ou de représentatif, les scientifiques émettront un ensemble d'hypothèses issues d'une ou plusieurs théories, dont l'opérationnalisation permettra d'en tester la robustesse. Dans cette optique, il n'y aura pas de résultats vrais ou faux. Certains résultats iront dans le sens ou non d'une hypothèse, et des fluctuations statistiques pourront confirmer la présence d'effets là où ces derniers sont objectivement absents (eg. *The Dead Salmon Study*). Toutefois, ces résultats ne seront pris en compte qu'en ce qu'ils peuvent apporter ou réfuter à une théorie, qui restera-t-elle plus ou moins robuste en fonction desdits résultats.

C'est pourquoi il est difficile de concevoir qu'un scientifique puisse ne pas se tromper, lorsqu'on se place dans son propre référentiel épistémologique. Toutefois, la vitesse des besoins d'actions et de prise de décision des instances médiatiques et politiques laisse peu de marge à l'erreur, et ces notions s'appliquent plus difficilement dans leur référentiel. Il y a donc un besoin de transmission de l'information scientifique qui puisse être fait en cohérence avec des notions plus absolues de vrai et de faux qu'en milieu scientifique, bien que cela se fasse nécessairement au détriment de la fidélité au message original.

La vulgarisation des résultats scientifiques est nécessaire pour permettre le changement de comportements, de politiques ou de normes, et les médias de masse sont par défaut le principal relais de l'information aux populations. Toutefois, comme toute vulgarisation implique une simplification, cette dernière peut d'une part être mal perçue des chercheurs eux-mêmes (D. R. Johnson et al., 2014), voire dériver en une contamination d'idées fausses, en ce

qu'elles ne seront pas fidèles à la source et à son éthique. Dans son article, Weingart (2002) pointe toutefois du doigt l'interdépendance nécessaire des médias et du monde universitaire.

Les médias n'évoluent pas dans la même temporalité, ni dans les mêmes besoins que les chercheurs académiques. Ils couvrent « *ce qui est existant, ce qui est actuel et ce qui est important* » (Dunwoody & Peters, 1993 ; cité dans Weingart, 2002). Ces critères de diffusion de l'information font bien partie des critères de recherche du monde scientifique, mais pas de diffusion des résultats de recherche. Ainsi, il y aura forcément une différence entre la couverture médiatique et la recherche scientifique. Suite à la crise du COVID-19, l'interdépendance entre ces deux mondes est devenue plus flagrante que jamais. Il faut cependant noter que la diffusion de la recherche dans la presse ne pourra jamais être *vraie*, c'est-à-dire fidèle à la source, car les médias de diffusion n'opèrent pas dans la même temporalité et complexité que la recherche.

L'utilisation des médias comme canaux d'information et de diffusion des résultats scientifiques a toutefois énormément évolué au cours des dernières décennies, notamment avec l'arrivée d'internet qui a permis d'alerter plus facilement sur des sujets sociaux, mais aussi d'être plus visibles auprès des médias eux-mêmes. Cette visibilité permet, en retour, une attractivité financière pour leur permettre de financer et de mettre en place divers projets d'amélioration (traitements, technologies, prises en charge, etc...). Par exemple, le scientifique James Hansen pu ainsi alerter l'opinion publique et politique des problèmes liés au dérèglement climatique et aux rythmes alarmants de la montée des eaux dans un article du New York Times<sup>8</sup> avant que son article ne soit publié dans la revue Science. La répercussion qu'eut le professeur Hansen permit ainsi le développement ultérieur des recherches en sciences environnementales.

Cette interdépendance peut toutefois conduire à des dérives des pratiques scientifiques pouvant remettre en cause la crédibilité du milieu académique. C'est ainsi par exemple qu'ont pu apparaître des pratiques telles que celles mentionnées dans le cas du « Climategate ». Cette affaire est apparue à la fin des années 2000, lors de la montée en puissance des inquiétudes liées au climat aux Etats-Unis. En 2009, des pirates informatiques ont réussi à pénétrer le serveur de la University of East Anglia aux Royaume-Unis, mettant à disposition des échanges privés ayant lieu au sein de l'Université. Bien que dans l'ensemble le contenu fut peu inquiétant, une série d'échanges d'universitaires montra des pratiques non éthiques au sein d'un laboratoire spécialisé dans l'étude du réchauffement climatique, la Climatic Research Unit.

Cette unité de recherche était très influente auprès des experts mondiaux s'intéressant aux problèmes climatiques (comme le Groupe d'Experts Intergouvernemental sur l'Evolution du Climat, ou GIEC). Parmi les accusations, on retrouvait des pratiques de manipulation de présentation de l'information afin de rendre leur discours plus impactant. Il y avait aussi des

pratiques de blocage du processus de revue par les pairs, qui permettait aux articles publiés de ne l'être que si ces derniers allaient dans le sens du réchauffement climatique, ou encore de manipulation des outils de mesure pour que les données aillent dans le sens du discours.

Ces pratiques furent jugées trop marginales pour être caractéristiques du milieu, et l'impact sur la littérature en elle-même fut assez restreint. Toutefois, les conséquences d'une telle affaire se révélèrent désastreuse pour l'opinion publique (Leiserowitz et al., 2013). L'éthique même de la profession en fut entachée et la crédibilité des sources scientifiques en fut diminuée. Une telle affaire reste cependant intéressante pour étudier l'ethos de la recherche scientifique. En effet, la communauté scientifique est souvent à l'origine de résultats impactant significativement la société (la découverte des plaques amyloïdes dans Alzheimer, le lien de causalité entre le tabagisme et le cancer), et il est assez simple d'envisager l'importance de la nécessité d'obtenir une réaction publique à ce genre de problèmes. On peut donc difficilement définir dans ce contexte le *vrai* d'une information, et on peut aussi comprendre le sentiment de nécessité de rendre l'information plus impactante pour qu'elle ait une véritable répercussion auprès du grand public.

La question du lien entre le devoir moral face au corps d'évidence et l'exigence tout aussi morale de rigueur méthodologique sont ici au cœur des enjeux de l'éthique scientifique, et continuent de faire débat (Grundmann, 2013) à une époque où les pressions économiques sur la recherche deviennent de plus en plus fortes (Hong & Walsh, 2009). Ces réflexions éthiques sont au cœur de ce que l'on peut considérer comme vrai ou non dans le flot d'information constant de l'ère du numérique. La validation par la communauté scientifique d'une vérité telle que définie ici, hors de sa propre réalité (scientifique ou commune) est donc actuellement nécessaire pour que le milieu académique continue de rester un acteur privilégié dans la création de recommandations et dans le processus de prise de décision.

### **1.1.6 - L'Amplification Sociale du Risque : Un background théorique de la propagation d'une information**

Comme nous avons pu le voir, un des points cruciaux dans la caractérisation de la véracité d'une information concerne la déformation nécessaire d'une information experte en une information accessible au grand public et aux politiques par les médias. Toutefois, cette déformation n'est pas toujours contrôlée et peut donner lieu à des problèmes majeurs par la suite, notamment en termes de difficultés d'éradication des croyances inexactes, ce qui sera développé dans la deuxième partie de cette introduction. Le framework d'amplification sociale du risque (ou SARF, pour Social Amplification of Risk Framework) fit partie des premières

tentatives d'analyse pluridisciplinaire pour évaluer la déformation d'information qui peut avoir lieu dans la communication de masse. Ce cadre théorique fut développé au sein d'un projet d'enfouissement de déchets nucléaires aux Etats-Unis à la fin des années 1980 et visait à intégrer des facteurs de niveau individuels sur la question de l'acceptabilité du projet avec les perspectives sociales et culturelles liées à la perception du risque. Le SARF est donc ici une tentative de collaboration visant à combler les disparités entre les théories en intégrant les connaissances issues de l'analyse du risque, des perceptions du risque et de la sociologie du risque (R. E. Kasperson et al., 1988). Il est à noter que le risque dont il est question dans ce framework n'est pas le risque physique ou vital. Le risque est ici davantage défini comme le résultat d'un processus d'interprétation individuelle ou groupale d'un ensemble de menaces.

Le SARF se décompose en deux principales étapes. La première concerne le filtrage et le formatage de l'information au moment de sa diffusion. L'information circule alors dans des réseaux de communication désignés sous le terme de « stations ». Ces stations peuvent être des acteurs sociaux, des universitaires, des institutions scientifiques, des politiciens, ou encore des médias de masse (J. X. Kasperson et al., 2003). La deuxième étape vise quant à elle à évaluer l'impact d'un risque en classant la diffusion des effets dans la société. Les impacts sociaux sont considérés comme des ricochets, diffusant leurs effets aux personnes directement concernées en premier, puis se répandant à la communauté locale, aux groupes professionnels, puis aux responsables et dirigeants avant d'impacter au final la société dans son ensemble. Les conséquences de ces ricochets se répercutent ensuite sur l'économie, la politique, et les notions de confiance. En retour, ces conséquences impactent elles-mêmes l'augmentation ou la diminution du risque lui-même (voir figure 2).

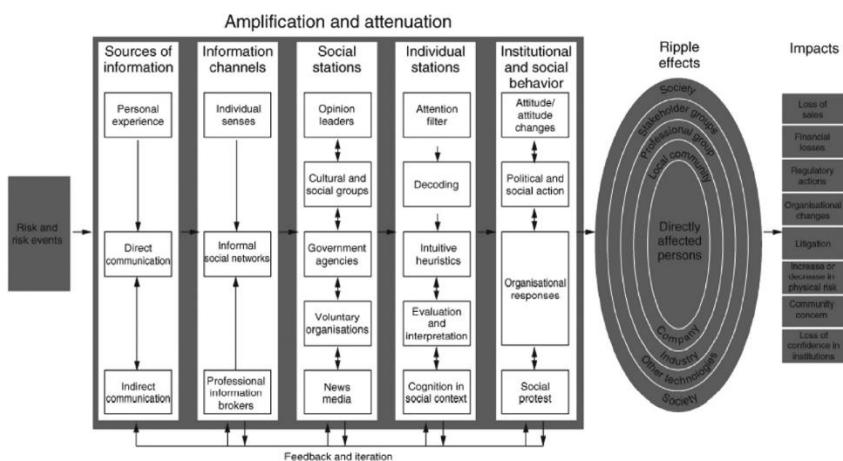


Fig. 1. The Social Amplification of Risk Framework (from Pidgeon et al., 2002, p. 14).

Figure 2 – Social Amplification Risk Framework (R. E. Kasperson et al., 1988)

Bien qu'il soit parfois jugé comme trop général ou peu informatif, le SARF sert davantage de cadre contextuel que de cadre prédictif. Son application au risque sert par exemple actuellement de référence au développement des théories la propagation des croyances au changement climatique (Mase et al., 2015) ou encore celles liées au vaccins (Carper, 2019). Dans le cadre de la propagation des fake-news, le SARF apparaît extrêmement approprié pour comprendre le phénomène de par sa prise en compte de la variété des facteurs sociétaux, inter et intra-individuels en jeu. Les fake-news sont en effet ici un facteur dans la chaîne de traitement des diverses informations pouvant amplifier la perception du risque d'un phénomène. Leurs impacts sur des phénomènes majeurs sont désormais avérés, que l'on se situe du point de vue politique (Allcott & Gentzkow, 2017) ou sanitaire (Carrieri et al., 2019). Les fake news sont une partie intégrante du processus d'amplification du risque. Notre but est donc ici d'essayer d'identifier les facteurs facilitant ou diminuant l'amplification du risque.

## 1.2 – Les mécanismes psychosociaux au cœur de la recherche d'informations

### 1.2.1 - Je reconnaiss donc je sais : Le biais d'exposition

Toutes les questions mentionnées ci-dessus peuvent-être considérées sous le prisme d'un élément majeur dans la chaîne de développement des véritudes : L'individu, et la façon dont il percevra et traitera l'information. L'analyse du traitement de l'information des individus fait partie intégrante des études conduites en psychologie depuis plusieurs décennies. Ce sont sur ces résultats que les techniques de stratégie de communication se sont fondées.

Dans son article, Schwarz et al. (2016) montre en quoi les techniques actuelles de réfutation des fausses informations seraient inefficaces, voire néfastes, en s'appuyant sur des concepts simples issus des travaux en psychologie cognitive et sociale. Il cite ainsi le format « *mythes vs. faits* », couramment employé pour rectifier une fake-news. Ce format peut en effet provoquer l'inverse de l'effet attendu et renforcer la fake-news en lui donnant davantage d'importance et de visibilité. Cet effet, appelé le biais d'exposition et publié par Robert Zajonc (1968), implique que le simple fait d'avoir été exposé à un stimulus dans le passé améliore de fait l'attitude d'un individu vis-à-vis de ce stimulus. L'effet d'exposition a été étudié très tôt, et déjà avant l'ère d'internet, on montrait qu'une information à laquelle des participants étaient exposés de manière répétitive semblait plus vraie qu'une information nouvelle (Arkes et al., 1989; Hasher et al., 1977).

L'exposition première à une information rend en effet cette dernière plus familière, et c'est ce sentiment de familiarité qui par la suite provoque l'impression qu'une information est plus vraie, cela même si on prévient les participants que la source n'est pas forcément crédible (Bacon, 1979; Begg et al., 1992). Ces résultats sont connus dans la littérature scientifique en psychologie sous le terme « *d'effet de la vérité illusoire* » (Hasher et al., 1977). De plus, les individus victimes de cet effet auront même tendance à confabuler sur le contexte d'apprentissage de l'information pour rendre cette dernière plus crédible, et à mal attribuer la source de cette information (Polage, 2012), résultant en la croyance en une fausse information.

Ainsi, si nous souhaitons développer un algorithme de reconnaissance des fake-news, il sera alors constructif de connaître les mécanismes au cœur du traitement de l'information des individus tels qu'identifiés par Schwarz et al. (2016). Ces mécanismes permettront en effet d'identifier quelles variables pourraient primer dans l'analyse du contenu d'une page web ou encore sous quelle forme présenter les résultats de cette analyse. Enfin, ils permettront aussi de mieux comprendre pourquoi les fake-news sont si répandues et tenaces sur internet.

### **1.2.2 - Les tenants psychologiques de la croyance en une information**

#### **L'impression de consensus social**

Lorsque nous devons évaluer la fiabilité d'une information, il est fréquent que nous nous tournions vers l'opinion générale. Les gens auront en effet davantage confiance en leurs affirmations si d'autres les soutiennent (Festinger, 1954), ou pourront encore se soumettre à l'avis majoritaire pour évaluer la véracité d'une information (Asch, 1956). Toutefois, plusieurs biais viennent gêner la perception que l'on se fera d'un consensus social. En effet, plus que l'opinion générale, c'est la fréquence d'apparition d'une information – et donc sa familiarité – que nous pouvons juger.

Pour étudier ce biais, Weaver et al. (2007) ont réalisé une série d'étude dans laquelle des participants entendaient une information selon 3 conditions :

- Ecoute d'une opinion une seule fois par un individu
- Ecoute d'une opinion trois fois par un même individu
- Ecoute d'une opinion une seule fois par trois individus

Sans surprise, les participants dans la condition de répétition (deuxième condition) et dans la condition de consensus social (troisième condition) trouvaient l'opinion plus fiable que ceux dans la première condition d'exposition simple. Toutefois, même si les participants dans la condition de consensus social donnaient la plus haute note de fiabilité de l'opinion, celle de la deuxième condition de répétition était une note égale à 80% de la note du consensus. Ces

résultats expriment ainsi clairement que la simple accessibilité d'une information en mémoire augmente la perception que nous aurons de sa fiabilité. Ils montrent aussi à quel point le simple fait d'entendre plusieurs fois une information issue d'une même source rend cette information presque aussi fiable que si elle avait été entendue de plusieurs sources différentes.

Ces résultats suggèrent donc que l'exposition répétée d'une information, peu importe sa véracité, suffit à la rendre crédible et peut donner l'illusion de la popularité de cette information. Ce biais peut-être d'autant plus critique sur les réseaux sociaux, où les algorithmes de recommandation de contenu enferment les utilisateurs dans des « bulles » d'information. Ces bulles ont tendance à exposer des contenus cohérents avec les intérêts de l'utilisateur et ne chercheront pas à le guider vers des informations contradictoires<sup>9</sup>.

### **La corroboration**

Avec l'accès à internet, quasiment toutes les possibilités de vérification de l'information et de l'accès à des sources diverses permettant de corroborer une information sont devenues possibles. Deux possibilités de vérification des sources ont été observées : Une approche complexe, demandant la vérification via plusieurs médias, la lecture en détail de l'information et la remontée à la source de l'information (voire la lecture d'articles scientifiques). La seconde approche pourrait être qualifiée d'automatique, sans efforts ou rapide : Si une information est supportée par plusieurs sources et preuves, il sera alors plus simple de trouver des preuves tierces et on pourra plus simplement citer de mémoire d'autres exemples corroborant cette information.

On pourrait donc se dire que la méthode la plus efficace pour encourager l'esprit critique d'un individu devant une information serait d'inciter à utiliser l'approche complexe, toutefois, les résultats empiriques montrent que cela pourrait aussi être délétère. En effet, l'abondance d'exemples corroborant une information peut diminuer l'influence de cette information. Ainsi, dans une étude où ils devaient se souvenir de deux ou dix exemples de situations où ils avaient été sûrs d'eux, des participants se sentaient plus confiants en eux-mêmes lorsqu'ils ne devaient fournir que deux exemples que lorsqu'ils devaient en fournir dix (Schwarz et al., 1991).

De la même manière, l'effort fournit pour réfuter une information est inversement corrélé avec la croyance d'un individu dans cette information (Sanna et al., 2002). Dans l'ensemble, ces résultats tendent vers une même conclusion. Une grande complexité et une grande quantité d'exemples seront moins efficaces pour corroborer ou réfuter une information qu'une poignée d'exemples simples et rapidement accessibles.

### **La cohérence avec les croyances initiales**

Le rôle de la consistance de l'information avec les croyances initiales dans la facilitation du traitement de l'information a été amplement documenté depuis les années 1950 (Festinger, 1962; Winkielman et al., 2011). Lorsqu'on fait face à une information incohérente, le temps de traitement de cette information est augmenté, et à l'inverse, les informations cohérentes sont traitées plus rapidement que des informations contrôle.

Par exemple, lorsqu'on demande à des participants « *Combien d'animaux de chaque espèce Moïse a-t-il emmené sur son arche ?* », la plupart des gens répondent « *Deux.* », bien qu'ils sachent que le protagoniste de cette histoire était Noé. Mais le thème de la question étant familier, la plupart des gens se concentrent sur ce qu'on leur demande plutôt que sur les détails de la question (Erickson & Mattson, 1981).

Toutefois, si l'on impose aux participants de répondre au même genre de question, mais dans un contexte où le traitement de l'information est rendu plus complexe (eg. Typologie dégradée), davantage de personnes reconnaissent qu'il y a une erreur (Song & Schwarz, 2008). Cela pointe vers la nécessité d'une cohérence structurelle (eg. L'aspect formel de la question) ainsi qu'une cohérence sémantique (eg. Respect d'une même thématique) pour fournir ce que l'on appelle une cohérence du traitement de l'information (Winkielman et al., 2011).

### **La cohérence interne**

Nous aurons davantage tendance à croire une information comme vraie si chaque élément de cette information s'insère dans une histoire cohérente (Schwarz et al., 2016). Par exemple, un verdict judiciaire aura davantage de chances d'innocenter un homme si les preuves sont racontées dans une histoire que si elles sont présentées dans un résumé factuel (Pennington & Hastie, 1992).

Ce besoin de cohérence interne est très similaire au fonctionnement de notre mémoire à long terme, et c'est de ces études qu'est apparu le champ d'étude des faux souvenirs (Loftus, 1996). Nous avons en effet naturellement tendance à vouloir « combler les blancs » et rajouter des détails à nos souvenirs, car ces derniers nous paraissent « vraisemblables » ou semblent découler de manière logique des autres informations dont nous disposons. Ce biais est si fort qu'il peut nous amener jusqu'à affirmer que l'on se souvienne d'un événement que l'on n'a jamais vécu ou d'une information que nous n'avons jamais perçue (Gerrie et al., 2006; Shaw & Porter, 2015).

### **La crédibilité de la source**

La question de la crédibilité de la source pour se fier à une information est probablement une des plus complexes à étudier, car c'est aussi celle étant la plus sujette aux biais individuels.

En effet, bien qu'une information soit considérée comme plus acceptable lorsqu'elle vient d'une source considérée comme fiable que lorsqu'elle vient d'une source jugée peu crédible (Petty & Cacioppo, 1986), la question de la légitimité de la source reste quant à elle sujette à la libre interprétation de chacun.

Ainsi, un individu sera jugé moins crédible si son accent diffère du nôtre (Lev-Ari & Keysar, 2010), si son apparence diffère de trop de la nôtre (Waytz, 2014) ou encore si son nom est trop compliqué à prononcer (Newman et al., 2014). Plus une personne nous semblera familière, plus nous aurons tendance à trouver cette source crédible, à moins que cette source ne soit familière pour de mauvaises raisons (eg. Donald Trump). Enfin, si le traitement de l'information est rendu simple par la source, le message sera alors plus simple à considérer comme vrai (Lev-Ari & Keysar, 2010), ce qui explique pourquoi un message accompagné d'une illustration est davantage accepté (Newman et al., 2012).

Pris dans leur ensemble, tous ces résultats tendent vers une seule et même direction : Un message qui sonne familier, consensuel, consistant, en accord avec les croyances d'un individu et dont la source semble familière, aura davantage tendance à être accepté par un individu sans que ce dernier ne se pose trop de questions. A l'inverse, les informations demandant un effort mental seront plus difficiles à traiter. Ces dernières pourront alors amener à une analyse critique si le traitement de l'information ne contredit pas les croyances initiales de la personne recevant l'information. Dans le cas contraire, le message aura alors davantage tendance à être rejeté.

C'est pour cela que l'article de Schwarz et al. (2016) incite à ne pas juxtaposer mythes et faits lors du démontage d'une fausse information. Cette juxtaposition a en effet tendance à répéter le message initial, peut potentiellement toucher de nouvelles audiences et peut amener à créer des clivages là où il ne devrait pas y en avoir. Par exemple, le fait d'exposer ce que pensent les climatosceptiques et ce qu'en pense la communauté scientifique aura davantage tendance à créer des scissions et à porter aux devants de la scène les arguments des sceptiques. Enfin, ces résultats indiquent aussi l'importance de la vulgarisation scientifique et de l'accessibilité tant financière que cognitive des résultats scientifiques qui à défaut de pouvoir être considérés comme « *vrais* » sont reproductibles et vérifiables.

C'est pour cette raison que nous sommes certains que si l'on souhaite éviter les méfaits de la désinformation, il faut porter davantage nos efforts sur la prévention que sur la critique d'une fake-news. Dans le cas d'un individu ignorant une information, un contexte d'apprentissage opposant mythes et réalité pourrait être bénéfique. Toutefois, comme nous l'avons vu, dans le cas de la propagation des fake-news, nous nous situons davantage dans le cas où un individu a déjà été influencé par une information fausse, mais qu'il aura davantage

tendance à croire vraie parce qu'il y aura déjà été exposé. Notre but est donc ici davantage préventif, et c'est dans le respect de cette optique préventive que nous pensons que l'algorithme que nous développerons devrait indiquer à quelqu'un visitant une page si cette dernière présente des signes classiques de désinformation ou non. De même, et sur une note plus globale, faire de la prévention sur les risques de la désinformation ne pourrait-être que bénéfique, les individus ayant en effet tendance à porter un regard plus critique sur les informations qu'on leur donne si ils ont été exposés à un message d'avertissement auparavant (Ecker et al., 2010).

### 1.2.3 - La recherche d'information à l'ère d'internet

Une des raisons pour lesquelles la société actuelle est appelée « *société de l'information* » est l'explosion en quelques décennies de l'accessibilité au contenu mondial de quasiment toute information possible. Bien que cette quantité ne change pas la manière dont on traite et évalue la crédibilité d'une information, elle a bien fait évoluer la fréquence avec laquelle nous sollicitons ces compétences (Flanagin & Metzger, 2008). Toutefois, le but des fake-news étant de se rapprocher autant que possible des sources d'informations fiables, le développement d'outils critiques d'évaluation commence à apparaître. En effet, lorsqu'auparavant, les informations étaient diffusées par une quantité restreinte d'éditeurs de contenus, l'accessibilité et la diversité des contenus font qu'il s'agit désormais pour tout un chacun de trouver l'information répondant au mieux à ses besoins, et donc celle que l'on trouve la plus fiable.

A partir de là, deux moyens s'offrent à tout un chacun pour évaluer la pertinence d'un contenu : S'engager dans une évaluation consciente et demandant des efforts, ou se baser sur des heuristiques permettant un traitement automatique et plus aisé. Les deux ne sont pas mutuellement exclusifs. L'évaluation consciente de la fiabilité d'une information en ligne repose sur 5 facteurs (Metzger, 2007). Tout d'abord, la *précision*, qui correspond au degré avec lequel un site web apporte plus ou moins d'erreurs. Ensuite, viennent l'*autorité* du site web (qui a écrit l'article, sur quelle plateforme ?), et l'*objectivité*, qui consiste à identifier si l'information présentée consiste en une opinion ou en un fait, si des conflits d'intérêts apparaissent, ainsi que les liens entre les sources et l'article. Enfin, l'*actualité* réfère à la mise à jour et à la pertinence contemporaine des informations et la *couverture* concerne la profondeur de traitement d'une information et son étendue.

Toutefois, la recherche en psychologie a depuis longtemps montré que l'engagement dans un processus contrôlé et coûteux en ressource n'est pas systématique chez l'être humain, et que la plupart des traitements sont automatisés (Bargh, 1994; Bargh et al., 2001, 2012;

Chaiken, 1980; Pyszczynski et al., 1999; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). Par exemple, une étude de Fogg et al. (2003) a montré qu'un des critères les plus importants pour évaluer la crédibilité d'un site web était l'aspect visuel et l'agencement des éléments sur le site web avant tout critère relatif au contenu de l'information. Enfin, les comportements rapportés de vérification à posteriori d'une information présentée sont très marginaux, les individus préférant généralement utiliser des stratégies de vérification requérant le moindre effort (Metzger, 2007).

Parmi les études sur l'évaluation de la crédibilité d'un contenu en ligne, une des premières fut celle de Sundar (2008), qui proposa le MAIN Model. Dans cet article, il proposa que les détails structurels, organisationnels et esthétiques d'un site web aidaient au déclenchement d'un traitement heuristique des informations de ce site web. Il proposa 4 facteurs déterminant les heuristiques d'évaluation : La modalité (texte, audio, vidéo), l'agency que l'on pourrait traduire par « contrôle ». Ce dernier réfère au contrôle perçu sur l'identité de la source de l'information (est-ce un post sur Twitter ? Un article scientifique ?). S'en suit l'interactivité, qui décrit le degré de contrôle et d'appropriation de l'information par le lecteur, et enfin la navigabilité, ou la facilité d'accès à l'information. Ces heuristiques restent encore en débat à l'heure actuelle, et d'autres sont encore développées (Metzger & Flanagin, 2013). Toutefois, il semblerait que la motivation soit un facteur clé dans le degré d'utilisation d'heuristiques ou de processus contrôlés employés pour évaluer une information (Metzger, 2007; Petty & Cacioppo, 1986). Ainsi, pris dans l'ensemble, ces résultats tendent à montrer le besoin d'un outil de vérification de l'information en ligne, car plusieurs facteurs peuvent déclencher un traitement heuristique de l'information. Toutefois, en plus d'une vérification externe, un entraînement individuel en faveur du développement d'un esprit critique et favorisant une recherche contrôlée de pertinence d'informations ne serait que bénéfique. Nous pourrons donc essayer d'intégrer des recommandations lors de la vérification automatique du contenu d'un site web.

#### **1.2.4 - La robustesse et le danger des fake-news**

Comme nous l'avons vu tout du long de cette introduction, les contraintes temporelles, cognitives et motivationnelles de tout un chacun limitent la capacité de recherche d'information, et donc la compréhension des différents enjeux sociaux, politiques ou économique contemporains. Ces contraintes provoquent comme effet le fait de ne pas toujours utiliser des processus coûteux d'évaluation de l'information, ce qui conduit parfois à lire et à croire des fausses informations, ou à minima, des informations déformées allant à l'encontre des faits. C'est ainsi par exemple que le candidat à la présidentielle américaine Donald J. Trump

relaya lui-même l'information depuis longtemps réfutée selon laquelle un lien entre vaccination et autisme existait<sup>10</sup>, car il l'avait vu sur les réseaux sociaux.

En dépit de cela, il est bien connu que le bon fonctionnement d'une démocratie dépend, entre autres, d'une population éduquée et bien informée. Les processus par lesquels les individus se forment des croyances et des opinions politiques, sociales ou médicales sont ainsi d'un intérêt public évident. Si une majorité de la population est susceptible de croire en une information factuellement erronée, les fake-news peuvent alors être à l'origine de décisions sociétales allant à l'encontre des intérêts des individus. Par exemple, suite à la controverse sur le lien entre vaccination et autisme dans les années 1990, l'Angleterre a vu son taux de couverture vaccinale chuter de 92% à 84% en 2002, allant jusqu'à 61% dans certains quartiers londoniens (Hussain et al., 2018). S'en est suivie une nouvelle épidémie de rougeole à la fin des années 2000 en Angleterre, mais aussi en France et aux Etats-Unis.

Un autre effet pervers de l'exposition à des fake-news concerne la force des convictions basées sur de fausses informations. Dans une étude menée par Leiserowitz et al. (2013), des individus ignorants sur le sujet du changement climatique montraient un faible support pour la cause des climatosceptiques. A l'inverse, les individus qui avaient reçu des fausses informations, et pensaient ainsi être les mieux informés sur le sujet, étaient ceux ayant les opinions les plus vivaces et robustes. Les coûts des fake news sont ainsi difficiles, voire quasiment impossibles, à ignorer dans notre société actuelle.

Différentes stratégies de correction ont été évaluées dans la littérature. Toutefois, dans l'ensemble, les stratégies montrent une certaine inefficacité à corriger des croyances basées sur de fausses informations. Cet échec à corriger de telles croyances est connu sous le terme d'*« effet d'influence continu »* (H. M. Johnson & Seifert, 1994). Dans l'étude de Johnson et Seigert (1994), on présentait une fausse histoire à des participants impliquant un incendie dans un entrepôt, dont on pense initialement que l'origine vient d'un placard où étaient stockés négligemment des barils d'essence et des peintures à l'huile.

Pour certains participants, l'histoire était ensuite rétractée via des phrases comme « *le placard s'est avéré être vide* ». On mesurait ensuite les références à l'essence et à la peinture dans une série de questions relatives à l'incendie. Les études utilisant ce paradigme ont montré de manière consistante que l'exposition à la phrase invalidant l'histoire n'avait que très rarement, voire jamais, un effet d'élimination sur la croyance en l'information initiale. Cet effet était présent et la réfutation d'une fake-news n'avait que peu d'effet, même si les individus croyaient, comprenaient et se souvenaient de la phrase d'invalidation (Ecker et al., 2010; Ecker, Lewandowsky, & Apai, 2011; Ecker, Lewandowsky, Swire, et al., 2011), voire pas d'effet (H.

M. Johnson & Seifert, 1994). Ce paradigme fut poussé pour étudier si la clarification de la rétractation pourrait avoir un effet plus efficace, mais de manière intéressante, en présentant une phrase telle que « *il n'y a jamais eu d'essence ni de peintures sur les lieux de l'incendie* », cette emphase sur la négation avait l'effet inverse, et les participants étaient alors d'autant plus enclins à croire en l'histoire (Seifert, 2002).

### 1.2.5 - Pourquoi préfère-t-on les histoires fausses ?

Parmi les explications à cette persistance des effets des fausses informations, quatre facteurs ont été retenus. Tout d'abord, la cohérence de la structure d'un événement fait partie des principales variables explicatives. Lorsqu'on raconte une histoire à un individu, ce dernier aura tendance à combler les vides présents dans l'histoire pour se créer sa propre version cohérente de la narrative. Nous cherchons naturellement à nous créer des représentations qui font sens, même si elles sont ou peuvent-être erronées, et nous préférerons toujours ces dernières à des histoires incomplètes (Ecker et al., 2010; Ecker, Lewandowsky, & Apai, 2011; H. M. Johnson & Seifert, 1994; Loftus, 1996). La correction d'une information peut en effet créer des vides dans les représentations mentales de l'histoire que se créent les individus, dont l'inconfort pourrait être supérieur à celui de savoir que la version que l'on s'est construite est partiellement incorrecte, mais cohérente.

Une autre possibilité vient d'un déficit des processus contrôlés de vérification de l'information. Tout d'abord, la confusion sur l'attribution de la source pourrait renforcer la croyance en l'information. Par exemple, un individu qui aura créé une narrative dans laquelle certaines informations sur l'origine de l'incendie venaient d'un rapport de police aura davantage de mal à réfuter ces arguments qu'un autre qui se souviendra que ces informations venaient d'une source moins fiable qu'un rapport de police. Enfin, la négation à posteriori d'une information peut rendre l'information plus confuse en mémoire, lorsque l'affirmation d'une réfutation peut-être plus simple à intégrer en mémoire. Par exemple, dire « *François n'aime pas le désordre* » est plus complexe à retenir que « *François aime l'ordre* » (Mayo et al., 2004).

Le troisième facteur, que nous avons vu plus tôt, concerne la familiarité de l'information, et la répétition provoquée par la négation d'une information. Enfin, un dernier facteur que l'on peut mentionner est le phénomène de réactance (Brehm & Brehm, 1981). Dans cette théorie, la réactance est « *un état motivationnel dirigé vers l'établissement d'une suppression ou d'une restriction de liberté* ». En somme, cet état décrit un mélange d'états cognitifs et émotionnels pouvant apparaître lorsqu'une restriction de libertés individuelles apparaît. Les techniques de communication persuasives peuvent ici être perçues comme des

menaces aux libertés individuelles (Steindl et al., 2015), les individus pouvant penser qu'on leur dise quoi penser. Les recherches dans ce domaine ont notamment été employées pour comprendre l'état psychologique de jurés assistant à un procès lorsqu'on présente à ces derniers des preuves allant à l'encontre de la narrative créée lors du procès (Lieberman & Arndt, 2000).

### **1.2.7 - Réduire l'impact des fake-news**

Nous avons pu montrer jusqu'ici que les fake-news sont un problème sociétal, intrusif et difficile à contrer. De plus, nous avons vu que les formats mythes contre réalité et la rétractation d'une information fausse ne permettaient pas de contrer les effets néfastes des fake-news. Toutefois, la recherche sur ces questions a déjà permis d'identifier trois facteurs influençant l'efficacité de la contradiction d'une fake-news (Lewandowsky et al., 2012).

#### **Les avertissements précédant l'exposition**

Les effets des fake-news peuvent être atténués si l'on expose de manière explicite les individus à un avertissement sur le fait que les informations présentées peuvent ne pas être fiables (Chambers & Zaragoza, 2001). Cependant, ces avertissements peuvent être moins efficaces que d'autres stratégies (Jou & Foreman, 2007) à moins de mettre l'emphase sur les dangers pour les individus des fausses informations (Ecker et al., 2010). Nous avons tous tendance à assumer par défaut la validité d'une information présentée, mais la présentation de cet avertissement permettrait d'impacter cet effet d'attente et d'allouer davantage de ressources à des processus contrôlés. Cela peut se constater car lorsqu'ils sont présentés face à un avertissement, les individus mettent plus de temps à traiter une information que lorsque l'avertissement n'est pas présent (Schul, 1993).

#### **La répétition des rétractations**

Les rétractations d'une fausse information peuvent être plus efficaces si ces dernières sont répétées, bien que les effets de l'exposition première à la fake-news reste quand même présent après la répétition des rétractations (Ecker, Lewandowsky, Swire, et al., 2011). Toutefois, l'effet d'une fausse information présentée une seule fois est le même que l'information ait été contredite une fois ou trois fois. Ainsi, même un encodage superficiel d'une fausse information sera complexe à manipuler, voire à révoquer de la mémoire. Les facteurs sous-tendant cet effet restent quant à eux encore complexes et mal compris. Quoi qu'il en soit, la répétition de la rétractation d'une fausse information reste un des leviers actionnables pour minimiser l'impact des fake-news.

### **Combler les trous de la narration**

Comme nous l'avons vu plus tôt, le fait de contredire une information peut créer un vide dans le récit cohérent qu'un individu se sera fait de l'histoire en question. La cohérence interne jouant un rôle primordial dans l'évaluation de la véracité d'une histoire (Pennington & Hastie, 1992; Schwarz et al., 2016), ce vide créé par la rétractation d'une information peut augmenter la motivation d'un individu à croire en une information fausse ou déformée, mais qui maintiendra la cohérence de la narrative. Des études ont montré que fournir une alternative aux trous narratifs d'une rétractation était une stratégie efficace pour effacer les effets d'une fausse information (H. M. Johnson & Seifert, 1994; Tenney et al., 2009). Par exemple, dire que « *aucun baril d'essence ou de peinture n'ont été trouvés, mais des preuves d'effraction ont été observées* » permet d'effacer l'effet d'influence continu. Toutefois, l'explication alternative doit alors être plausible, déterminante et revenir sur la raison de la suspicion de la fausse information en premier lieu.

En résumé, l'effet d'influence continu de la désinformation peut-être grandement atténué en combinant l'avertissement sur la qualité de l'information, la répétition de cet avertissement ou de la rétractation des fausses informations, et en fournissant des explications alternatives sur l'information en question. Nous pensons qu'en développant un algorithme de détection des fake-news, nous pourrions agir sur les deux premiers leviers. Le dernier reste quant à lui plus complexe à automatiser, car il implique la vérification d'une histoire, l'identification des éléments trompeurs et le remplacement de ce ou ces éléments par des éléments exacts, convaincants et probants. La section suivante visera à établir l'état de l'art des recherches sur l'automatisation de la détection des fake-news.

## **1.3 – L'état de l'art des algorithmes de détection des fake news**

On définit la détection automatique des fake-news comme le processus de catégorisation des informations selon un continuum de véracité, avec une mesure associée de certitude. Avec l'accessibilité grandissante de machines dotées d'une puissance de calcul importante pour des prix désormais abordables, la qualité des algorithmes de machine learning permet désormais d'obtenir de bons résultats sur des tâches de classification d'images, de détection de voix, et de traitement du langage. Depuis les événements comme l'élection de Donald J. Trump aux USA ou la vote en faveur du Brexit, la détection automatique de fake news est devenue un enjeu important de la recherche et plusieurs algorithmes sont désormais développés pour prévenir la publication de contenu faux.

### 1.3.1 - Aux origines : La détection de spam

La détection d'emails de spam fut une des premières applications des algorithmes automatiques de détection de contenu frauduleux. Ces algorithmes utilisent des techniques de machine learning permettant de classifier du texte comme étant du spam ou du contenu légitime. Ils impliquent le pré-traitement du texte, l'extraction de caractéristiques (souvent appelées *features*) via des méthodes comme les sacs de mots, et la sélection de ces features basée sur ceux menant aux meilleures performances dans la performance de l'algorithme sur un jeu d'entraînement. Ces features sont ensuite classifiés en utilisant divers classifieurs comme les K Nearest Neighbours (KNN), les Support Vector Machines (SVM) ou les Naive Bayes Classifiers.

Comme pour la détection de fake news, le but de ces algorithmes est de séparer des exemples de textes véridiques d'exemples de textes fallacieux. Parmi les différents algorithmes, le classifieur naïf de Bayes obtient de bons résultats (Androutsopoulos et al., 2000), bien que sa difficulté soit de s'adapter à de nouveaux types de spams (Sasaki & Shinnou, 2005), pour lesquels l'utilisation des KNNs permettait une adaptation rapide du modèle en obtenant des résultats similaires à ceux obtenus avec des algorithmes robustes en traitement du langage comme les SVMs. Toutefois, plusieurs des études publiées rapportant l'efficacité des algorithmes de détection de spam ne rapportaient que les résultats sans donner d'indications sur le contenu permettant aux algorithmes de classifier un mail comme étant un spam ou non (eg. Androutsopoulos et al., 2000 ; Debarr & Wechsler, 2009 ; Sasaki & Shinnou, 2005).

On pourrait ainsi se dire que le spam ayant du contenu fallacieux, ces algorithmes permettraient d'aider à la détection de fake-news, car ce dernier reste un problème de traitement automatique de langage caractéristique d'une intention de duperie. Un des épisodes de fake-news avant l'arrivée de « l'ère de la post-vérité » avait même été qualifié de « *spam social* » (Waldrop, 2017). Toutefois, la différence fondamentale concerne le fait que le spam soit un contenu non sollicité et générant davantage d'agacement que d'attrait. A la différence, les fake-news sont un contenu dont la valeur émotionnelle, l'aspect et la proximité avec l'actualité sociétale en font un contenu vers lequel les gens seront attirés. De plus, le spam est souvent en lien avec des questions financières, visant souvent à obtenir les informations bancaires des victimes. A l'inverse, comme nous l'avons vu plus haut, les fake-news ont-elles un but manipulateur et trompeur qui peut à l'occasion amener vers des buts mercantiles (eg. *Achat de produits déguisés sous l'appellation de médecines alternatives*), mais pas uniquement.

### 1.3.2 - Déetecter les fake-news à l'ère des réseaux sociaux

De manière surprenante, bien que les fake-news aient plus d'une décennie, l'intérêt pour la détection automatique de fake-news s'est particulièrement développé depuis la campagne pour les présidentielles américaines, comme on peut le voir sur la figure 3. Cette période, qui a vu l'élection du président Trump ainsi que le vote en faveur du Brexit, a permis de mettre en lumière un simple fait : Celui du fait que les médias dits « *traditionnels* » tels que la télévision, la radio ou les journaux n'étaient désormais plus les seuls vecteurs d'information de la population, et que les réseaux sociaux étaient désormais un vecteur puissant de partage de l'information. Ces articles peuvent être originaires des réseaux sociaux eux-mêmes, mais ils peuvent aussi être le relai d'articles originaires de sites web officiels ou de blogs non soumis à une déontologie journalistique.

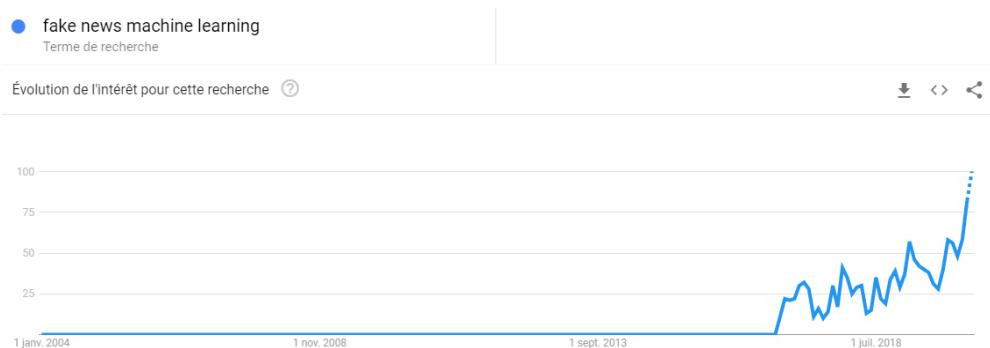


Figure 3 – Recherches Google sur les termes *fake-news* et *machine learning*

Toutefois, nous ne nous intéresserons pas à la propagation de ces contenus sur les réseaux sociaux, qui s'intéressent à d'autres enjeux que ceux de ce mémoire. L'analyse sur les réseaux sociaux permet en effet entre autres de détecter les individus à risque de croire et de partager de fausses informations, et d'analyser l'interprétation d'un individu d'un article en vérifiant dans le cas d'un retweet avec commentaire ou d'un partage Facebook avec commentaire l'adéquation entre le contenu rédigé par l'utilisateur, le titre de l'article et le contenu de l'article. Les principales compagnies de réseaux sociaux utilisent ainsi des systèmes de détection dont l'algorithme est basé sur le signalement interne par les utilisateurs d'un contenu fallacieux (Strickland, 2018). Ce système est efficace et a montré que la source était un élément majeur dans l'édition d'une fake-news, toutefois, il est exclusif à l'hôte d'un réseau social. Le centre d'intérêt de ce mémoire est lui de voir au sein de l'article partagé, quelles seront les features qui pourraient être les plus à même d'expliquer le fait qu'un individu redirigé sur cette page puisse croire au contenu présenté dessus.

### **1.3.3 - Les stratégies actuelles pour combattre les fake-news**

Il a été mentionné dans la partie précédente que Facebook, et les plateformes de réseaux sociaux en général, utilisaient une méthode de détection des fake-news semi-automatisée, car elle se base entre autres sur un input humain, le signalement, pour labelliser une information comme fausse ou non. Ce n'est toutefois pas la seule méthode employée, ni la plus efficace dans tous les contextes. Celle-ci reste adaptée à une plateforme comme Facebook qui dispose déjà d'un algorithme de recommandation de contenu et d'une base d'utilisateurs conséquente. Leur algorithme de détection doit donc prendre en compte ce premier biais dans le traitement de l'information pour adapter leur détection en conséquence. Deux autres approches peuvent-être décrite.

Tout d'abord, l'approche de l'intervention humaine et professionnelle, impliquant une vérification de l'information et sa correction via un article plus ou moins détaillé. On peut noter parmi ces initiatives l'International Fact Checking Network, qui permet de rapporter les articles en anglais et en allemand qui propageraient des fake-news. Plusieurs autres sites web de vérification de l'information ont eux aussi vu le jour de la part des médias eux-mêmes comme le Washington Post Fact Checker, les New York Times Fact Checks et même en France avec les Décodeurs du journal Le Monde ou le site Factuel de l'Agence France Presse. Ensuite viennent les interventions automatiques faites par les algorithmes basés principalement sur leur contenu, leur propagation et les algorithmes d'apprentissage continu.

Une analyse de la chronologie de la vérification des faits en France nous indique cependant que beaucoup de travail reste à faire pour le contenu en français<sup>11</sup>. En effet, chaque journal commence à monter depuis une dizaine d'années son équipe de vérification de l'information. Toutefois deux problèmes apparaissent de cette méthode. Si aucune automatisation n'est implantée, le temps qui sera pris pour détecter une information comme fausse ou vraie sera alors plus long, ce qui laisse davantage de marge aux fake-news pour se répandre. Ensuite, cela indique que le fact-checking en profondeur n'est pas systématique au sein d'une rédaction et qu'un journaliste peu coutumier de l'équipe de fact-checking peut ainsi lui-même rédiger un contenu qui n'aura pas été vérifié, que cela soit à dessein ou non.

De plus, le seul outil à disposition en France pour vérifier le contenu est le Décodex, dont la seule application d'automatisation se base uniquement sur l'URL d'un site qui aura été catégorisé manuellement comme propageant régulièrement de faux contenus, parfois de faux contenus ou comme étant un site parodique. Bien que cette méthodologie permette un tri rapide et efficace, elle sera aveugle à des articles propageant des fake-news mais sur un site fiable.

C'est pourquoi il est important de tenter d'extraire les informations internes à un contenu pouvant indiquer sa fiabilité.

#### 1.3.4 - L'extraction de features

En suivant tout ce que nous avons développé jusqu'ici, nous pouvons en déduire les features suivant à extraire pour classifier les fake-news :

- **Sources de l'article** : D'où viennent les sources de l'article ? Peuvent-elles être catégorisées comme fiable ? Y a-t-il un grand nombre de sources dans l'article ?
- **Titre** : Le résumé de l'article avec son éventuel sous-titre, dont le but est d'attirer le lecteur
- **Corps de texte** : Le contenu effectif de l'article

Les features seront extraites de ces 3 composants, et nous utiliserons prioritairement les features linguistiques pour le développement de l'algorithme. On pourrait rajouter les features visuelles en ajoutant les illustrations en image ou en vidéo des articles, mais pour des raisons matérielles et temporelles, ces dernières ne pourront être étudiées.

Concernant les features linguistiques, comme nous l'avons vu, le but des fake-news est d'attirer et d'alerter le lecteur. Dans ce but, on peut supposer que le langage employé soit différent de celui des sources vérifiées, que l'on peut attendre plus factuel et moins émotionnel. Plusieurs autres features peuvent être extraits des informations textuelles, regroupées comme suit par Reis et al. (2019) :

- 1) **Features linguistiques** : Caractéristiques pouvant être extraites au niveau des phrases. Ces dernières regroupent les mots ou groupes de mots les plus souvent employés ensemble (comme des expressions par exemple)
- 2) **Features lexicales** : Ces dernières concernent des caractéristiques comme le nombre de mots uniques, la fréquence de chaque mot dans le texte, le nombre de référence à soi ou à autrui, etc.
- 3) **Features psycholinguistiques** : Ces caractéristiques concernent le caractère psychologique de la langue. Cela implique par exemple leur valence ou leur intensité émotionnelle.
- 4) **Features sémantiques** : Les features sémantiques renvoient au sens des mots utilisés.

Toutefois, bien que de nombreux algorithmes de détection des fake-news aient été développés à partir des caractéristiques linguistiques (eg. Oshikawa et al., 2018; Rubin et al., 2016), le travail mené par Ruchansky et al. (2017) a permis de montrer que la combinaison de

plusieurs facteurs restait la méthode la plus efficace pour détecter des fake-news. Travaillant sur les données de réseaux sociaux, les chercheurs ont pu intégrer à leur modèle le pattern de propagation des articles sur les réseaux sociaux. N'ayant pas ces données sous la main, nous garderons à l'esprit l'importance de ne pas conserver uniquement les caractéristiques linguistiques pour pouvoir développer notre algorithme, et c'est pourquoi nous chercherons à exploiter au maximum les données disponibles sur un article en y incorporant les sources et les titres, afin de voir si ces derniers pourraient avoir un impact dans la détection de fake-news, ce travail n'ayant à notre connaissance pas encore été mené dans la littérature. De plus, la plupart des études menées l'ont été sur la propagation des fake-news sur les réseaux sociaux, et non sur l'analyse des articles de fake-news eux-mêmes.

## **2 - Présentation et acquisition des données**

### **2.1 – La base de données du Décodex**

#### **2.1.1 - « Un premier pas vers la vérification de masse de l'information »**

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<sup>12</sup>. 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. Toutefois, ce travail restant de l'investigation en profondeur ne permet pas nécessairement de se maintenir au contact de l'actualité, ce pourquoi le Décodex fut développé. Le but affiché par les auteurs est « d'offrir à chaque internaute les moyens de repérer les plus évidentes d'entre elles [i.e. *les fake-news*], et d'être averti lorsqu'il consulte un site connu pour diffuser de fausses informations ».

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 ».

Le choix de cette base pour développer un algorithme provient de plusieurs facteurs. D'une part, le travail du Décodex est régit par une charte explicite et accessible<sup>13</sup>. L'ensemble de la méthodologie est transparent et leur modèle économique ainsi que les actionnaires responsables du financement de leur travail sont énoncés clairement, ce qui nous permet de considérer la classification des Décodeurs comme se rapprochant d'une information sinon objective, du moins aussi peu soumise aux biais énoncés plus haut que possible. Plusieurs critiques ont en effet été faites à l'encontre de cette base (Venturini et al., 2018), cependant, aucun autre travail de la sorte n'a jamais été réalisé et mis à libre disposition du public en France. Par exemple, d'après l'outil Google Scholar, 72 articles scientifiques, rapports et papiers ont déjà fait référence à la base de données du Décodex depuis 2017.

Le Décodex est ainsi un outil de choix pour obtenir des données grâce au travail méticuleux de l'équipe en charge de ce projet, mais aussi parce qu'il est un des rares existants en France. On pourra aussi citer Factuel, qui est le site de fact-checking de l'AFP et regorge

d'informations intéressantes pour décrypter l'information et consiste en une série d'articles approfondis vérifiant l'actualité. CheckNews, un moteur de recherche de question-réponses géré par les journalistes de Libération, 20 Minutes Fake-Off, une rubrique du journal 20 Minutes et HoaxBuster, qui est une plateforme collaborative de fact-checking. Toutefois, la plupart de ces ressources consistent en des articles de fond sur un sujet et même le moteur de recherche de Libération ne fait que renvoyer un internaute vers des articles de Libération. Parmi toutes ces ressources, aucune ne présente les caractéristiques de liberté de l'information ou de recherche d'automatisation du fact-checking que celles à la base des Décodeurs.

Le développement de l'automatisation du fact-checking est aujourd'hui en plein essor, et il est certain que plusieurs outils apparaîtront dans les prochaines années à venir. Toutefois, il est à noter que dans l'ensemble, la plupart des outils concernent uniquement les pays et articles anglophones (eg. Google Fact Checking), et que les outils utilisables en France restent encore très limités. C'est pourquoi nous avons décidé d'essayer d'exploiter au mieux les ressources mises à disposition par le Décodex afin d'avoir un outil d'automatisation à disposition pour les pages web francophones.

## 2.2.2 - Méthodologie : Téléchargement et mise en forme des données

Le script qui a été utilisé pour récupérer les données peut être retrouvé en Annexe 1. Nous en détaillerons la méthodologie ici. Tout d'abord, les données ont été captées depuis l'URL du site web du monde permettant de télécharger le fichier de données brutes au format JSON ([http://s1.lemde.fr/mmpub/data/decodelex/hoax/hoax\\_debunks.json](http://s1.lemde.fr/mmpub/data/decodelex/hoax/hoax_debunks.json)). Le fichier est divisé en deux niveaux. Le premier niveau correspond au périmètre d'intérêt des données. Le second niveau est détaillé juste après :

- *Debunks* : Comprend, pour chaque hoax : Un identifiant de hoax, un résumé du hoax, sa teneur en vérité, un résumé du debunk, un lien vers le debunk.
- *Hoaxes* : Comprend une suite de liens vers l'article de hoax d'origine, chacun associé à son identifiant de hoax. Ainsi un même hoax peut avoir plusieurs articles.

Ces deux informations ont tout d'abord été séparées dans deux tables différentes, correspondant aux hoaxes et aux debunks. Nous avons finalement joint ces deux tables par leur identifiant afin de se retrouver avec la table 1.

debunk_id		hoax_link	true_false
0	1	<a href="https://www.facebook.com/CorentinFNJ/posts/250...">https://www.facebook.com/CorentinFNJ/posts/250...</a>	FAUX
1	1	<a href="http://www.paulomouvementcitoyen.com/2017/02/u...">http://www.paulomouvementcitoyen.com/2017/02/u...</a>	FAUX
2	1	<a href="https://www.blog.sami-aldeeb.com/2017/02/06/pr...">https://www.blog.sami-aldeeb.com/2017/02/06/pr...</a>	FAUX
3	1	<a href="https://francaisdefrance.wordpress.com/2016/12...">https://francaisdefrance.wordpress.com/2016/12...</a>	FAUX
4	1	<a href="https://www.facebook.com/10212204315883498/pos...">https://www.facebook.com/10212204315883498/pos...</a>	FAUX
5	1	<a href="https://www.facebook.com/387841388254774/posts...">https://www.facebook.com/387841388254774/posts...</a>	FAUX
6	1	<a href="https://twitter.com/MONSTERLOVE696/status/9761...">https://twitter.com/MONSTERLOVE696/status/9761...</a>	FAUX
7	1	<a href="https://www.lemonde.fr/les-decodeurs/article/2...">https://www.lemonde.fr/les-decodeurs/article/2...</a>	VRAI
8	2	<a href="https://www.facebook.com/439281732799406/posts...">https://www.facebook.com/439281732799406/posts...</a>	FAUX
9	2	<a href="https://www.facebook.com/PorteTesCouilles2/pos...">https://www.facebook.com/PorteTesCouilles2/pos...</a>	FAUX

Table 1 – Extrait de la dataframe après nettoyage de la base du Décodex

La catégorisation créée par l'équipe du Décodex comportait cependant plusieurs types de classement d'une information. On retrouvait ainsi les catégories « FAUX », « CONTESTABLE », « DOUTEUX », « TROMPEUR », « C'est plus compliqué », « Trompeur » et « PRUDENCE ». Nous avons regroupé les catégories « Trompeur » et « DOUTEUX » sous « TROMPEUR » et « C'est plus compliqué », « Prudence » et « PRUDENCE » sous « CONTESTABLE ». A la sortie, nous avons donc 4 manières de catégoriser une information : « FAUX », « VRAI », « CONTESTABLE » et « TROMPEUR ». Tous les articles présents dans la partie *debunks* du document, et ayant fait l'objet d'une investigation plus approfondie ont ici été considérés comme vrais, dans le sens où la synthèse du travail de recherche que ces derniers ont demandé est fidèle aux sources des informations, et s'approche ainsi davantage de la réalité qu'un autre.

Nous avons ensuite analysé de manière plus fine nos données. Dans un premier temps, une vérification des doublons nous a permis de voir que seuls 9 articles parmi les 13450 articles présents dans la base de données, étaient répétés entre 2 et 4 fois. Nous avons ensuite cherché à voir quelles étaient les sources les plus fréquemment utilisées pour chaque catégorie. Ces résultats sont présentés dans la figure 4. Comme il existait 1186 sites web différents, nous avons conservé (pour des questions de lisibilité) les 10 premiers sites web utilisés, et concaténé les autres sources dans la catégorie « Autres ». Comme on peut le voir sur le graphique, 77.75% des sources utilisées par le Décodex pour trouver une fake-news étaient issues des principaux réseaux sociaux existant actuellement (Facebook en représente 71.27%, suivi par Twitter à 4,89%, le reste étant occupé par Youtube et Reddit).

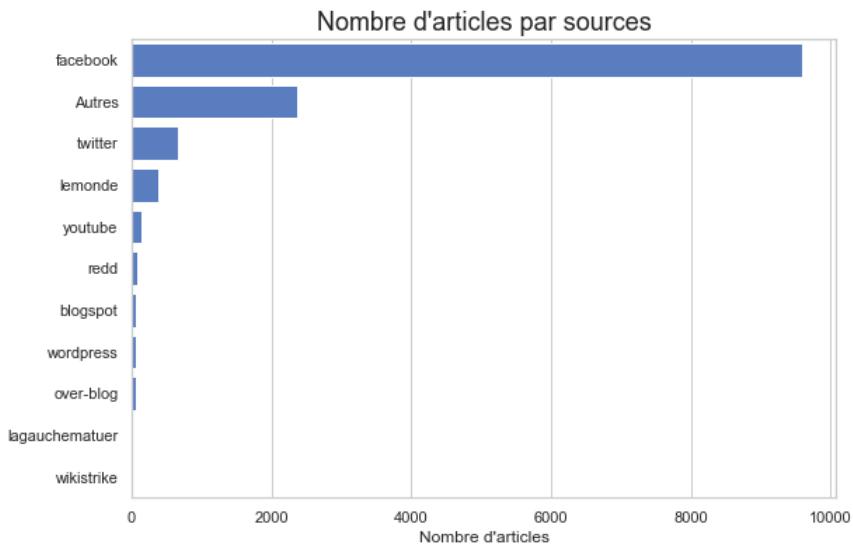
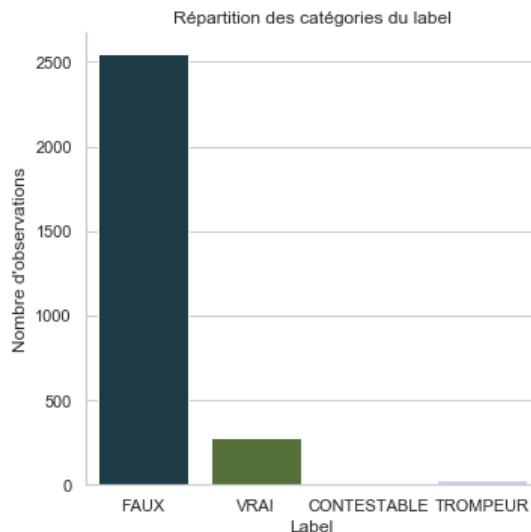


Figure 4 – Distribution des articles en fonction des sources

Les réseaux sociaux sont toutefois ici problématiques pour plusieurs raisons. Tout d'abord, ils ne font pas partie du but de cette analyse. La structure d'un article de blog ou d'un site web est fondamentalement différente de celle d'un post sur un réseau social, et ces derniers font désormais partie d'un objet d'étude et de problématiques à part entière, l'Analyse des Réseaux Sociaux. De plus, notre objectif est ici de créer un outil de web scraping afin de récolter des données. Cependant, faire ceci sur un réseau social est plus compliqué, ces derniers protégeant souvent leurs sites web contre le web scraping<sup>14</sup>, et la collecte de données sur les réseaux sociaux étant plus complexe à faire sur ces sites en raison de la potentialité de récolte de données à caractère privé. Nous détaillerons tout cela dans la partie suivante dédiée à la construction du web scraper et à la définition de son cadre légal. Bien que cela nous fasse perdre plus des trois quarts de nos observations (soit 10469), nous aurons toujours suffisamment de sites web à récupérer (2997 liens au total) pour pouvoir les exploiter dans notre algorithme. Enfin, nous avons voulu vérifier le déséquilibre entre les classes de notre label (notre catégorisation en « VRAI », « FAUX », « CONTESTABLE », « TROMPEUR »). Cette répartition peut être vue en figure 5. Il apparaît que les catégories « CONTESTABLE » et « TROMPEUR » ne pourront pas être suffisamment importantes pour avoir du sens en termes d'interprétabilité. Nous avons donc décidé de les enlever à notre analyse.



*Figure 5 – Déséquilibre de classes entre les différents types d'articles*

## 2.2 – Le web scraping et l’acquisition de données sur le web

### 2.2.1 - Le cadre légal du web scraping

Le web scraping est défini comme le processus d’utilisation d’outils technologiques à but d’extraire automatiquement et d’organiser les données du web, dans le but de les analyser ultérieurement (Krotov & Silva, 2018). Le principe est simple : Un script ou un programme accède à une 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.

De manière générale, le web scraping est une méthode existante à l’heure actuelle dans une zone juridique grise. En effet, la quantité de questions entourant cette pratique est assez vaste, car un web scraper ne fait rien qu’un humain ne saurait faire : Naviguer sur un site web. Il le fait simplement plus rapidement. De plus, les usages peuvent parfois être à visée de profits, mais aussi à but académique. Cela pose aussi la question de la propriété des informations affichées sur le web, car les données d’un site web sont dans l’absolu publiquement accessibles (à moins que ce dernier ne soit protégé par un abonnement). Dans un article du Bloomberg

Bureau of National Affairs<sup>15</sup>, deux avocats spécialistes dans le droit du numérique James Snell et Nicola Menaldo ont expliqué que leurs expertises et affaires leur avaient permis d'identifier 5 catégories de plaintes couramment soulevées sur le sujet du web scraping :

- Violation du droit d'auteur
- Violation de contrat, si le site web spécifie clairement dans son contrat d'utilisation interdire l'usage d'un web scraper
- Violation d'une loi spécifique au gouvernement fédéral américain (Computer Fraud and Abuse Act)
- Utilisation d'un web scraper dans le cadre de la collecte d'information pour envoyer des emails non sollicités
- Détournement d'information à but d'appropriation

C'est ainsi par exemple qu'une des affaires les plus connues en matière de web scraping opposait le site web LinkedIn à la société hiQ Labs (*hiQ Labs v. LinkedIn, no. 17-CS-03301-EMC, 2017 WL 3473663*). Dans cette affaire, la compagnie hiQ collectait des données utilisateurs disponibles publiquement sur le site LinkedIn, et dont leur business model était centré autour de l'analyse de ces données. LinkedIn avait envoyé une première lettre à hiQ leur demandant d'arrêter leur activité, suite à quoi hiQ les poursuivit en retour en leur déclarant ne pas avoir violé de lois, clamant n'avoir collecté que des données accessibles publiquement. La balance penchait davantage en faveur de hiQ dans l'affaire, la plainte de LinkedIn apparaissant suspecte dans cette histoire, car eux-mêmes traquaient les données de leurs utilisateurs et commençaient à restreindre de manière rétroactive l'accès aux données publiques en vue du développement de leur service « Recruteur », un produit compétitif à celui de hiQ. La compagnie hiQ finit par gagner le premier procès, la politique et les intentions de contrôle complet de LinkedIn sur leurs données utilisateurs étant considérées comme contraires à l'intérêt public.

En Europe, c'est en 2017 qu'un des cas les plus importants est apparu lorsque la Commission Européenne a dû trancher dans une affaire opposant les « Fintechs », ou entreprises numériques de services financiers aux banques. Les fintechs proposaient en effet différents services (gestion des budgets, paiements automatisés) à leurs utilisateurs contre l'acceptation d'un service tiers pour qu'ils puissent se connecter à leurs comptes. Même dans ce cas, le jugement de la commission ne fut pas unilatéral. Dans le cadre de la manipulation de données sensibles impliquant notamment l'utilisation de cartes de crédit, le webscraping était interdit. Toutefois, les banques devaient fournir en retour des API (*Application Programming Interface*) pour permettre un accès facilité aux données.

Dans l'ensemble, Krotov et Silva (2018) ont identifié 4 principes légaux généraux entourant le cadre légal du web scraping :

- **Conditions d'utilisation** : Les conditions d'utilisation d'un site web ne sont pas suffisantes pour déclarer qu'une violation de contrat a eu lieu dans le cas où un individu utiliserait un scraper sur le site lorsque les conditions du contrat l'interdit. Il faut en effet pour cela le consentement libre et éclairé du l'utilisateur ainsi qu'un accord explicite (en cliquant sur « J'accepte » par exemple) pour qu'il puisse y avoir violation de contrat, mais rien n'oblige un individu à les accepter.
- **Matériel sous droit d'auteur** : Scraper et republier des données ou informations explicitement possédées par le propriétaire d'un site web est considéré comme une violation de droit d'auteur.
- **But du web scraping** : Le scraping de contenu « premium » par exemple à des fins de revente ou de redistribution dudit contenu est prohibé par la loi.
- **Dégâts faits au site web** : Si le web scraping surcharge ou cause des dégâts à l'infrastructure hébergeant le site web ou au site web lui-même, le responsable peut alors être passible d'une amende. Cependant, le dommage doit être matériel et simple à démontrer.

Enfin, en France, le cadre légal du web scraping reste aussi flou qu'ailleurs, aucune loi n'entourant la pratique encore. Toutefois, et comme les autres pays européens, un des cadres législatifs les plus importants entourant cette pratique reste le Règlement Général pour la Protection des Données, qui interdit le traitement de données à caractère personnel de tout citoyen européen sans le consentement libre, explicite et éclairé de celui-ci, dans le cadre d'un traitement prédefinit et accepté par la Commission Nationale de l'Informatique et des Libertés.

En dehors de cela, la seule loi pouvant entourer cette pratique est l'Article 323-3 du code pénal, disant « *Le fait d'introduire frauduleusement des données dans un système de traitement automatisé, d'extraire, de détenir, de reproduire, de transmettre, de supprimer ou de modifier frauduleusement les données qu'il contient est puni de cinq ans d'emprisonnement et de 150 000€ d'amende* ». Toutefois cet article pourrait difficilement s'appliquer au web scraping qui extrait, plutôt que d'introduire ou de modifier des données, sans prendre en compte la notion de fraude. D'autant plus qu'un scraper fait le même travail qu'un simple navigateur web : Il se connecte à un serveur web pour récupérer les données d'un site et les traiter.

Dans l'ensemble, ces informations peuvent être résumées de la manière suivante : Il est difficile de dire que le web scraping n'est pas légal, encore plus dans le cadre de ce mémoire par exemple qui est réalisé dans un cadre académique. Toutefois, rien n'interdit les sites web à

rendre l'accès à leur site plus complexe en utilisant des CAPTCHA par exemple, ou en fournissant des services tiers comme une API et en interdisant le web scraping parce que cette API existe (c'est le cas pour l'API Twitter Developer par exemple).

### 2.2.2 - Le cadre éthique du web scraping

Le cadre légal entourant la pratique du web scraping est donc assez flou, bien que dans un périmètre définit notamment par le cadre du RGPD. Le cadre éthique du web scraping est quant à lui 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** : En combinant les données collectées d'un site web avec d'autres sources, un chercheur peut de manière non intentionnelle révéler l'identité de ceux à l'origine de la donnée.
- **Le respect de la confidentialité des organisations et du secret des affaires** : Les organisations ont tout autant le droit à la confidentialité de leurs affaires que les individus. Par exemple, on peut estimer sans en avoir l'intention le revenu d'un site web ou d'une organisation en récoltant les données d'un site proposant des annonces d'emploi, ce qui peut entacher la réputation de cette entreprise par la suite.
- **Diminuer la valeur produite par une organisation** : Les résultats de l'analyse d'un web scraper peuvent par exemple, générer une baisse d'intérêt pour les organisations à l'origine des données publiées.

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 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.

C'est dans le respect de ces buts que nous n'avons souhaité travailler que sur des éléments comme les titres, les articles et les sources, sans inclure d'autres données comme les méthodes analytiques d'un site web ou ses espaces publicitaires. De plus, nous ferons attention ici à ne présenter que des données agrégées et ne permettant pas de cibler une personne en

particulier. Enfin, si ces données étaient amenées à être mises à disposition en Open Source, nous nous assurerons que cela ne risque de causer de tort ni aux auteurs des sites web, ni aux créateurs du Décodex.

### **2.2.3 - Acquisition des données : Développement du Web Scraper**

Le web scraper a été développé avec Python 3.7.4 et le script d'exécution peut être trouvé en Annexe 2. Nous avons tout d'abord importé 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 ont ensuite été créées, dans lesquelles ont été insérées les données récoltées. L'algorithme suivait la logique suivante :

- Importer les données du Décodex pré traitées dans une dataframe.
- Déclarer le moteur utilisé pour le web scraping (dans notre cas, Chromium 80.0 en combinaison avec le navigateur Brave 1.5)
- Déclarer une variable égale à l'index de la dataframe le plus petit ne contenant pas de données web scrapées (row\_number). 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.
- 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

### **2.2.4 - Nettoyage des données issues du web scraper**

Suite au webscraping, des erreurs sont apparues 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. Le script de nettoyage des données peut être trouvé en annexe 3. Dans un premier

temps, toutes les observations contenant une liste de mots couramment rencontrés face à des erreurs de serveur (erreur 404, erreur 503) ont été supprimés, ce qui représentait 550 observations sur les 2816 originales. Toutes les observations tombées en erreur pendant le webscraping et ayant renvoyé un objet de type `NoneType` ont aussi été supprimées (12 observations). Enfin, les observations où les trois colonnes accueillant les données issues du web scraping étaient vides ont été filtrées. Au final, 692 erreurs de webscraping ont été filtrées, laissant 2183 observations.

Ensuite, les articles en langues autres que le français ont été filtrés. Afin d'éviter toute perte non nécessaire d'information, ce filtre a cependant été réalisé à la main. Cela a permis d'enlever les erreurs restantes non détectées dans la partie suivante (page non scrapée par la présence d'un CAPTCHA, page de connexion, ...). Les thèmes ainsi filtrés concernaient principalement une rumeur autour de la santé d'Hillary Clinton pendant les élections américaines ou la dangerosité du port du maillot du Football Club Barcelona aux Emirats Arabes Unis. Deux passes ont été ainsi réalisées pour éviter toute erreur d'inattention et 268 observations ont été ainsi supprimées. La dataframe après nettoyage comporte 1856 observations. La répartition entre les catégories Vrai et Faux est présente en figure 6. Un pattern régulier a aussi été trouvé dans le corps de texte retrouvé « *Amazon va fermer tous ses sites en\xa0France pour cinq jours Amazon :* » qui a été supprimé.

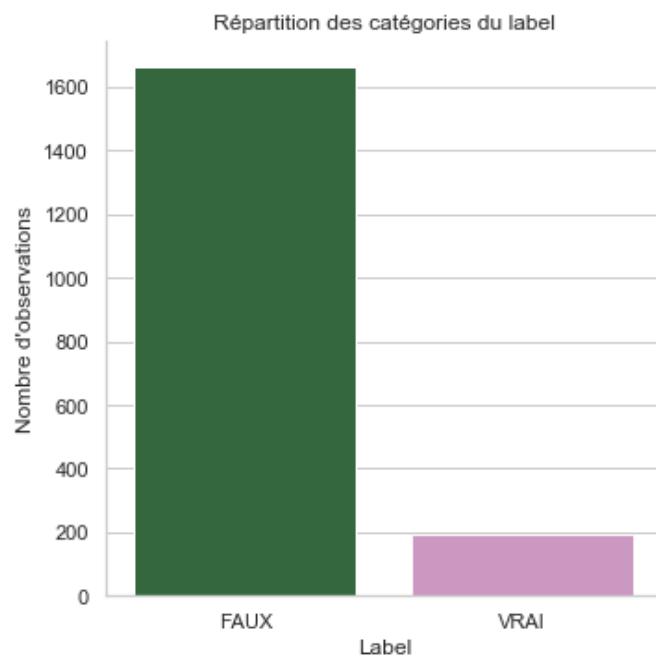


Figure 6 – Représentation du déséquilibre de classes après nettoyage des données

## 3 – L’algorithme de classification de fake-news

Plusieurs méthodes et challenges apparaissent pour classifier les données ci-dessus.

Dans l’ensemble, le workflow suivant a été suivi :

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

### 3.1 – Nettoyer les données acquises

Les colonnes correspondant au titre et au corps du texte auront le même type de traitement de texte, lorsque la colonne correspondant aux sources subira comme seul nettoyage l’extraction du Domain Name System (DNS). Ce dernier permettra de prendre en compte le nom des URLs cités comme source afin de voir si les sources présentes dans un article peuvent modifier la précision de notre algorithme. Le script de nettoyage des données est présent en annexe 4. L’algorithme de nettoyage du titre et du texte fonctionne comme suit :

- Convertir toutes les chaînes de caractère en minuscule
- Convertir le symbole « % » en mot « pourcents »

- Supprimer tout ce qui ressemble à un URL. Ceci a déjà été fait pendant la phase de webscraping, mais cette étape a été réitérée pour prévenir toute erreur qui aurait pu avoir lieu dans notre code.
- Suppression de tous les caractères spéciaux
- Supprimer les chiffres isolés par deux espaces
- Remplacer tous les espaces multiples par un simple espace
- Supprimer les espaces en début et en fin de chaîne
- Concaténer les colonnes titre, corps et sources en une seule

La partie suivante consistait d'une part à lemmatiser le texte, c'est-à-dire à convertir chaque mot à sa racine (« *mangées* », « *mangé* » et « *mangent* » deviennent « *manger* »), puis à enlever les stopwords. Ces derniers sont une liste de mots sans signification particulière comme « le », « la », « les », « un », ... Qui sont toutefois utilisés comme indices dans le processus de lemmatisation, raison pour laquelle cette étape a été ajoutée par la suite. Tous les stopwords sont issus de la librairie `stop_words` qui, après un comparatif des différentes librairies de Python de stopwords était la plus exhaustive. Elle proposait toutefois certains mots à enlever qui ont été conservés (« bon », « comment », « juste », ...), car leur catégorisation en stopwords semblait peu pertinente dans le cadre de cette analyse. En effet, comme nous analysons des formes syntaxiques qui pourraient être caractéristiques des fake-news, il fallait conserver le maximum de mots ayant un sens, même s'ils sont ambigus comme « bon », qui peut n'avoir aucun sens particulier (« Bon, ce n'était pas la première fois ») comme signifier la qualité « bon ». Ont aussi été filtrées toutes les observations dont la longueur de l'article était inférieure à 15 mots après suppression des stop-words, ce nombre ayant été choisi d'après une étude évaluant la longueur moyenne des phrases chez des blogueurs en fonction de leur âge et du genre (Goswami et al., 2009). La distribution des mots par article peut être retrouvée en figure 7. De tels articles peuvent en effet être davantage qualifiés comme des phrases ou des commentaires davantage que des articles, et cela nous a ainsi permis de supprimer 140 observations supplémentaires que l'on pourrait qualifier de peu pertinentes.

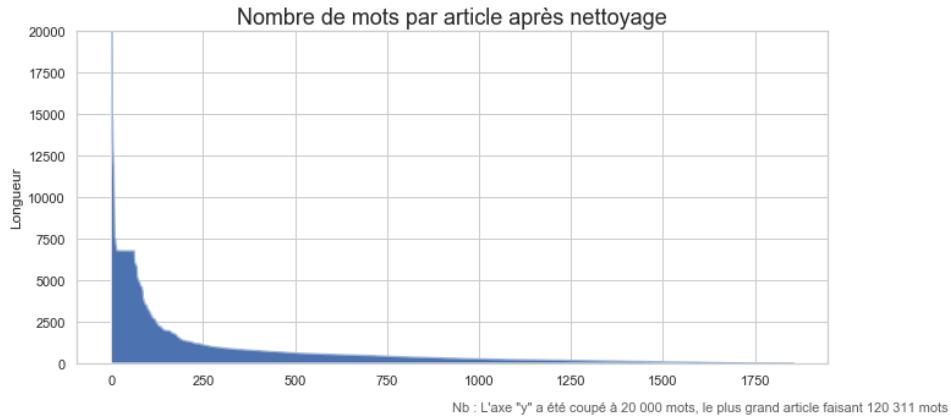


Figure 7 – Distribution du nombre de mots par article

Finalement, une première représentation visuelle des articles peut être retrouvée en figure 8. Ce regroupement permet d'observer déjà certaines différences entre les deux catégories d'information. Il semblerait en effet que les termes « *faux* » et les verbes d'action (« *agir* », « *vouloir* », « *pouvoir* ») fassent davantage partie du champ lexical des fake-news. Ces informations seront aussi utiles pour essayer d'améliorer la performance de notre algorithme plus tard en filtrant certains mots qui pourraient être fréquents mais peu utiles, comme par exemple le verbe *être*, qui est le plus présent dans chaque catégorie d'article. Toutefois, une première passe sera faite avant de nettoyer à nouveau nos données.



Figure 8 – Nuage de mots pour chaque catégorie d'article

## 3.2 – L'algorithme de détection des fake-news

### 3.2.1 - Le problème du déséquilibre de classes

Comme nous l'avons vu précédemment, un des problèmes majeurs rencontré avec le jeu de données obtenu est le sous-échantillonnage de la classe « vraie information », avec une répartition d'environ 90% d'observations correspondant à la classe « fake-news » (1534), et 10% à la classe « vraie information » (182). Les algorithmes réagissent différemment aux problèmes de déséquilibre de classes. Par exemple, le Linear SVC, qui est une forme de

Support-Vector Machine, gère assez bien les déséquilibres de classes avec l'hyperparamètre  $C$ . Cet hyperparamètre détermine la pénalité liée à la mauvaise classification d'une observation en assignant un poids à chaque classe de sorte que  $C_j = C * w_j$ , où  $C$  correspond à la valeur de la pénalité de mal-classification,  $w_j$  qui est un poids inversement proportionnel à la fréquence de la classe  $j$  et  $C_j$  est la valeur de  $C$  pour la classe  $j$ . Ainsi, plus une classe est fréquente, moins son poids sera important. A l'inverse, un algorithme comme la régression logistique aura beaucoup plus de difficultés à gérer ces problèmes (King & Zeng, 2001) et devrait être évité lorsque de tels problèmes apparaissent.

La prise en compte du déséquilibre de classe entre aussi en jeu dans le choix de la métrique qui sera utilisée pour analyser l'efficacité du modèle. Les métriques couramment employées pour évaluer un modèle de machine learning sont :

- **Accuracy** (exactitude) : L'exactitude nous permet de savoir parmi les observations catégorisées comme positives (ou faisant partie d'une classe  $k$ ), combien ont été correctement catégorisées. Elle se calcule de la manière suivante :

$$Accuracy = \frac{VraisPositifs}{VraisPositifs + FauxPositifs}$$

- **Recall** (rappel) : Le rappel nous permet d'évaluer le nombre d'éléments de notre classe  $k$  reconnus parmi ceux ayant été correctement reconnus et ceux ayant été mal labellisés. Autrement dit, de toutes nos données issues de la classe  $k$ , combien ont été renvoyées par notre algorithme ? Il se calcule ainsi :

$$Recall = \frac{VraisPositifs}{VraisPositifs + FauxNégatifs}$$

- **F1-Score** : Le score F1 est un juste milieu entre la précision et le rappel. C'est une moyenne harmonique pondérée dont le score va de 0 (mauvais score) à 1 (bon score).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Cette mesure peut être balancée en 2 autres mesures, la F0.5, qui met davantage de poids sur la précision, moins sur le rappel, et la F2 qui en met davantage sur le rappel.

- **ROC AUC** : L'aire sous la courbe ROC est une représentation visuelle du compromis entre taux de vrais positifs (en y) et taux de faux positifs (en x). Si la classification d'un modèle est aléatoire, les courbes de probabilité pour chaque catégorie suivent alors les mêmes valeurs, et la courbe ROC résultante sera donc une droite parfaite. Si notre modèle distingue correctement vrais positifs et faux positifs, le taux de vrais positifs tendra alors au plus vite vers 1 au plus qu'augmente le taux de faux positif. La courbe

sera alors similaire à une courbe d'apprentissage. Cette dernière est à proscrire en cas de déséquilibre de classe (Saito & Rehmsmeier, 2015).

- **PR-AUC** : La courbe PR-AUC combine la visualisation de la précision et du rappel en une seule visualisation. Cela permet de savoir à partir de quel taux de rappel (en x) notre précision (en y) chute. Cette courbe est davantage informative dans le cas de données extrêmement déséquilibrées (Saito & Rehmsmeier, 2015).

En utilisant un arbre de décision disponible en ligne<sup>16</sup>, les métriques auxquelles nous devrions prêter attention dans l'analyse de la qualité de notre modèle sont l'accuracy et le F1-Score. Nous préterons cependant davantage d'attention au score F1, qui peut se révéler moins sujet aux problèmes de déséquilibre de classe que l'exactitude. En effet, même sans modèle, notre répartition en 90-10 peut provoquer des cas comme celui où un modèle aléatoire peut mieux prédire qu'un modèle entraîné, car ce dernier aura alors une performance de 90% sur la classe dominante, cet effet est aussi appelé « Accuracy Paradox ». C'est pourquoi nous utiliserons aussi la précision, qui correspond à la proximité (ou la variabilité) des erreurs des observations au sein de chaque classe, ainsi que le recall pour voir le nombre de faux négatifs.

Enfin, les données vont aussi conditionner l'algorithme qui sera utilisé. Les développeurs de la librairie de machine learning Sci-Kit learn ont ainsi mis en ligne un arbre de décision<sup>17</sup> sur lequel nous nous appuierons pour sélectionner les algorithmes appropriés à notre analyse. Notre échantillon comporte ainsi plus de 50 observations. Nous souhaitons prédire une catégorie à partir de données labellisées, toutefois nous disposons de moins de 100 000 observations. L'arbre de décision nous indique ainsi que les algorithmes les mieux adaptés sont les suivants : Linear SVC, puis Naïve Bayes si le premier n'est pas concluant.

Finalement, nous essaierons aussi de régler le problème du déséquilibre de classe avec d'autres outils que la sélection des outils les mieux adaptés. En plus de l'acquisition de données supplémentaires (qui ne sera pas faite ici, car l'expertise demandée pour trouver un article factuel ne fait pas partie du spectre de ce travail), il existe ainsi d'autres méthodes permettant de gérer différemment ces problèmes, couramment nommées techniques de rééchantillonnage :

- **SMOTE** (*Synthetic Minority Over-sampling Technique*) : Conçue par Chawla et al. (2002), cette méthode consiste à suréchantillonner chaque classe afin que la ou les classes en minorité soient équivalentes à la classe majoritaire en utilisant un algorithme de plus proches voisins pour générer des données synthétiques.
- **Nearmiss** : Cette méthode peut être considérée comme l'opposé de la technique SMOTE. Chaque classe majoritaire est ici sous-échantillonnée en utilisant les k plus

proches voisins pour correspondre à la taille de l'échantillon minoritaire (Mani & Zhang, 2003).

Ces deux approches seront testées afin de voir si elles contribueront à améliorer notre algorithme ou non.

### 3.2.2 - Préparation du pipeline de l'algorithme

Comme cela a été précédemment développé, deux algorithmes principaux seront utilisés : Le Linear SVC puis le Naïve Bayes. Nous comparerons les résultats obtenus avec ces deux algorithmes lors de l'application de la méthode Nearmiss pour le sous échantillonnage et SMOTE pour le suréchantillonnage. Le script utilisé est décrit en Annexe 5. Pour exécuter l'algorithme, le pipeline suivant a été appliqué :

- **Vectorisation du texte** : L'application d'un algorithme quel qu'il soit implique le traitement des données sous forme numérique. Dans le cadre de l'analyse de texte, cette numérisation consiste à transformer un texte en un vecteur de nombres, l'approche couramment utilisée est appelée « bag of words ». Les observations ne changent pas dans cette opération et se rapportent au même texte, mais les colonnes deviennent alors des features où chaque feature se rapporte à un mot et contient la fréquence ou le compte d'apparition de chaque mot. Les deux fonctions les plus couramment employées sont le CountVectorizer et le TfidfVectorizer.

Le TfidfVectorizer a l'avantage de donner une représentation normalisée de la fréquence totale d'apparition de chaque mot en calculant la fréquence dans le document inverse. Son intérêt est donc prononcé lorsqu'on considère qu'un mot fréquent puisse ne pas fournir de véritable gain d'information, et lorsque la taille du vocabulaire augmente. C'est pourquoi nous utiliserons ici le Tf-Idf, car la taille de vocabulaire entre chaque classe est différente, alors que nous souhaitons avoir une représentation normalisée de nos données dans chaque classe.

- **Sélection des features** : Comme chaque features sont faites à partir de mots, la tâche de classification peut avoir lieu sur un grand nombre de features (de la taille du vocabulaire totale). Il faut ainsi un moyen de réduire le nombre de dimensions afin d'améliorer les capacités computationnelles de l'algorithme et d'éviter de faire du surapprentissage. Les méthodes de sélection de features sont nombreuses et font l'objet d'intenses recherches au sein de la communauté académique (Alelyani et al., 2017). On recense 4 méthodes principales de sélection de features. Les *filtres* réalisent des analyses statistiques sur l'espace des features pour déterminer des sous-ensembles, les *wrappers* sélectionnent plusieurs sous-ensembles puis les classifient, l'approche *embedded* inclut la phase de sélection pendant l'entraînement de la

classification et l'approche *hybride* cherche à combiner les avantages des filtres et des wrappers. En se basant sur l'article d'Alelyani et al. (2017), nous avons décidé d'utiliser le Chi<sup>2</sup> comme filtre de features.

- **Application du classifieur :** Comme nous l'avons discuté dans la partie précédente, nous essaierons deux classifieurs : Linear SVC et Naïve Bayes.

### 3.2.3 - Exécution du pipeline de l'algorithme

Le premier pipeline consistait donc à appliquer le TfidfVectorizer, pour lequel les unigrammes et les bigrammes étaient extraits, puis à sélectionner les K meilleures features avant d'appliquer le LinearSVC. Toutefois, il manquait un élément déterminant dans les hyperparamètres du modèle : La valeur du K. Le premier pipeline a donc été lancé dans une boucle, sans précision des hyperparamètres, afin de déterminer la valeur la mieux adaptée au K. Les résultats sont visibles en figure 9.

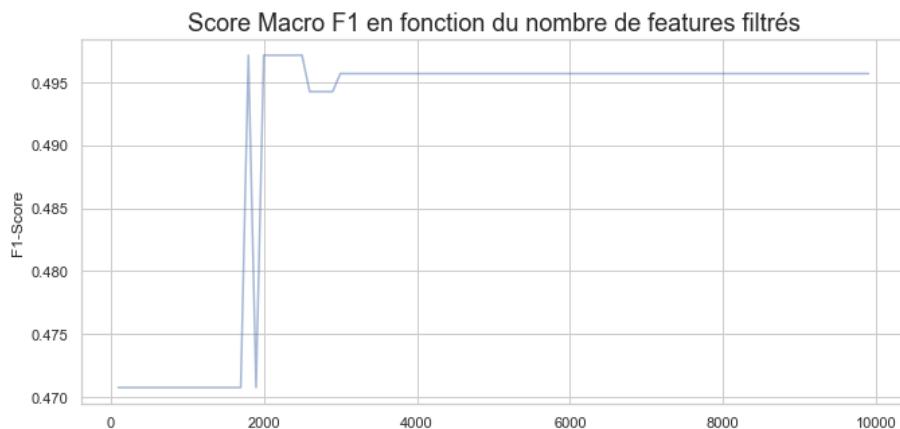


Figure 9 – F1 Score (macro) en fonction du nombre de features retenus

Afin de l'évaluer, le score F1 a été utilisé. La bibliothèque sklearn propose cependant 3 différents scores F1 :

- **Weighted F1 :** Ce score est issu du calcul du score F1 de chaque classe indépendamment, auxquels sont assignés un poids dépendant du nombre de labels correctement classifiés pour chaque classe :  $F1_{weighted} = F1_{vrai} * Weight_{vrai} + F1_{faux} * Weight_{faux}$ . Ce score aura donc tendance à favoriser les classes majoritaires.
- **Micro F1 :** Ce score utilise le nombre global de vrais positifs, faux négatifs et faux positifs pour calculer le score F1 directement :  $F1_{class1+class2}$ . Ce score est intéressant

car il ne favorise aucune classe en particulier. Il est équivalent à l'accuracy du modèle de base.

- **Macro F1** : Ce score additionne les scores F1 séparément pour chaque classe mais sans utiliser de poids, résultant en une plus grande pénalisation lorsque le modèle performe mal avec la classe minoritaire.  $F1_{class1} + F1_{class2}$

Nous éviterons ici d'utiliser les Micro et Weighted F1 pour évaluer la qualité de notre modèle à cause du déséquilibre de classe et lui préférerons le Macro F1. L'utilisation de cette métrique permet d'évaluer la performance du modèle en fonction de sa performance pour chaque article. Le but principal ici est d'avoir un modèle qui puisse reconnaître de manière adéquate les faux articles et les vrais articles, quitte à ce qu'il ne puisse pas rappeler tous les articles labellisés vrais. Le but de cet algorithme est en effet d'identifier avec précision un article faux, quitte à suridentifier un article vrai comme étant faux. Un excès de méfiance dans le cadre de la détection de fake news est en effet à préférer à un excès de confiance (et donc de faux négatifs).

D'après le graphique précédent, le nombre K de features optimal à sélectionner est de 2000 features. Le score F1 montre quant à lui que l'on pourrait considérer le modèle comme peu informatif, car son score de classification n'est plus plus important que la chance (49,7%). Toutefois, les métriques du modèle présentées en table 2 montrent que ce dernier pourrait être plus informatif que si l'on se fiait à un seul score.

	precision	recall	f1-score	support
FAUX	0.892128	1.000000	0.942989	306.000000
VRAI	1.000000	0.026316	0.051282	38.000000
accuracy	0.892442	0.892442	0.892442	0.892442
macro avg	0.946064	0.513158	0.497136	344.000000
weighted avg	0.904044	0.892442	0.844487	344.000000

Table 2 – Résultats du premier modèle LinearSVC sans sur/sous échantillonnage

Le premier modèle a ainsi correctement identifié les articles vrais et faux qu'il a remontés, mais il n'a remonté qu'un nombre très limité d'articles vrais. C'est donc un modèle qui a davantage tendance à classifier trop d'articles dans la catégorie « Faux ». Toutefois, lorsqu'il catégorise un article dans la catégorie « Vrai », il le fait correctement.

Le deuxième modèle créé visait à gérer le problème de déséquilibre de classes en utilisant une méthode de suréchantillonnage (SMOTE) et de sous-échantillonnage (Nearmiss) afin d'en comparer les performances. La méthode Nearmiss s'est cependant révélée peu utile

dans l'amélioration des performances du modèle et la méthode SMOTE a ainsi été retenue. Les résultats sont présentés en table 3. Cette méthode a permis d'améliorer la performance de rappel des articles vrais au détriment des articles faux. Cependant cet algorithme permet d'obtenir un modèle plus robuste avec un score F1 de 0.57. Une perte considérable de précision pour la catégorie « vrai » de ce modèle montre toutefois que l'algorithme a eu des difficultés à grouper les articles vrais dans la même classe. Les performances du modèle ont cependant pu être améliorées en ajustant la valeur des hyperparamètres  $C$ , le paramètre de régularisation qui permet une marge d'erreur légèrement plus importante dans le processus de classification, ainsi qu'en demandant à l'algorithme d'ajuster le poids de chaque classe par sa fréquence inverse. Le choix de cet hyperparamètre s'est fait sur la base de la figure 10.

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
FAUX	0.906355	0.885621	0.895868	306.00000
VRAI	0.222222	0.263158	0.240964	38.00000
accuracy	0.816860	0.816860	0.816860	0.81686
macro avg	0.564288	0.574389	0.568416	344.00000
<b>weighted avg</b>	<b>0.830782</b>	<b>0.816860</b>	<b>0.823524</b>	<b>344.00000</b>

Table 3 - Résultats du second modèle LinearSVC avec suréchantillonnage SMOTE

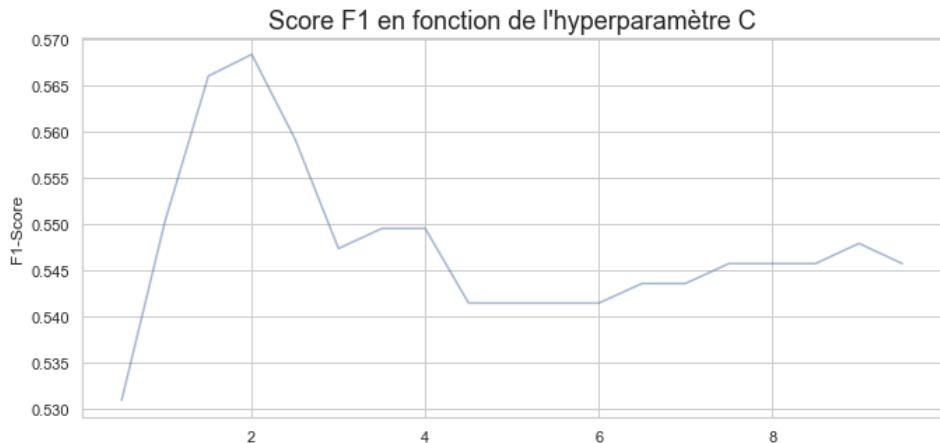


Figure 10 – Score F1 en fonction de l'hyperparamètre C

Le deuxième algorithme mis en place pour essayer d'améliorer les performances du modèle est l'utilisation d'un algorithme naïf de Bayes à la place du SVM couramment employé en classification de texte. Les résultats de cet algorithme présentés en table 4 n'ont toutefois pas permis de dépasser ceux du LinearSVC.

	precision	recall	f1-score	support
FAUX	0.906863	0.604575	0.725490	306.000000
VRAI	0.135714	0.500000	0.213483	38.000000
accuracy	0.593023	0.593023	0.593023	0.593023
macro avg	0.521289	0.552288	0.469487	344.000000
weighted avg	0.821678	0.593023	0.668931	344.000000

Table 4 – Résultats du modèle bayésien avec suréchantillonnage SMOTE

Un résumé des 3 modèles développés est disponible en table 5. On peut y voir que comme attendu, le premier modèle a classifié toutes les observations dans la catégorie majoritaire « FAUX », ce qui lui permet de jouir d'une importante précision, ce qui se fait au détriment du rappel des informations vraies. Le modèle numéro 2 avec suréchantillonnage avec la méthode SMOTE a quand a-t-il obtenu de meilleurs résultats, bien que seul un quart des articles vrais ont été catégorisés dans leur propre catégorie. On peut considérer ce modèle comme un « modèle trop prudent ». Il est toutefois pertinent d'accepter la classification dans cette catégorie lorsqu'on parle de fausses informations sur internet. Enfin, le dernier modèle utilisant l'algorithme de Bayes a quant à lui mieux performé sur la prédiction des articles vrais, 50% ayant été mal classifiés. Cependant, cela s'est fait au détriment de la qualité de la précision des articles faux qui est de 60%. Cela nous conforte dans l'idée que bien que le second modèle puisse être trop prudent, il arrive à retenir un grand nombre d'articles faux.

Linear SVC simple		Linear SVC SMOTE		Naïve Bayes	
predicted:FAUX	predicted:VRAI	predicted:FAUX	predicted:VRAI	predicted:FAUX	predicted:VRAI
actual:FAUX	306	0	271	35	185
actual:VRAI	37	1	28	10	19

Table 5 – Matrices de confusion de chaque modèle

## 4 – Conclusions et perspectives

### 4.1 – Classifier les fake-news

L’objectif initial de ce projet était de développer un algorithme de classification qui permettrait de détecter une fake-news en utilisant comme données les pages web identifiées comme contenant des fake news présentes dans la base développée par le Décodex. Ce travail cherchait à apporter une nouvelle piste de réflexion en proposant une solution automatisée de détection de fake-news sur des articles francophones. Le but était aussi de développer une réflexion transverse sur l’approche de la détection de fake-news, en questionnant la notion du vrai et du faux, ainsi qu’en rapportant ce phénomène à son impact social et psychologique. Finalement, nous souhaitions ouvrir la question des fake-news aux articles de site webs et de blogs et non aux seuls réseaux sociaux, qui concentrent une partie majeure de l’attention au sein de la communauté académique.

Du point de vue psycho-social, il apparaît que l’on peut conclure à certaines caractéristiques propres aux fake-news sans nécessairement avoir besoin de recourir au développement d’un algorithme en amont. Le besoin d’attirer l’attention et de rester en mémoire sont deux caractéristiques assez simples à reconnaître pour une fake-news. L’attention est en général davantage attirée par des stimuli à valence négative que par des stimuli neutres, et ces informations restent souvent bien plus longtemps en mémoire et de manière plus robustes que les informations factuelles (Goldsmith & Dhar, 2013). Le marché des médias étant régit par le besoin d’attirer et de retenir l’attention, cette tâche peut se retrouver plus complexe car ces derniers utiliseront aussi un vocabulaire négatif, impactant et favorisant un traitement heuristique des informations (Chaiken, 1980; Mase et al., 2015; Schwarz et al., 1991). Le but des propagateurs de fake-news ayant des similarités avec celui des médias (enjeux financiers), la seule distinction claire entre les deux types d’informations est relative à l’intention (Verstraete et al., 2017).

Il apparaît des résultats de ce travail que *ceteris paribus*, l’intention n’est pas aisément distinguable à partir du contenu produit par des manipulateurs. Cela va sans dire, les problèmes méthodologiques liés au développement de l’algorithme ont ajouté davantage de bruit à nos données, et il est ainsi impossible d’affirmer que l’on ne saurait distinguer un article de fake-news d’un article détaillé de manière automatique. La piste de recherche sur la détection des intentions reste cependant une voie ouverte pouvant s’avérer prometteuse.

Ce travail a toutefois permis de proposer des pistes pour mieux comprendre la prévalence et la persistance des fake-news. Plusieurs méthodes issues de la littérature

permettent aussi d'adopter une position plus critique et objective sur les meilleurs moyens d'éviter la propagation des fake-news.

Un des résultats les plus présents est l'importance de la prévention des fausses informations. Une fausse information faite pour proposer une narrative cohérente et marquante qui s'intègre rapidement en mémoire. Cependant, une fois mémorisée, la remplacer par une autre narrative devient alors une tâche particulièrement complexe. Il est donc important qu'il y ait une information de masse sur les effets des fake-news et de leur ténacité. Pour citer Lewandowsky et al. (2012): « *It is important for the general public to [...] understand...]* that people may ‘throw mud’ because they know it will ‘stick’ (tr. “Il est important que le public comprenne que certaines personnes jettent de la boue, car ils savent qu'elle tâchera quelqu'un”). Il est donc important que la prévention et l'éducation à l'esprit critique se développe, car les stratégies de correction de biais de croyance sont complexes et souvent sans effet. Ce pan de la culture est d'ailleurs rempli d'idées préconçues qui permettraient dans la croyance populaire d'éradiquer facilement une fausse information, telle que le format « mythes vs faits », toutefois la revue que nous avons fait sur cette littérature montre que les stratégies « de bon sens » restent inefficaces voire contreproductives la plupart du temps.

Du point de vue classification automatique, le travail requis est extrêmement complexe et nécessitera encore beaucoup de développements. Toutefois, il représente la première tentative de classification de contenu d'articles de fake-news en langue française, ce qui n'a à notre connaissance pas encore été réalisé. Une fois développé, et amélioré, une possibilité d'extension serait de récupérer le modèle créé, et de l'insérer dans une application qui pourrait scraper un site web, puis analyser son contenu pour le classifier ensuite en utilisant ce modèle. Le but serait ici de proposer une analyse à priori du contenu d'un article qui permettrait à un individu d'être averti d'être potentiellement exposé à un article trompeur avant, au moment ou juste après la lecture de son contenu.

L'approche transverse permet aussi d'ajouter des précisions relatives aux difficultés actuellement existantes dans le champ du Machine Learning appliqué à la détection de fake-news (Oshikawa et al., 2018). Il apparaît en effet que les études visant à classifier les fausses et vraies informations confondent souvent des termes relatifs aux fake-news, ce qui peut rendre l'analyse des résultats plus opaque. La question de l'évaluation de l'exactitude d'une information est en effet un problème récurrent, et bien souvent cette évaluation est relative uniquement à un fait d'actualité, alors que les fausses informations peuvent aussi concentrer des thèmes plus généraux. La dépendance à l'actualité des algorithmes de détection de fake-news est ainsi un jalon qu'il faut prendre en compte et tenter d'éviter dans les futures recherches

sur ce sujet. Enfin, l'un des buts de ce travail était aussi de souligner l'importance de la prise en compte de multiples informations dans la caractérisation automatique d'une fake-news. Ces dernières sont en effet toujours inscrites dans un contexte spatial, temporel, lexical et formel bien défini, et la prise en compte du seul texte peut se révéler insuffisante pour obtenir des résultats satisfaisants. Prendre en compte les sources d'un texte, son auteur ou sa date peut aider à améliorer la pertinence des algorithmes (Ruchansky et al., 2017). Ce travail n'a pas permis de satisfaire l'utilisation de toutes ces variables périphériques, mais visait aussi à rappeler leur importance.

Les résultats obtenus avec le présent travail permettent aussi de mettre en lumière qu'il est possible de développer une méthode de classification automatique des fausses informations à partir de données francophones, bien que ces dernières soient plus restreintes que les différents corpus anglophones à disposition. Une collaboration renforcée avec l'équipe de travail du Décodex pourrait ainsi être une direction à prendre afin de renforcer la qualité des données utilisées dans l'algorithme.

Un autre questionnement a aussi pu être soulevé quant aux buts de tels algorithmes. L'idéal est en effet d'avoir un modèle permettant de classifier les articles véridiques et non informatifs dans leurs catégories respectives. L'expérience montre qu'il faut parfois faire des choix sur la finalité du modèle, car ce dernier ne permettra pas nécessairement d'obtenir une classification parfaite. Une balance entre le nombre de faux articles considérés comme vrai doit donc être opérée avec le nombre de faux articles classifiés comme faux. En l'occurrence, le but d'un tel modèle, pour être cohérent avec la littérature sur le sujet, doit prioriser la précision de la classification des fake-news, même si cela implique que des articles véridiques doivent être considérés comme faux. L'importance de la prévention des fake-news ne saurait être trop soulignée, et le principe de précaution doit être placé avant toute autre considération. Notre modèle ne permet cependant pas de classifier suffisamment de bonnes informations (environ un quart), et c'est pourquoi nous ne pouvons pas le considérer comme informatif. Un principe de précaution trop important peut en effet avoir un effet délétère sur l'utilisateur qui pourrait alors considérer le modèle comme non pertinent et décider de revenir à une consommation normale d'informations sans faire attention à la pertinence et à l'exactitudes des « vérités » qui pourraient y être formulées. Une autre possibilité serait de l'utiliser sur une variété de sites web issus de cette même base afin d'assurer une « veille sur les fausses informations » afin d'être prévenu au plus tôt dès qu'un certain nombre d'articles paraissent sur un sujet dans plusieurs médias et contiennent des caractéristiques propres aux fake-news.

## 4.2 – Limites méthodologiques

Une des premières limites méthodologiques rencontrée dans ce travail concerne le déséquilibre de classe entre les articles scrapés vrais et faux. La répartition était d'environ 9 articles faux pour un article vrai, ce qui a mené l'algorithme au paradoxe d'exactitude (ou accuracy paradox). Ce paradoxe surgit lorsque l'algorithme reflète seulement la distribution de classe sous-jacente et fausse la qualité du modèle, car la plupart des algorithmes de machine learning supposent une distribution égale des observations au sein des différentes classes du label. Un biais de préférence surgit alors pour la classe majoritaire, ce qui impacte négativement la classe minoritaire. L'exactitude du modèle devient alors un indicateur à proscrire, car ce dernier se base sur la quantité de vrais positifs et de vrais négatifs prédits par l'algorithme. Ainsi, si les données avaient contenu 95 d'observations dans la classe faux articles pour 5 vrais articles, en classifiant toutes les données dans la catégorie faux il aurait eu une exactitude de  $\frac{95+0}{100} * 100 = 95\%$ . En utilisant des métriques différentes, une autre interprétation des résultats de chaque modèle a pu être présentée. Toutefois, une possibilité serait dans un premier temps de collecter davantage de données appartenant à la première classe.

Un autre problème inhérent aux données de la première classe concerne les articles qui ont été sélectionnés pour appartenir à la catégorie « VRAI » dans cette étude. En effet, dans la base de données du Décodex, les liens vers les articles présentant une version plus travaillée des rumeurs et fausses informations redirigeaient toujours, et de manière logique, vers des articles du monde. Même si les résultats du modèle avaient été différents et plus justes en termes de classification, ce dernier aurait tout de même dû être pris avec précaution, car il aurait alors été possible que l'algorithme ait réalisé une classification distinguant les articles faux des articles contenant des marques syntaxiques ou sémantiques propres aux rédacteurs des articles du Monde. De plus amples travaux devraient donc consister en une première phase de recherche et d'étiquetage des articles, avec une classification utilisant la méthode des juges afin d'éviter qu'un seul individu soit responsable de cette classification et que ses propres biais n'entrent en compte et faussent eux-mêmes les résultats sur les données d'entrée du modèle. Ce travail devrait aussi s'assurer de la diversité des sources, afin de ne pas réaliser de surapprentissage sur les données d'un seul site, comme cela a pu être le cas avec les articles issus de la catégorie « VRAI ».

Une fois ce travail préparatoire réalisé, il serait possible que les résultats issus d'un classifieur comme le Support Vector Machine ou le Naïve Bayes deviennent alors bien plus concluants. Le but serait ici d'améliorer la qualité des données en entrée avant de passer à la

phase algorithmique. Les méthodes de suréchantillonnage et de sous-échantillonnage utilisées ont tout de même été utiles et ont permis d'améliorer la qualité du modèle, mais la limite principale à l'efficacité d'un algorithme est et restera la qualité des données en entrée et du travail de pré-traitement qu'il y aura eu sur ces dernières. Un tel travail de classification pourrait aussi être très bénéfique à la communauté académique, qui pourrait alors s'appuyer sur de telles données pour développer un algorithme spécifique aux fake-news en français.

Une autre possibilité serait aussi de travailler sur les données autres que sémantiques. La librairie nltk dispose ainsi d'un outil permettant de récupérer plusieurs types d'informations sur un corpus de texte. Cette librairie de traitement du langage naturel permet bien plus qu'une liste de stop-words, un outil de découpe de mots ou un outil de lemmatisation, mais elle permet aussi de récupérer plusieurs informations qui pourraient s'avérer très utiles dans l'évaluation d'un article web. Ces informations concernent les caractéristiques syntaxiques du texte, qui pourraient aussi permettre de reconnaître de manière encore plus fine des textes faux en montrant une utilisation supérieure de verbes transitifs par exemple, ou de noms propres. Il faut toutefois conserver à l'esprit le fait que cette bibliothèque, et ce traitement de catégorisation est extrêmement coûteux en termes de ressources et demanderait des configurations matérielles supérieures à celle d'un ordinateur classique, notamment en termes de RAM disponible et de puissance de calcul du processeur.

Enfin, un travail plus approfondi sur la partie web scraping aurait aussi pu permettre une meilleure qualité de l'information en entrée. Ce processus d'automatisation nous a ainsi permis de récupérer du texte dans sa globalité, toutefois cette récupération était dépendante de plusieurs facteurs. Tout d'abord, elle supposait que le formatage des pages était similaire, et que les titres étaient bien identifiés par les balises HTML <h1>, <h2> et <h3>. De plus, le texte récupéré contenait aussi d'autres informations qui pourraient ne pas apparaître pertinentes pour cette étude précise, comme par exemple la présence des licences auxquelles étaient soumises les sites web, certains textes de publicité ou encore des textes issus de menus de navigation. Une expertise en technologie web devrait ainsi améliorer la qualité des informations récupérées. La prise en compte des différents éléments d'une page web, comme les scripts JavaScript ou la mise en forme CSS étant bien au-delà des compétences requises pour ce travail, ceci n'a ainsi pas pu faire partie de la partie développement du scraper.

## **4.3 – Perspectives futures**

Ce travail a permis de poser certaines fondations sur les limites et difficultés de la détection de fake-news écrites en français sur les plates-formes autres que les réseaux sociaux, et plusieurs voies pour prolonger s'ouvrent afin d'améliorer ce processus. Comme dit précédemment, un des travaux fondateurs qui pourrait être réalisé dans le domaine serait une classification manuelle s'appuyant sur les données du Décodex. Ces dernières fournissent déjà une base considérable relative aux articles faux, et un travail d'extension de cette base fournirait des données qui pourraient être valorisées par tout un pan de la communauté académique.

Les recherches futures devraient aussi s'appuyer systématiquement sur un ensemble de recherches pluridisciplinaires, afin d'intégrer tous les enjeux relatifs à ce problème. Plusieurs problèmes épistémologiques sont liés aux fake-news, et ces dernières ouvrent la voie à plusieurs possibilités d'extension : De la recherche sur les intentions d'auteur à la prévention de la propagation des fausses informations. De plus, l'efficacité d'un algorithme de machine learning sur la détection des fake-news ne pourra être véritablement évaluée que si ce dernier est pris en compte dans une série d'études expérimentales montrant si l'utilisation d'un tel modèle permet véritablement d'apporter une plus-value psychologique et sociale dans la propagation et la rétention des fake-news.

Une autre possibilité serait quant à elle d'investiguer plus en profondeur les questions des sources contenues dans un article. Parvenir à les identifier afin de collecter les données présentes sur les sites en lien avec un article faux pourrait aussi permettre d'améliorer la qualité des classifications faites.

## **4.4 – Conclusion**

Avec l'ère de l'information est apparue l'ère de la post-vérité. Avec la profusion d'informations contradictoires, le travail d'évaluation de la pertinence de l'information s'est complexifié. Les enjeux des diffuseurs d'information n'ont pas aidé à simplifier ce travail, car leurs enjeux économiques favorisent la diffusion d'informations simples à traiter par les consommateurs, impactantes pour rester en mémoire et suffisamment cohérentes pour leur fournir rapidement une explication satisfaisante sur un événement particulier. Le problème des fake-news est leur capacité de mimétisme dans ce processus de traitement de l'information, avec la seule différence qui concerne l'intention de l'auteur et la fiabilité du contenu à sa source. Cette simple différence peut cependant avoir des effets dramatiques, et biaiser la représentation du monde d'un individu. Cet effet est d'autant plus critique que la plupart des stratégies de

correction des informations erronées échouent, ce qui oblige à porter un intérêt redoublé pour la prévention des fake-news, et l'éducation dès le plus jeune âge à l'esprit critique.

Nous avons essayé ici de proposer une base de travail qui permettrait de développer un algorithme de classification des fake-news pour des articles rédigés en français. La plupart du travail sur cette question se faisant en effet sur des articles anglophones. Pour cela, nous avons travaillé avec la base de données du Décodex, un outil développé par des journalistes du quotidien Le Monde, qui avait déjà classifié des articles présentant des fake-news en proposant des articles plus factuels. Pour cela, nous avons récupéré le contenu de chaque article de cette base en utilisant un outil de web scraping. Nous avons ensuite développé un algorithme qui utiliserait le contenu de ces articles pour tenter de trouver des features discriminant un article vrai d'un article faux. Cependant, la qualité du contenu des articles scrapés ainsi que les problèmes de déséquilibre de classe n'ont pas permis d'arriver à un résultat suffisamment satisfaisant pour être exploitable, car notre modèle était « trop prudent », c'est-à-dire qu'il avait tendance à catégoriser trop souvent un article comme appartenant à la classe des faux articles. Nous avons cependant proposé d'autres pistes d'amélioration de ce travail, discuté des données, des métriques d'évaluation et proposons des pistes de directions futures dans la recherche sur la détection de fake-news en français. Enfin, nous soulignons particulièrement l'importance première que la recherche sur ce sujet ne soit pas spécifique à un domaine, mais réunisse l'expertise de plusieurs domaines, allant de l'intelligence artificielle à l'ingénierie en passant par les sciences humaines comme la psychologie et la sociologie.

# Webographie

- 1 : Site de Oxford Dictionaries - <https://languages.oup.com/word-of-the-year/word-of-the-year-2016>
- 2 : Site de PolitiFact : <https://www.politifact.com/truth-o-meter/article/2015/dec/21/2015-lie-year-donald-trump-campaign-misstatements/>
- 3 : Comparaison PolitiFact Trump/Clinton : <https://www.politifact.com/truth-o-meter/lists/people/comparing-hillary-clinton-donald-trump-truth-o-met/>
- 4 : Statistiques ethniques des meurtres aux Etats-Unis pour l'année 2014 : [https://ucr.fbi.gov/crime-in-the-u-s/2014/crime-in-the-u-s-2014/tables/expanded-homicide-data/expanded\\_homicide\\_data\\_table\\_6\\_murder\\_race\\_and\\_sex\\_of\\_victim\\_by\\_race\\_and\\_sex\\_of\\_offender\\_2014.xls](https://ucr.fbi.gov/crime-in-the-u-s/2014/crime-in-the-u-s-2014/tables/expanded-homicide-data/expanded_homicide_data_table_6_murder_race_and_sex_of_victim_by_race_and_sex_of_offender_2014.xls)
- 5 : « Accuracy is for snake-oil pussies » <https://www.theguardian.com/politics/2016/apr/20/accuracy-is-for-snake-oil-pussies-vote-leaves-campaign-director-defies-mps>
- 6 : The truth, the whole truth and... Wait, how many truths are there ? <https://theconversation.com/the-truth-the-whole-truth-and-wait-how-many-truths-are-there-6955>
- 7 : <https://www.latimes.com/opinion/op-ed/la-oe-baum-lazer-how-to-fight-fake-news-20170508-story.html>
- 8 : James Hansen à propos du réchauffement climatique en 1988 :  
<https://www.nytimes.com/1988/03/29/science/temperature-for-world-rises-sharply-in-the-1980-s.html>
- 9 : La personnalisation du web nous enferme-t-elle dans notre bulle ?  
<https://www.slate.fr/story/39977/web-bulle-personnalisation-google>
- 10 : Trump claims vaccines and autism are linked but his own experts vehemently disagree  
<https://www.independent.co.uk/news/world/americas/trump-vaccines-autism-links-anti-vaxxer-us-president-false-vaccine-a8331836.html>
- 11 : Le fact-checking en France en une chronologie : <https://larevuedesmedias.ina.fr/le-fact-checking-en-france-en-une-chronologie>
- 12 : Premier article sur le Décodex : [https://www.lemonde.fr/les-decodeurs/article/2017/02/02/le-decodex-un-premier-pas-vers-la-verification-de-masse-de-l-information\\_5073130\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2017/02/02/le-decodex-un-premier-pas-vers-la-verification-de-masse-de-l-information_5073130_4355770.html)
- 13 : Charte du Décodex : [https://www.lemonde.fr/les-decodeurs/article/2014/03/10/la-charte-des-decodeurs\\_4365106\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2014/03/10/la-charte-des-decodeurs_4365106_4355770.html)
- 14 : Le problème du web scraping de données Facebook : <https://www.octoparse.com/blog/5-things-you-need-to-know-before-scraping-data-from-facebook>
- 15 : Web Scraping in an Era of Big Data 2.0 :  
<https://www.perkinscoie.com/images/content/1/5/v2/156775/Snell-web-scraping-BNAI.pdf>
- 16 : Step-By-Step Framework for Imbalanced Classification Projects :  
<https://machinelearningmastery.com/framework-for-imbalanced-classification-projects/>
- 17 : Sci-Kit learn - Choosing the right estimator : [https://scikit-learn.org/dev/tutorial/machine\\_learning\\_map/index.html](https://scikit-learn.org/dev/tutorial/machine_learning_map/index.html)

# Bibliographie

- Alelyani, S., Tang, J., & Liu, H. (2017). Feature Selection for Clustering : A Review. *ACM Computing Surveys (CSUR)*, 50(6), 45.
- Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), 211-236. <https://doi.org/10.1257/jep.31.2.211>
- An Enquiry Concerning Human Understanding*. David Hume. (1748). 139.
- Androutsopoulos, I., Koutsias, J., Chandrinou, K. V., Palioras, G., & Spyropoulos, C. D. (2000). An evaluation of Naive Bayesian anti-spam filtering. *arXiv:cs/0006013*. <http://arxiv.org/abs/cs/0006013>
- Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, 2(2), 81-94. <https://doi.org/10.1002/bdm.3960020203>
- Asch, S. E. (1956). Studies of independence and conformity : I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9), 1-70. <https://doi.org/10.1037/h0093718>
- Bacon, F. T. (1979). Credibility of Repeated Statements : Memory for Trivia. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 241-252.
- Bargh, J. A. (1994). The four horsemen of automaticity : Awareness, intention, efficiency, and control in social cognition. *Handbook of Social Cognition*, 1, 1-40.
- Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will : Nonconscious activation and pursuit of behavioral goals. *Journal of personality and social psychology*, 81(6), 1014.
- Bargh, J. A., Schwader, K. L., Hailey, S. E., Dyer, R. L., & Boothby, E. J. (2012). Automaticity in social-cognitive processes. *Trends in Cognitive Sciences*, 16(12), 593-605. <https://doi.org/10.1016/j.tics.2012.10.002>
- Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of Processes in Belief : Source Recollection, Statement Familiarity, and the Illusion of Truth. *Journal of Experimental Psychology: General*, 121, 13.
- Brehm, S. S., & Brehm, J. W. (1981). *Psychological reactance : A theory of freedom and control*. (Academic Press).
- Burkhardt, J. M. (2017). History of Fake News. *Library Technology Reports*, 53(8), 5-9.
- Carper, L. (2019). What influences our decision to vaccinate? The social amplification of risk framework and vaccination. *Proceedings of the International Crisis and Risk Communication Conference*, 2(2019), 12-14. <https://doi.org/10.30658/icrcc.2019.3>
- Carrieri, V., Madio, L., & Principe, F. (2019). Vaccine hesitancy and (fake) news : Quasi-experimental evidence from Italy. *Health Economics*, 28(11), 1377-1382. <https://doi.org/10.1002/hec.3937>
- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5), 752.
- Chambers, K. L., & Zaragoza, M. S. (2001). Intended and unintended effects of explicit warnings on eyewitness suggestibility : Evidence from source identification tests. *Memory & Cognition*, 29(8), 1120-1129. <https://doi.org/10.3758/BF03206381>

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357. <https://doi.org/10.1613/jair.953>
- Debarr, D., & Wechsler, H. (2009). Spam Detection using Clustering, Random Forests and Active Learning. "Spam Detection using Clustering, Random Forests, and Active Learning", *CEAS 2009 – Sixth Conference on Email and Anti-Spam*, Mountain View, 16–17.
- Devaux, P., Thyssen-Rutten, N., & Popper, K. (1973). *La Logique de la découverte scientifique* (Payot).
- Dufour, F. (2018). *Les réalités de la réalité—Deuxième partie : Vers une compréhension des facteurs responsables des progrès de la science moderne* (Fritz Dufour, Vol. 2).
- Ecker, U. K. H., Lewandowsky, S., & Apai, J. (2011). Terrorists brought down the plane!—No, actually it was a technical fault : Processing corrections of emotive information. *Quarterly Journal of Experimental Psychology*, 64(2), 283-310. <https://doi.org/10.1080/17470218.2010.497927>
- Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory : Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, 18(3), 570-578. <https://doi.org/10.3758/s13423-011-0065-1>
- Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, 38(8), 1087-1100. <https://doi.org/10.3758/MC.38.8.1087>
- Ellerton, P. (2012, mai 11). The truth, the whole truth and ... wait, how many truths are there? *The Conversation Media Trust*. <https://theconversation.com/the-truth-the-whole-truth-and-wait-how-many-truths-are-there-6955>
- Erickson, T. D., & Mattson, M. E. (1981). From words to meaning : A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 20(5), 540-551. [https://doi.org/10.1016/S0022-5371\(81\)90165-1](https://doi.org/10.1016/S0022-5371(81)90165-1)
- Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, 7(2), 117-140. <https://doi.org/10.1177/001872675400700202>
- Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.
- Flanagin, A. J., & Metzger, M. J. (2008). Digital Media and Youth : Unparalleled Opportunity and Unprecedented Responsibility. *Digital Media*, 5-27.
- Fogg, B. J., Soohoo, C., Danielson, D. R., & Marable, L. (2003). How Do Users Evaluate the Credibility of Web Sites ? A Study with Over 2,500 Participants. *Proceedings of the 2003 Conference on Designing for User Experiences*, 1-15.
- Fuller, S. (2018). What Can Philosophy Teach Us About the Post-truth Condition. In *Post-Truth, Fake News : Viral Modernity & Higher Education* (p. 13-26). Springer Berlin Heidelberg.
- Gallie, W. B. (1956). IX.—Essentially Contested Concepts. *Proceedings of the Aristotelian Society*, 56(1), 167-198. <https://doi.org/10.1093/aristotelian/56.1.167>
- Gerrie, M. P., Belcher, L. E., & Garry, M. (2006). ‘Mind the gap’ : False memories for missing aspects of an event. *Applied Cognitive Psychology*, 20(5), 689-696. <https://doi.org/10.1002/acp.1221>

- Goldsmith, K., & Dhar, R. (2013). Negativity bias and task motivation: Testing the effectiveness of positively versus negatively framed incentives. *Journal of Experimental Psychology: Applied*, 19(4), 358-366. <https://doi.org/10.1037/a0034415>
- Goswami, S., Sudeshna, S., & Maryus, R. (2009). *Stylometric Analysis of Bloggers' Age and Gender*. 4.
- Grundmann, R. (2013). "Climategate" and The Scientific Ethos. *Science, Technology, & Human Values*, 38(1), 67-93. <https://doi.org/10.1177/0162243911432318>
- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1), 107-112. [https://doi.org/10.1016/S0022-5371\(77\)80012-1](https://doi.org/10.1016/S0022-5371(77)80012-1)
- Hong, W., & Walsh, J. P. (2009). For Money or Glory? Commercialization, Competition, and Secrecy in the Entrepreneurial University. *The Sociological Quarterly*, 50(1), 145-171. <https://doi.org/10.1111/j.1533-8525.2008.01136.x>
- Hussain, A., Ali, S., Ahmed, M., & Hussain, S. (2018). The Anti-vaccination Movement : A Regression in Modern Medicine. *Cureus*. <https://doi.org/10.7759/cureus.2919>
- Johnson, D. R., Ecklund, E. H., & Lincoln, A. E. (2014). Narratives of Science Outreach in Elite Contexts of Academic Science. *Science Communication*, 36(1), 81-105. <https://doi.org/10.1177/1075547013499142>
- Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect : When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), 1420-1436. <https://doi.org/10.1037/0278-7393.20.6.1420>
- Jou, J., & Foreman, J. (2007). Transfer of learning in avoiding false memory : The roles of warning, immediate feedback, and incentive. *Quarterly Journal of Experimental Psychology*, 60(6), 877-896. <https://doi.org/10.1080/17470210600831184>
- Kasperson, J. X., Kasperson, R. E., Pidgeon, N., & Slovic, P. (2003). The social amplification of risk : Assessing fifteen years of research and theory. In N. Pidgeon, R. E. Kasperson, & P. Slovic (Éds.), *The Social Amplification of Risk* (1<sup>re</sup> éd., p. 13-46). Cambridge University Press. <https://doi.org/10.1017/CBO9780511550461.002>
- Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., Kasperson, J. X., & Ratick, S. (1988). The Social Amplification of Risk : A Conceptual Framework. *Risk Analysis*, 8(2), 177-187. <https://doi.org/10.1111/j.1539-6924.1988.tb01168.x>
- Keyes, R. (2004). *The Post-Truth Era : Dishonesty and Deception in Contemporary Life*. St. Martin's Publishing Group. <https://books.google.fr/books?id=f0Kvm3KObXoC>
- King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), 27.
- Krotov, V., & Silva, L. (2018). *Legality and Ethics of Web Scraping*. 6.
- Larson, H. J., de Figueiredo, A., Xiaohong, Z., Schulz, W. S., Verger, P., Johnston, I. G., Cook, A. R., & Jones, N. S. (2016). The State of Vaccine Confidence 2016 : Global Insights Through a 67-Country Survey. *EBioMedicine*, 12, 295-301. <https://doi.org/10.1016/j.ebiom.2016.08.042>
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094-1096. <https://doi.org/10.1126/science.aao2998>

- Lee, T. M., Markowitz, E. M., Howe, P. D., Ko, C.-Y., & Leiserowitz, A. A. (2015). Predictors of public climate change awareness and risk perception around the world. *Nature Climate Change*, 5(11), 1014-1020. <https://doi.org/10.1038/nclimate2728>
- Leiserowitz, A. A., Maibach, E. W., Roser-Renouf, C., Smith, N., & Dawson, E. (2013). Climategate, Public Opinion, and the Loss of Trust. *American Behavioral Scientist*, 57(6), 818-837. <https://doi.org/10.1177/0002764212458272>
- Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6), 1093-1096. <https://doi.org/10.1016/j.jesp.2010.05.025>
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction : Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, 13(3), 106-131. <https://doi.org/10.1177/1529100612451018>
- Lieberman, J. D., & Arndt, J. (2000). Understanding the limits of limiting instructions : Social psychological explanations for the failures of instructions to disregard pretrial publicity and other inadmissible evidence. *Psychology, Public Policy, and Law*, 6(3), 677-711. <https://doi.org/10.1037/1076-8971.6.3.677>
- Loftus, E. F. (1996). Memory distortion and false memory creation. *The Bulletin of the American Academy of Psychiatry and the Law*, 24(3), 281-295.
- Mancuso, K., Hauswirth, W. W., Li, Q., Connor, T. B., Kuchenbecker, J. A., Mauck, M. C., Neitz, J., & Neitz, M. (2009). Gene therapy for red-green colour blindness in adult primates. *Nature*, 461(7265), 784-787. <https://doi.org/10.1038/nature08401>
- Mani, I., & Zhang, J. (2003). kNN approach to unbalanced data distributions : A case study involving information extraction. *Proceedings of workshop on learning from imbalanced datasets*, 126. <https://sci2s.ugr.es/keel/pdf/specific/congreso/jzhang.pdf>
- Mase, A. S., Cho, H., & Prokopy, L. S. (2015). Enhancing the Social Amplification of Risk Framework (SARF) by exploring trust, the availability heuristic, and agricultural advisors' belief in climate change. *Journal of Environmental Psychology*, 41, 166-176. <https://doi.org/10.1016/j.jenvp.2014.12.004>
- Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent" : Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, 40(4), 433-449. <https://doi.org/10.1016/j.jesp.2003.07.008>
- Metzger, M. J. (2007). Making sense of credibility on the Web : Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13), 2078-2091. <https://doi.org/10.1002/asi.20672>
- Metzger, M. J., & Flanagan, A. J. (2013). Credibility and trust of information in online environments : The use of cognitive heuristics. *Journal of Pragmatics*, 59, 210-220. <https://doi.org/10.1016/j.pragma.2013.07.012>
- Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, 19(5), 969-974. <https://doi.org/10.3758/s13423-012-0292-0>
- Newman, E. J., Sanson, M., Miller, E. K., Quigley-McBride, A., Foster, J. L., Bernstein, D. M., & Garry, M. (2014). People with Easier to Pronounce Names Promote Truthiness of Claims. *PLoS ONE*, 9(2), e88671. <https://doi.org/10.1371/journal.pone.0088671>

- Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A Survey on Natural Language Processing for Fake News Detection. *ArXiv Preprint*, 11.
- Pennington, N., & Hastie, R. (1992). Explaining the evidence : Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology*, 62(2), 189-206. <https://doi.org/10.1037/0022-3514.62.2.189>
- Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion : Central and peripheral routes to attitude change*. Springer New York.
- Polage, D. C. (2012). Making up History : False Memories of Fake News Stories. *Europe's Journal of Psychology*, 8(2), 245-250. <https://doi.org/10.5964/ejop.v8i2.456>
- Pyszczynski, T., Greenberg, J., & Solomon, S. (1999). A dual-process model of defense against conscious and unconscious death-related thoughts : An extension of terror management theory. *Psychological review*, 106(4), 835.
- Reis, J. C. S., Correia, A., Murai, F., Veloso, A., Benevenuto, F., & Cambria, E. (2019). Supervised Learning for Fake News Detection. *IEEE Intelligent Systems*, 34(2), 76-81. <https://doi.org/10.1109/MIS.2019.2899143>
- Rivenc, F. (1992). *Logique et fondements des mathématiques. Anthologie (1850—1914)*. Payot.
- Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, 7–17. <https://doi.org/10.18653/v1/W16-0802>
- Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI : A Hybrid Deep Model for Fake News Detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17*, 797-806. <https://doi.org/10.1145/3132847.3132877>
- Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3). <https://doi.org/10.1371/journal.pone.0118432>
- Sanna, L. J., Schwarz, N., & Stocker, S. L. (2002). When debiasing backfires : Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3), 497-502. <https://doi.org/10.1037/0278-7393.28.3.497>
- Sasaki, M., & Shinnou, H. (2005). Spam detection using text clustering. *2005 International Conference on Cyberworlds (CW'05)*, 4 pp. - 319. <https://doi.org/10.1109/CW.2005.83>
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing : I. Detection, search, and attention. *Psychological review*, 84(1), 1.
- Schul, Y. (1993). When Warning Succeeds : The Effect of Warning on Success in Ignoring Invalid Information. *Journal of Experimental Social Psychology*, 29(1), 42-62. <https://doi.org/10.1006/jesp.1993.1003>
- Schwarz, N., Bless, H., Strack, F., Klumpp, G., & et al. (1991). Ease of retrieval as information : Another look at the availability heuristic. *Journal of Personality and Social Psychology*, 61(2), 195-202. <https://doi.org/10.1037/0022-3514.61.2.195>
- Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade : Lessons from cognitive psychology. *Behavioral Science & Policy*, 2(1), 85-95. <https://doi.org/10.1353/bsp.2016.0009>
- Seifert, C. M. (2002). The continued influence of misinformation in memory : What makes a correction effective? In *Psychology of Learning and Motivation* (Vol. 41, p. 265-292). Elsevier. [https://doi.org/10.1016/S0079-7421\(02\)80009-3](https://doi.org/10.1016/S0079-7421(02)80009-3)

- Shaw, J., & Porter, S. (2015). *Constructing Rich False Memories of Committing Crime*. 1-11.
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127-190.
- Song, H., & Schwarz, N. (2008). Fluency and the Detection of Misleading Questions : Low Processing Fluency Attenuates the Moses Illusion. *Social Cognition*, 26(6), 791-799. <https://doi.org/10.1521/soco.2008.26.6.791>
- Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding Psychological Reactance : New Developments and Findings. *Zeitschrift Für Psychologie*, 223(4), 205-214. <https://doi.org/10.1027/2151-2604/a000222>
- Strickland, E. (2018). AI-human partnerships tackle « fake news » : Machine learning can get you only so far-then human judgment is required - [News]. *IEEE Spectrum*, 55(9), 12-13. <https://doi.org/10.1109/MSPEC.2018.8449036>
- Sundar, S. S. (2008). The MAIN Model : A Heuristic Approach to Understanding Technology Effects on Credibility. *Digital Media*, 73-100.
- Tenney, E. R., Cleary, H. M. D., & Spellman, B. A. (2009). Unpacking the Doubt in “Beyond a Reasonable Doubt” : Plausible Alternative Stories Increase Not Guilty Verdicts. *Basic and Applied Social Psychology*, 31(1), 1-8. <https://doi.org/10.1080/01973530802659687>
- Venturini, T., Jacomy, M., Bounegru, L., & Gray, J. (2018). Visual Network Exploration for Data Journalists. In S. A. Eldridge & B. Franklin (Éds.), *The Routledge Handbook of Developments in Digital Journalism Studies* (1<sup>re</sup> éd., p. 265-283). Routledge. <https://doi.org/10.4324/9781315270449-21>
- Verstraete, M., Bambauer, D. E., & Bambauer, J. R. (2017). Identifying and Countering Fake News. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3007971>
- Waldrop, M. M. (2017). News Feature : The genuine problem of fake news. *Proceedings of the National Academy of Sciences*, 114(48), 12631-12634. <https://doi.org/10.1073/pnas.1719005114>
- Waytz, A. (2014). The mind in the machine : Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 5.
- Weaver, K., Garcia, S. M., Schwarz, N., & Miller, D. T. (2007). Inferring the popularity of an opinion from its familiarity : A repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology*, 92(5), 821-833. <https://doi.org/10.1037/0022-3514.92.5.821>
- Weingart, P. (2002). The moment of truth for science : The consequences of the ‘knowledge society’ for society and science. *EMBO Reports*, 3(8), 703-706. <https://doi.org/10.1093/embo-reports/kvf165>
- Winkielman, P., Huber, D. E., Kavanagh, L., & Schwarz, N. (2011). *Fluency of consistency : When thoughts fit nicely and flow smoothly*. 36.
- Woleński, J. (2019). *Semantics and Truth* (Vol. 45). Springer International Publishing. <https://doi.org/10.1007/978-3-030-24536-8>
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2, Pt.2), 1-27. <https://doi.org/10.1037/h0025848>

## Annexes

### Annexe 1 – Transformation des données du Décodex en dataframe

#### 1 - decodex\_data\_preprocessing

April 25, 2020

```
[1]: # Import libraries
import json
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import tldextract
import urllib.request

from collections import Counter
from datetime import datetime

[2]: # Open the Décodex data which can be downloaded at
# http://s1.lemede.fr/mmpub/data/decodex/hoax/hoax_debunks.json
decodex_url = 'http://s1.lemede.fr/mmpub/data/decodex/hoax/hoax_debunks.json'
with urllib.request.urlopen(decodex_url) as f:
    df = json.load(f)

[3]: # The file is a dictionary with 2 main keys : debunks and hoaxes
# debunks = debunk_number['What we think', 'True or False', 'What is known', ↴
#   'Link to a more factual article']
debunks = df['debunks']
# hoaxes = {'hoax.address.com' : 'debunk_number'}
hoaxes = df['hoaxes']
print('First debunk:', debunks['1'], '\nFirst Hoax:', list(hoaxes.keys())[0])

First debunk: ["Ce prêtre a-t-il été agressé à Avignon en février 2017, sans que les médias n'en parlent ?", "FAUX", "La rumeur fait référence à une vraie agression, mais qui date de mai 2013. Les faits ont alors été largement évoqués par de nombreux médias régionaux et nationaux.", "https://www.lemonde.fr/les-decodeurs/article/2017/02/27/des-militants-denoncent-l-omerta-des-medias-sur-l-agression-d-un-pretre-a-avignon-qui-date-de-2013_5086287_4355770.html"]

First Hoax: https://www.facebook.com/CorentinFNJ/posts/250959068694828

[4]: # We want the hoaxes to have a more regular format like {'hoax_number': ['hoax1. ↴
#   address.com', 'hoax2.address.com']}
```

```

hxs = {}
for key, value in hoaxes.items():
    if value in hxs:
        hxs[value].append(key)
    else:
        hxs[value]=[key]

```

[5]: # Now our hoaxes dict is comprised of keys which correspond to the debunk\_id  
# followed by a list of the links to the hoax posts  
hxs['1']

[5]: ['https://www.facebook.com/CorentinFNJ/posts/250959068694828',  
'http://www.paulomouvementcitoyen.com/2017/02/un-pretre-agresse-a-avignon.html',  
'https://www.blog.sami-aldeeb.com/2017/02/06/pretre-agresse-a-avignon-si-ca-avait-ete-un-imam/',  
'https://francaisdefrance.wordpress.com/2016/12/12/agression-dun-moine-hier-soir-silence-total-heureusement-il-y-a-internet/',  
'https://www.facebook.com/10212204315883498/posts/10212140163199721',  
'https://www.facebook.com/387841388254774/posts/383701835335396',  
'https://twitter.com/MONSTERLOVE696/status/976170425272160258']

[6]: # Now we can convert our hxs dict to a DataFrame  
df\_hxs = pd.DataFrame.from\_dict(hxs, orient='index')  
df\_hxs = df\_hxs.stack().to\_frame().reset\_index().drop('level\_1', axis=1)  
df\_hxs.columns = ['debunk\_id', 'hoax\_link']  
df\_hxs.head(5)

	debunk_id	hoax_link
0	1	https://www.facebook.com/CorentinFNJ/posts/250...
1	1	http://www.paulomouvementcitoyen.com/2017/02/u...
2	1	https://www.blog.sami-aldeeb.com/2017/02/06/pr...
3	1	https://francaisdefrance.wordpress.com/2016/12...
4	1	https://www.facebook.com/10212204315883498/pos...

[7]: # We can do the same for the debunks dict  
df\_dbnk = pd.DataFrame.from\_dict(debunks, orient='index',  
 columns = ['hoax\_summary', 'true\_false',  
 'debunk\_brief', 'debunk\_link'])  
df\_dbnk.head(5)

	hoax_summary	true_false	\
1	Ce prêtre a-t-il été agressé à Avignon en févr...	FAUX	
2	Emmanuel Macron a-t-il déclaré vouloir faire p...	FAUX	
3	Ce pompier a-t-il eu « l'œil crevé par une rac...	FAUX	
4	Un élu PS a-t-il été relaxé après un excès de ...	FAUX	
5	Les médias ont-ils caché l'affaire des « frais...	FAUX	

```
debunk_brief \
1 La rumeur fait référence à une vraie agression...
2 Il s'agit d'un article du site parodique BuzzB...
3 Cette rumeur qui a circulé en février 2017 est...
4 Contrairement à ce que laisse entendre la rume...
5 Contrairement à ce que les rumeurs qui circule...
```

```
debunk_link
1 https://www.lemonde.fr/les-decodeurs/article/2...
2 https://crosscheck.firstdraftnews.com/checked...
3 https://www.lemonde.fr/les-decodeurs/article/2...
4 https://www.lemonde.fr/les-decodeurs/article/2...
5 https://www.lemonde.fr/les-decodeurs/article/2...
```

```
[8]: # Since the index value is based on our dict key, we can now pass it as an id
      ↵column
df_dbnk['debunk_id'] = df_dbnk.index
df_dbnk.head(5)
```

```
hoax_summary true_false \
1 Ce prêtre a-t-il été agressé à Avignon en févr... FAUX
2 Emmanuel Macron a-t-il déclaré vouloir faire p... FAUX
3 Ce pompier a-t-il eu « l'œil crevé par une rac... FAUX
4 Un élu PS a-t-il été relaxé après un excès de ... FAUX
5 Les médias ont-ils caché l'affaire des « frais... FAUX
```

```
debunk_brief \
1 La rumeur fait référence à une vraie agression...
2 Il s'agit d'un article du site parodique BuzzB...
3 Cette rumeur qui a circulé en février 2017 est...
4 Contrairement à ce que laisse entendre la rume...
5 Contrairement à ce que les rumeurs qui circule...
```

```
debunk_link debunk_id
1 https://www.lemonde.fr/les-decodeurs/article/2... 1
2 https://crosscheck.firstdraftnews.com/checked... 2
3 https://www.lemonde.fr/les-decodeurs/article/2... 3
4 https://www.lemonde.fr/les-decodeurs/article/2... 4
5 https://www.lemonde.fr/les-decodeurs/article/2... 5
```

```
[9]: # Now we need to merge the true_false column to the hoaxes DataFrame
df_webs = pd.merge(df_hxs, df_dbnk[['debunk_id', 'true_false']], on =
      ↵'debunk_id', how = "left")
df_webs.head(5)
```

```
[9]:   debunk_id                      hoax_link true_false
 0      1 https://www.facebook.com/CorentinFNJ/posts/250... FAUX
 1      1 http://www.paulomouvementcitoyen.com/2017/02/u... FAUX
 2      1 https://www.blog.sami-aldeeb.com/2017/02/06/pr... FAUX
 3      1 https://francaisdefrance.wordpress.com/2016/12... FAUX
 4      1 https://www.facebook.com/10212204315883498/pos... FAUX
```

```
[10]: # Find the unique values in the true_false column
print('Unique values in true_false categorization before processing:', df_webs.
     ↪true_false.unique())

# Replace these to have fewer categories
df_webs = df_webs.replace(
    'C'est plus compliqué', 'CONTESTABLE').replace(
    'Trompeur', 'TROMPEUR').replace(
    'Prudence', 'CONTESTABLE').replace(
    'PRUDENCE', 'CONTESTABLE').replace(
    'DOUTEUX', 'TROMPEUR')

print('Unique values in true_false categorization after processing:', df_webs.
     ↪true_false.unique())
```

```
Unique values in true_false categorization before processing: ['FAUX'
'CONTESTABLE' 'DOUTEUX' 'TROMPEUR' 'C'est plus compliqué'
'Trompeur' 'PRUDENCE']
Unique values in true_false categorization after processing: ['FAUX'
'CONTESTABLE' 'TROMPEUR']
```

```
[11]: # Add the link to the articles which tell the true story
# We'll label those as 'VRAI'
pd.options.mode.chained_assignment = None
tmp_dbnk = df_dbnk[['debunk_id', 'debunk_link']]
tmp_dbnk.columns = ['debunk_id', 'hoax_link']
tmp_dbnk['true_false'] = 'VRAI'
```

```
[12]: # Then we can bind the rows of this df to the df_webs
df_webs = df_webs.append(tmp_dbnk, sort = False)
df_webs['debunk_id'] = pd.to_numeric(df_webs['debunk_id'])
df_webs = df_webs.sort_values('debunk_id').reset_index(drop=True)

df_webs.head(10)
```

```
[12]:   debunk_id                      hoax_link true_false
 0      1 https://www.facebook.com/CorentinFNJ/posts/250... FAUX
 1      1 http://www.paulomouvementcitoyen.com/2017/02/u... FAUX
 2      1 https://www.blog.sami-aldeeb.com/2017/02/06/pr... FAUX
 3      1 https://francaisdefrance.wordpress.com/2016/12... FAUX
```

```

4      1 https://www.facebook.com/10212204315883498/posts... FAUX
5      1 https://www.facebook.com/387841388254774/posts... FAUX
6      1 https://twitter.com/MONSTERLOVE696/status/9761... FAUX
7      1 https://www.lemonde.fr/les-decodeurs/article/2... VRAI
8      2 https://www.facebook.com/439281732799406/posts... FAUX
9      2 https://www.facebook.com/PorteTesCouilles2/pos... FAUX

```

```
[13]: # Now that we have a clean dataframe with all the links to the articles
# we can check if some articles are present more than once
ids = df_webs['hoax_link']
duplicate_articles = df_webs[ids.isin(ids[ids.duplicated()])]
dup_articles_list = dict(Counter(duplicate_articles['hoax_link']))
print('Nombre d\'articles uniques présents dans le Décodex :', len(set(df_webs['hoax_link'])), '\n', '\x1b[1m', 'Articles présents en doublon dans la base :', '\x1b[0m')
for link, count in dup_articles_list.items() :
    print('Lien :', link, '- Nombre d\'occurrences :', count)
```

Nombre d'articles uniques présents dans le Décodex : 13450  
 Articles présents en doublon dans la base :

Lien : [https://www.lemonde.fr/les-decodeurs/article/2017/03/01/intox-en-serie-sur-des-pages-facebook-d-extreme-droite\\_5087453\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2017/03/01/intox-en-serie-sur-des-pages-facebook-d-extreme-droite_5087453_4355770.html) - Nombre d'occurrences : 3  
 Lien : [https://www.lemonde.fr/les-decodeurs/article/2016/09/09/les-invraisemblables-intox-sur-la-sante-de-hillary-clinton\\_4995400\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2016/09/09/les-invraisemblables-intox-sur-la-sante-de-hillary-clinton_4995400_4355770.html) - Nombre d'occurrences : 3  
 Lien : [https://www.lemonde.fr/les-decodeurs/article/2017/04/12/bombardement-chimique-de-khan-cheikhoun-en-syrie-les-intox-a-1-epreuve-des-faits\\_5110175\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2017/04/12/bombardement-chimique-de-khan-cheikhoun-en-syrie-les-intox-a-1-epreuve-des-faits_5110175_4355770.html) - Nombre d'occurrences : 2  
 Lien : [https://www.lemonde.fr/les-decodeurs/article/2017/04/21/etudes-bidons-rumeurs-boules-puantes-le-grand-n-importe-quoi-de-fin-de-campagne\\_5115237\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2017/04/21/etudes-bidons-rumeurs-boules-puantes-le-grand-n-importe-quoi-de-fin-de-campagne_5115237_4355770.html) - Nombre d'occurrences : 4  
 Lien : [https://www.lemonde.fr/les-decodeurs/article/2017/07/11/antivaccins-des-mensonges-dans-un-debat-legitime\\_5159187\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2017/07/11/antivaccins-des-mensonges-dans-un-debat-legitime_5159187_4355770.html) - Nombre d'occurrences : 4  
 Lien : [https://www.lemonde.fr/les-decodeurs/article/2016/11/07/les-intox-du-fn-sur-les-privileges-des-migrants-face-aux-francais\\_5026857\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2016/11/07/les-intox-du-fn-sur-les-privileges-des-migrants-face-aux-francais_5026857_4355770.html) - Nombre d'occurrences : 2  
 Lien : <https://www.bfmtv.com/mediaplayer/video/ouragan-irma-il-y-a-un-manque-de-maturite-un-manque-du-sens-du-deploiement-de-l-etat-gilbert-collard-979363.html> - Nombre d'occurrences : 2  
 Lien : [https://www.lemonde.fr/les-decodeurs/article/2018/09/17/rafales-de-canulars-et-videos-detournees-sur-la-tempete-florence-et-le-typhon-mangkhut\\_5356361\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2018/09/17/rafales-de-canulars-et-videos-detournees-sur-la-tempete-florence-et-le-typhon-mangkhut_5356361_4355770.html) - Nombre d'occurrences : 2  
 Lien : [https://www.lemonde.fr/les-decodeurs/article/2019/03/11/equipage-feminin-turbulences-filmees-fausses-photos-trois-intox-sur-le-crash-d-ethiopian-airlines\\_5434477\\_4355770.html](https://www.lemonde.fr/les-decodeurs/article/2019/03/11/equipage-feminin-turbulences-filmees-fausses-photos-trois-intox-sur-le-crash-d-ethiopian-airlines_5434477_4355770.html) - Nombre d'occurrences : 3

```
[14]: # Here, we check for the most used sources of information
# First, we extract the url using tldextract, then we count the number of ↵
# occurrences for each of these
most_used_links = Counter(ids.apply(lambda url: tldextract.extract(url).domain))
mul_df = pd.DataFrame.from_dict(most_used_links, orient='index').reset_index()
mul_df.columns = ['url', 'count']
mul_df = mul_df.sort_values(by=['count'], ascending=False)
mul_df.head(10).style.hide_index()
```

[14]: <pandas.io.formats.style.Styler at 0x236a4674c48>

```
[15]: print('Nombre de sources différentes :', len(mul_df))

# Since there are many different information sources, we will only keep the 10 ↵
# most represented and
# concatenate the rest under the 'Autres' category
mul_df_plot = mul_df.copy()
mul_df_plot['url'] = np.where(mul_df_plot['count'] < 30, 'Autres', ↵
    mul_df_plot['url'])
mul_df_plot = mul_df_plot.groupby(by=['url']).sum().sort_values(by=['count'], ↵
    ascending=False).reset_index()
mul_df_plot['perc']= mul_df_plot['count']*100/mul_df_plot['count'].sum()
```

Nombre de sources différentes : 1186

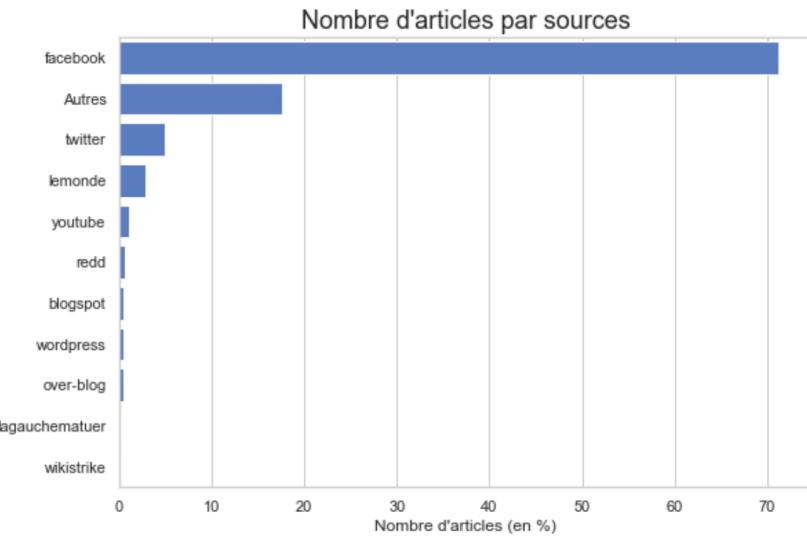
```
[16]: sns.set(style = 'whitegrid')

# Initialize the matplotlib figure
f, ax = plt.subplots(figsize = (9, 6))

# Plot the information sources
sns.set_color_codes('muted')
sns.barplot(x = 'perc', y = 'url', data = mul_df_plot,
            label = 'Total', color = 'b').set_title('Nombre d\'articles par ↵
sources', fontsize = 18)

# Add an informative axis label
ax.set(ylabel = '',
       xlabel = 'Nombre d\'articles (en %)')
```

[16]: [Text(0, 0.5, ''), Text(0.5, 0, "Nombre d'articles (en %)")]



```
[17]: # Here we'll drop the rows where the dns is equal to Facebook, Twitter, Youtube or Reddit
# First, we add the url column to df_webs
df_webs['url_dns'] = df_webs['hoax_link'].apply(lambda url: tldextract.extract(url).domain)

# Then we filter the rows containing the websites we do not want to scrape.
rows_to_rm = ['facebook', 'twitter', 'youtube', 'redd']
df_webscraping = df_webs[~df_webs['url_dns'].isin(rows_to_rm)].reset_index().drop('index', axis=1).drop('url_dns', axis=1)
df_webscraping.head(5)

# We also filter the debunk_ids which are only equal to 1 (which means there were only social networks)
# links in the hoaxes document and only the debunk remains)
df_webscraping = df_webscraping.groupby('debunk_id')
df_webscraping = df_webscraping.filter(lambda x: len(x) > 1)
```

```
[18]: # Finally, we must take a look at our label to see the class imbalance and
sns.set(style = 'whitegrid')

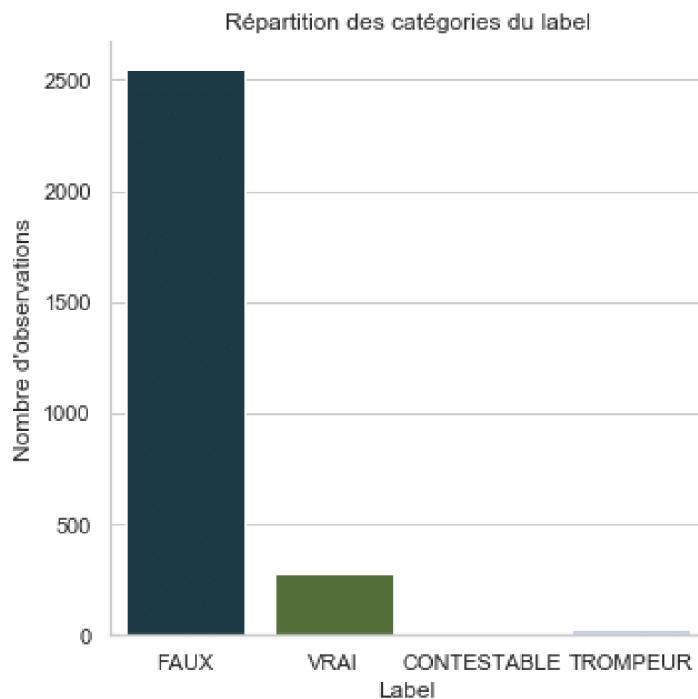
# Plot the information sources
sns.catplot(x = 'true_false',
```

```

        kind = 'count',
        palette = 'cubebehelix',
        data=df_webscraping).set(title = 'Répartition des catégories du',
        ↪label',
        xlabel = 'Label',
        ylabel = 'Nombre d\'observations')

```

[18]: <seaborn.axisgrid.FacetGrid at 0x236a45d8108>



```

[19]: # We can safely remove the "Contestable" and "Trompeur" categories, which won't
      ↪be informative in this analysis
searchfor = ['TROMPEUR', 'CONTESTABLE']
df_webscraping = df_webscraping[~df_webscraping.true_false.str.contains('||'.
      ↪join(searchfor))]

# We also filter the debunk_ids which are only equal to 1
df_webscraping = df_webscraping.groupby('debunk_id')

```

```

df_webscraping = df_webscraping.filter(lambda x: len(x) > 1)

df_webscraping.head(5)

[19]:   debunk_id                      hoax_link true_false
0          1 http://www.paulomouvementcitoyen.com/2017/02/u... FAUX
1          1 https://www.blog.sami-aldeeb.com/2017/02/06/pr... FAUX
2          1 https://francaisdefrance.wordpress.com/2016/12... FAUX
3          1 https://www.lemonde.fr/les-decodeurs/article/2... VRAI
4          2 https://www.buzzbeed.com/macron-veut-faire-pay... FAUX

[20]: filename = './data/df_webscraping_' + datetime.now().strftime("%Y-%m-%d-%H%M%S".
    ↵'csv')
      df_webscraping.to_csv(filename, encoding = 'utf-8', header = True, index = None)

```

## Annexe 2 – Web Scraping

### 2 - decodex\_webscraping

April 25, 2020

```
[1]: import pandas as pd
import numpy as np
import nbprogress
import re

from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from datetime import datetime

[2]: # Import our file
df = pd.read_csv('data/df_webscraping_2020-04-13-111815.csv', header = 0, encoding = 'utf-8', sep = ',')
df.head(5)

[2]:   debunk_id          hoax_link true_false
0           1 http://www.paulomouvementcitoyen.com/2017/02/u... FAUX
1           1 https://www.blog.sami-aldeeb.com/2017/02/06/pr... FAUX
2           1 https://francaisdefrance.wordpress.com/2016/12... FAUX
3           1 https://www.lemonde.fr/les-decodeurs/article/2...
4           2 https://www.buzzbeed.com/macron-veut-faire-pay... FAUX

[3]: # Here, we add the columns which we'll populate with the web scraper
df['title'] = np.nan
df['body'] = np.nan
df['sources'] = np.nan

[4]: def webscrape_decodex(df) :

    driver_path = 'app/chromedriver.exe'
    brave_path = 'C:/Program Files (x86)/BraveSoftware/Brave-Browser/Application/brave.exe'

    chrome_options = Options()
    chrome_options.binary_location = brave_path

    row_number = np.where(pd.isnull(df['body']))[0][0]
```

1

```

erreurs = 0

for idx, row in nbprogress.log(df.iterrows(), size = len(df)) :
    if idx < row_number :
        pass
    else :
        try :
            browser = webdriver.Chrome(executable_path = driver_path, options = chrome_options)
            url = row['hoax_link']
            browser.get(url)
            html_source = browser.page_source
            browser.quit()
            soup = BeautifulSoup(html_source, 'html.parser')

            #####
            ### Title Section ###
            #####

            try :
                title = soup.find('h1').text.strip()
            except :
                title = ''
            try :
                subtitle = soup.find('h2').text.strip()
            except :
                subtitle = ''

            complete_title = title + subtitle

            #####
            ### Text Content Section ###
            #####

            text = ''
            for p_tag in soup.find_all('p') :
                text = text + ' ' + p_tag.text.strip()
            text = text.strip()

            #####
            ### URL List Section ###
            #####

            url_list = ''
            for link in soup.body.find_all('a', href=True):
                if re.search(r'http', link['href']) :
                    url_list = link['href'] + ', '

```

```

url_list = url_list.strip()
url_list = url_list[:-1]

# Finally, we append those values in our dataframe

df.loc[row_number, 'title'] = complete_title
df.loc[row_number, 'body'] = text
df.loc[row_number, 'sources'] = url_list

row_number += 1

except :
    df.loc[row_number, 'title'] = 'Connection refused'
    df.loc[row_number, 'body'] = 'Connection refused'
    df.loc[row_number, 'sources'] = 'Connection refused'

erreurs += 1

print('Nombre d\'erreurs rencontrées pendant le webscraping : ', erreurs)

filename = './data/webscraped_data_' + datetime.now().
˓→strftime("%Y-%m-%d-%H%M%S.json")
df.to_json(filename)
# Then open the file with pd.read_json (r'Path where you saved the JSON
˓→file\FileName.json')
```

[5]: webscrape\_decode(df)

VBox(children=(HTML(value=''), IntProgress(value=0, max=2816)))

Nombre d'erreurs rencontrées pendant le webscraping : 12

## Annexe 3 – Traitement des données scrapées

### 3 - webscraped\_data\_analysis

April 25, 2020

```
[128]: import pandas as pd
import numpy as np
import seaborn as sns

from datetime import datetime

[105]: # Import webscraped data
df = pd.read_json ('r'data/webscraped_data.json')
df.head(5)

[105]:      debunk_id          hoax_link true_false \
0           1 http://www.paulomouvementcitoyen.com/2017/02/u... FAUX
1           1 https://www.blog.sami-aldeeb.com/2017/02/06/pr... FAUX
2           1 https://francaisdefrance.wordpress.com/2016/12... FAUX
3           1 https://www.lemonde.fr/les-decodeurs/article/2... VRAI
4           2 https://www.buzzbeed.com/macron-veut-faire-pay... FAUX

                           title \
0             Un prêtre agressé à Avignon:
1             Your connection is not private
2             Francaisdefrance's Blog12/12/2016
3   Des militants dénoncent l'omerta des médias su...
4   Macron veut faire payer un loyer aux propriéta...

                           body \
0  action citoyenne, sécurité et citoyenneté Prêt...
1  Attackers might be trying to steal your inform...
2  Soyons solidaire de notre Église, A diffuser, ...
3  Consulter le journal Sur les sites de vente en ...
4  BuzzBeed Site à vocation parodique (dans la me...

                     sources
0  https://www.over-blog.com/cookies
1
2  https://subscribe.wordpress.com/
3  https://lemonde.fr/confidentialite/
4  https://www.buzzbeed.com/sabonner
```

1

```
[106]: # Create a list referencing all failures to scrape data
rows_to_rm = ["We couldn't find the page you are looking for",
              "Your connection is not private",
              "This site can't be reached",
              "404Page non trouvée !",
              "Erreur 404!BLABLA & CHAT EN DIRECT",
              "Not found, error 404Popular Right Now",
              "70news.wordpress.com is no longer available.",
              "Oops! That page can't be found.Recent Posts",
              "Erreur 404",
              "404 Not Found",
              "Not Found",
              "Page Not Found",
              "404Page inexistante",
              "La page que vous avez demandé n'existe pas",
              "Please turn JavaScript on and reload the page.",
              "Oops! Something went wrong here",
              "Page not found (404)",
              "The server can not find the requested page",
              "Error 404",
              "Le blog a été supprimé",
              "Oops! That page can't be found.",
              "403 Forbidden",
              "Apologies, but the page you requested could",
              "404 - Page Not Found",
              "Page non trouvée",
              "Page not found",
              "You may not be able to visit this page because",
              "Oops",
              "Ooops",
              "404 ERROR",
              "Erreur de serveur404",
              "page can't be found",
              "page couldn't be found",
              "is for sale",
              "is no longer available",
              "Cette page n'est malheureusement pas disponible",
              "Cette page n'est pas disponible",
              "503 Service Temporarily Unavailable",
              "Je ne suis pas un robot",
              "This page isn't available",
              "It Looks like the page you are looking for",
              "Erreur de serveur404",
              "UNE ERREUR EST SURVENUE",
              "la page ne peut pas être trouvée",
              "Cette page est introuvable",
              "la page que vous recherchez",
```

```

    "Page désactivée",
    "Error 404"]

# Remove all rows containing previous strings
df_clean = df[~df['title'].str.contains(' | '.join(rows_to_rm), na=False)]
df_clean.head(5)

[106]:   debunk_id                               hoax_link true_false \
0          1 http://www.paulomouvementcitoyen.com/2017/02/u...      FAUX
2          1 https://francaisdefrance.wordpress.com/2016/12...      FAUX
3          1 https://www.lemonde.fr/les-decodeurs/article/2...
4          2 https://www.buzzbeed.com/macron-veut-faire-pay...
5          2 https://crosscheck.firstdraftnews.com/checked-...

                                title \
0           Un prêtre agressé à Avignon:
2           Francaisdefrance's Blog12/12/2016
3  Des militants dénoncent l'omerta des médias su...
4  Macron veut faire payer un loyer aux propriéta...
5  Macron a-t-il déclaré vouloir faire payer un l...

                                body \
0  action citoyenne, sécurité et citoyenneté Prêt...
2  Soyons solidaire de notre Église, A diffuser, ...
3  Consulter le journal Sur les sites de vente en ...
4  BuzzBeed Site à vocation parodique (dans la me...
5  First Draft and \nGoogle News Lab Travailler e...

                                sources
0  https://www.over-blog.com/cookies
2  https://subscribe.wordpress.com/
3  https://lemonde.fr/confidentialite/
4  https://www.buzzbeed.com/sabonner
5  https://firstdraftnews.com

[107]: print("Nombre d'observations supprimées suite au filtrage des erreurs 404 :",
         ~df.shape[0] - df_clean.shape[0])

Nombre d'observations supprimées suite au filtrage des erreurs 404 : 550

[108]: # Remove all rows which failed to be scraped and are equal to a NoneType
df_clean = df_clean.replace(to_replace='None', value=np.nan).dropna()

[109]: df_clean.shape

[109]: (2254, 6)

```

```
[110]: # Remove all rows which have empty values in the title and body columns
cols = ['title', 'body']
df_clean = df_clean[~df_clean[cols].replace('', np.nan).isin([np.nan])].
          ~all(axis=1)]
```

```
[111]: df_clean.shape
```

```
[111]: (2124, 6)
```

```
[112]: print("Nombre d'observations supprimées suite au premier filtrage :", df.
          ~shape[0] - df_clean.shape[0])
```

Nombre d'observations supprimées suite au premier filtrage : 692

```
[114]: df_clean = df_clean.drop(df.index[[17, 19, 20, 26, 34, 50, 52, 55, 57, 58, 59,
                                         ~60, 62, 64, 65, 66, 68, 70, 71, 73,
                                         76, 108, 118, 133, 163, 165, 225, 255, 257,
                                         ~289, 346, 385, 386, 387, 389, 390, 391,
                                         392, 393, 395, 396, 397, 407, 409, 411, 412,
                                         ~413, 414, 415, 416, 417, 418, 419, 420,
                                         421, 439, 440, 441, 466, 467, 468, 469, 470,
                                         ~489, 495, 497, 523, 536, 541, 542, 564,
                                         570, 571, 625, 634, 637, 643, 656, 686, 756,
                                         ~842, 857, 859, 860, 861, 862, 863, 873,
                                         894, 904, 912, 932, 961, 1052, 1090, 1095,
                                         ~1097, 1101, 1124, 1153, 1154, 1156, 1157,
                                         1165, 1167, 1210, 1211, 1218, 1219, 1231,
                                         ~1246, 1250, 1251, 1255, 1256, 1259, 1269,
                                         1270, 1273, 1275, 1277, 1278, 1347, 1351,
                                         ~1378, 1379, 1382, 1386, 1393, 1395, 1396,
                                         1397, 1399, 1401, 1410, 1422, 1424, 1425,
                                         ~1428, 1436, 1437, 1444, 1445, 1446, 1463,
                                         1468, 1470, 1478, 1482, 1501, 1502, 1504,
                                         ~1509, 1510, 1512, 1513, 1514, 1515, 1516,
                                         1518, 1526, 1527, 1530, 1609, 1611, 1617,
                                         ~1643, 1682, 1714, 1738, 1739, 1740, 1741,
                                         1743, 1744, 1745, 1827, 1828, 1829, 1832,
                                         ~1833, 1839, 1841, 1842, 1845, 1852, 1853,
                                         1870, 1911, 1977, 1978, 1979, 1980, 1981,
                                         ~1985, 1994, 2002, 2009, 2024, 2027, 2030,
                                         2034, 2046, 2064, 2072, 2074, 2110, 2126,
                                         ~2127, 2128, 2180, 2193, 2312, 2323, 2334,
                                         2336, 2338, 2367, 2371, 2381, 2389, 2536,
                                         ~2537, 2538, 2539, 2552, 2659, 2744, 2747,
                                         2748, 2750, 2751, 2752, 2755, 2793, 2799]])
```

```
[115]: df_clean.shape
```

```
[115]: (1888, 6)
```

```
[116]: list(df_clean.body.loc[[2711]])
```

```
[116]: ['Amazon va fermer tous ses sites en\xa0France pour cinq jours Amazon : «\xa0La crise nous fait basculer dans les nouveaux Temps modernes\xa0» Le congé s'échange comme une monnaie «\xa0Le gouvernement rompt l'équilibre entre l'activité de l'entreprise et le droit au repos des travailleurs\xa0» Notre-Dame, Saint-Denis... Faut-il reconstruire les monuments détruits\xa0? Incendie de Notre-Dame : pourquoi ces vidéos ne prouvent rien sur les origines du feu Au cœur du chantier de Notre-Dame, « mine d'or pour les chercheurs\xa0» Les masques faits maison sont-ils efficaces\xa0contre le coronavirus\xa0? «\xa0Celles qui se salissent les mains pour les autres\xa0» Roberto Saviano\xa0: «\xa0La faiblesse, c'est de se croire invincible\xa0» «\xa0Gardons-nous de tomber dans une réactivité maladive, viro-induite, sociale et politique\xa0» La Chine, elle aussi, doit annuler des dettes africaines Le comédien Maurice Barrier est mort du\xa0Covid-19 Coulon, Philippon, Reid... Leur livre de chevet en temps de confinement «\xa0OKGB\xa0: le sabre et le bouclier\xa0»\xa0: de la Tcheka au FSB, une histoire secrète rouge sang Pour les festivals, « ce coup d'arrêt risque d'avoir une incidence durable\xa0» Les nuits berlinoises en sommeil S'aimer comme on se confine\xa0: «\xa0Trois semaines à se raconter nos vies. Nos envies. Nos corps\xa0» Cinq cocktails (sans alcool) à réaliser et à siroter avec les enfants Confinés en couple\xa0: cultiver le désir malgré la promiscuité Newsletters du monde Applications Mobiles Abonnement Suivez Le Monde Le Monde utilise des cookies pour vous offrir une expérience utilisateur de qualité, mesurer l'audience, optimiser les fonctionnalités des réseaux sociaux et vous proposer des publicités personnalisées. En poursuivant votre navigation sur ce site, vous acceptez l'utilisation de cookies dans les conditions prévues par notre politique de confidentialité. En savoir plus et gérer les cookies.]
```

```
[117]: # Remove all patterns related to the following message
df_clean['body'] = df_clean['body'].str.replace(r'Consulter le journal Amazon va \u202afermer tous ses sites en\xa0France pour cinq jours Amazon :', '')
df_clean['body'] = df_clean['body'].str.replace(r'Amazon va fermer tous ses \u202asites en\xa0France pour cinq jours Amazon :', '')
```

```
[120]: df_clean = df_clean.drop(df.index[[49, 51, 63, 290, 292, 373, 380, 1020, 1092,\u202a
    1119, 1141, 1155, 1208, 1342, 1344, 1345,
    1346, 1357, 1358, 1360, 1372, 1373, 1385,\u202a
    1400, 1442, 1607, 1645, 2172, 2226, 2638,
    2699, 2771]])
```

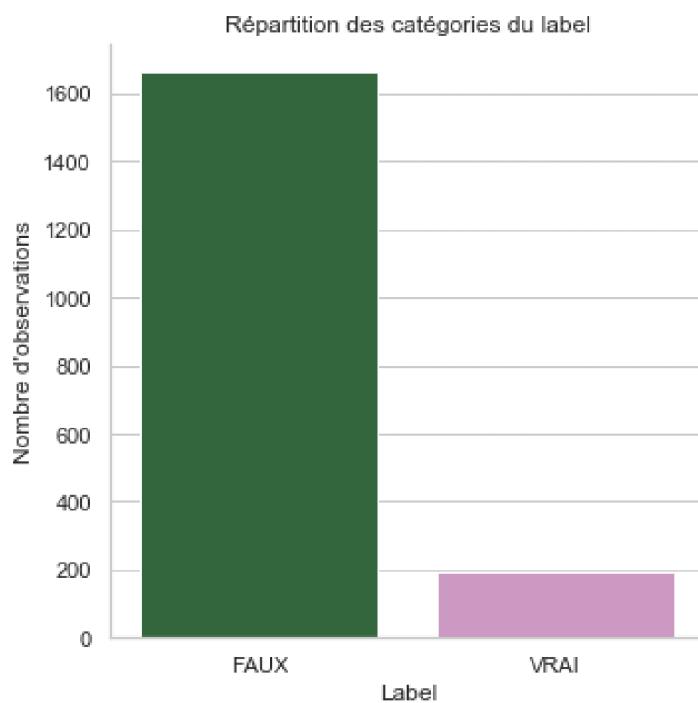
```
[121]: print("Nombre d'observations supprimées suite au premier filtrage :", df.\u202ashape[0] - 692 - df_clean.shape[0])
```

```
Nombre d'observations supprimées suite au premier filtrage : 268
```

```
[124]: # Finally, we must take a look at our label to see the class imbalance
sns.set(style = 'whitegrid')

# Plot the information sources
sns.catplot(x = 'true_false',
            kind = 'count',
            palette = 'cubeHelix',
            data = df_clean).set(title = 'Répartition des catégories du label',
                                  xlabel = 'Label',
                                  ylabel = 'Nombre d\'observations')
```

```
[124]: <seaborn.axisgrid.FacetGrid at 0x1cf842559c8>
```



```
[125]: df_clean.shape
```

```
[125]: (1856, 6)
```

## Annexe 4 – PreProcessing

### 4 - ml\_cleaning

April 25, 2020

```
[1]: import pandas as pd
import re
import tldextract
import spacy
import nbprogress
import matplotlib.pyplot as plt
import nltk
import seaborn as sns

from wordcloud import WordCloud
from tqdm import tqdm, tqdm_notebook
from stop_words import get_stop_words
from nltk.tokenize import word_tokenize

[2]: # Import webscraped data
df = pd.read_json(r'data/webscraped_cleaned_2020-04-19-131331.json')
df = df[['title', 'body', 'sources', 'true_false']].reset_index(drop=True)
df.head(5)

[2]: title \
0 Un prêtre agressé à Avignon:
1 Francaisdefrance's Blog12/12/2016
2 Des militants dénoncent l'omerta des médias su...
3 Macron veut faire payer un loyer aux propriéta...
4 Macron a-t-il déclaré vouloir faire payer un l...

[2]: body \
0 action citoyenne, sécurité et citoyenneté Prêt...
1 Soyons solidaire de notre Église, A diffuser, ...
2 Consulter le journal Sur les sites de vente en ...
3 BuzzBeed Site à vocation parodique (dans la me...
4 First Draft and \nGoogle News Lab Travailler e...

[2]: sources true_false
0 https://www.over-blog.com/cookies FAUX
1 https://subscribe.wordpress.com/ FAUX
2 https://lemonde.fr/confidentialite/ VRAI
```

1

```

3     https://www.buzzbeed.com/sabonner      FAUX
4           https://firstdraftnews.com       VRAI

[3]: # Sources cleaning
df['sources'] = df['sources'].apply(lambda url: tldextract.extract(url).domain)

[4]: # Title cleaning
df['title_clean'] = df['title'].str.lower()
# Converts strings to lowercase
df['title_clean'] = df['title_clean'].str.replace(r'\%', ' pourcents')
# Converts the % to "percents"
df['title_clean'] = df['title_clean'].str.replace(r'http[s]?\\S+', ' ')
# Removes URLs
df['title_clean'] = df['title_clean'].str.replace(r'^ \\w+0-9', ' ')
# Removes all special characters
df['title_clean'] = df['title_clean'].str.replace(r'^\\.d{1}|\\W\\d{1}\\W', ' ')
# Removes digits isolated with whitespaces
df['title_clean'] = df['title_clean'].str.replace(r' +', ' ')
# Replaces 2+ whitespaces with single whitespace
df['title_clean'] = df['title_clean'].str.strip()
# Removes whitespaces and newlines at the beginning/end of strings
df.drop('title', axis=1, inplace=True)
df.head(5)

```

	body	sources	\
0	action citoyenne, sécurité et citoyenneté Prêt...	over-blog	
1	Soyons solidaire de notre Église, A diffuser, ...	wordpress	
2	Consulter le journal Sur les sites de vente en ...	lemonde	
3	BuzzBeed Site à vocation parodique (dans la me...	buzzbeed	
4	First Draft and \nGoogle News Lab Travailler e...	firstdraftnews	

	true_false	title_clean
0	FAUX	un prêtre agressé à avignon
1	FAUX	francaisdefrance s blog12 12 2016
2	VRAI	des militants dénoncent l omerta des médias su...
3	FAUX	macron veut faire payer un loyer aux propriéta...
4	VRAI	macron a t il déclaré vouloir faire payer un l...

```

[5]: # Body cleaning
df['body_clean'] = df['body'].str.lower()
# Converts strings to lowercase
df['body_clean'] = df['body_clean'].str.replace(r'\%', ' pourcents')
# Converts the % to "percents"
df['body_clean'] = df['body_clean'].str.replace(r'http[s]?\\S+', ' ')
# Removes URLs

```

```

df['body_clean'] = df['body_clean'].str.replace(r'\w+0-9', ' ')
    ↵#Removes all special characters
df['body_clean'] = df['body_clean'].str.replace(r'^\d{1}|\W\d{1}\W', ' ')
    ↵#Removes digits isolated with whitespaces
df['body_clean'] = df['body_clean'].str.replace(r' +', ' ')
    ↵#Replaces 2+ whitespaces with single whitespace
df['body_clean'] = df['body_clean'].str.strip()
    ↵#Removes whitespaces and newlines at the beginning/end of strings
df.drop('body', axis=1, inplace=True)
df.head(5)

```

```

[5]:      sources true_false \
0      over-blog      FAUX
1      wordpress      FAUX
2      lemonde        VRAI
3      buzzbeed       FAUX
4  firstdraftnews    VRAI

                           title_clean \
0      un prêtre agressé à avignon
1      francaisdefrance s blog12 12 2016
2  des militants dénoncent l omerta des médias su...
3  macron veut faire payer un loyer aux propriéta...
4  macron a t il déclaré vouloir faire payer un l...

                           body_clean
0  action citoyenne sécurité et citoyenneté prêtr...
1  soyons solidaire de notre église a diffuser il...
2  consulterle journal sur les sites de vente en ...
3  buzzbeed site à vocation parodique dans la mes...
4  first draft and google news lab travailler ens...

```

```

[6]: # Then, we concatenate all the text columns in a single column and make the
      ↵dataframe tidy
df['text'] = df[['title_clean', 'body_clean', 'sources']].agg(' '.join, axis=1)
df.drop(['title_clean', 'body_clean', 'sources'], axis=1, inplace=True)
df = df[df.columns[[1, 0]]]
df.head(5)

```

```

[6]:      text true_false
0  un prêtre agressé à avignon action citoyenne s...      FAUX
1  francaisdefrance s blog12 12 2016 soyons solid...      FAUX
2  des militants dénoncent l omerta des médias su...      VRAI
3  macron veut faire payer un loyer aux propriéta...      FAUX
4  macron a t il déclaré vouloir faire payer un l...      VRAI

```

```
[7]: # Here, we lemmatize the text
nlp = spacy.load("fr_core_news_sm", disable = "ner")

# Increase nlp memory
nlp.max_length = 2000000

# Instantiate tqdm
tqdm.pandas(tqdm_notebook)

# First, we apply the text processing unit on the text
df['lemmatized'] = df['text'].progress_map(lambda x: nlp(x))

# Then we retrieve the lemmas and concatenate the words back into sentences
df['lemmas'] = df['lemmatized'].apply(lambda x : " ".join([token.lemma_ for
    token in x]))
df.drop(['lemmatized', 'text'], axis=1, inplace=True)

# The resulting dataframe was saved to the temp_lemmas.json file
```

```
C:\Users\edaveau\AppData\Local\Continuum\anaconda3\lib\site-
packages\tqdm\std.py:666: FutureWarning: The Panel class is removed from pandas.
Accessing it from the top-level namespace will also be removed in the next
version
    from pandas import Panel
100%|     | 1856/1856 [08:49<00:00,  3.51it/s]
```

```
[6]: #df = pd.read_json (r'data/temp_lemmas.json')

# Instantiate tqdm
tqdm.pandas(tqdm_notebook)

# Import the stopwords from the stop_words package
stop_words_pack = get_stop_words('fr')

# Prepare a list of words we'll remove from the aforementioned stopwords
to_remove = ['aucun', 'autre','bon', 'comment', 'dedans', 'dehors', 'droite', 'd\'ébut', 'encore', 'force',
            'haut','hors', 'ici', 'juste', 'maintenant', 'moins', 'nom', 'nouveau', 'nouveaux', 'personne',
            'personnes','plupart', 'pourquoi', 'quand', 'que', 'quel', 'quelle', 'quelles', 'quels', 'qui',
            'sans', 'seulement', 'tandis', 'tellement', 'valeur', 'être']

# Remove these words from the list
stopWords = list(set(stop_words_pack)-set(to_remove))

# The stopwords can more easily be removed if we tokenize them
```

```

df["token"] = df.apply(lambda row: word_tokenize(row['lemmas']), axis=1)

# Then we keep a list of all the words not in the stopwords list
df["stop"] = df["token"].progress_map(lambda x: [word for word in x if word not in stopWords])

# We join everything back into sentences
df["text"] = df["stop"].progress_map(lambda x: " ".join(x))
df.drop(['lemmas', 'token', 'stop'], axis=1, inplace=True)
df.head(5)

df.to_json('data/lemmas.json')

```

100% | 1856/1856 [00:08<00:00, 210.03it/s]  
100% | 1856/1856 [00:00<00:00, 14098.25it/s]

```

[12]: df_length = pd.read_json(r'data/lemmas.json')

df_length['tokens'] = df_length['text'].apply(word_tokenize)
df_length['length'] = df_length['tokens'].apply(len)
df_length = df_length.sort_values('length', ascending = False).reset_index()

```

```

[36]: sns.set(style = 'whitegrid')

# Initialize the matplotlib figure
f, ax = plt.subplots(figsize = (11, 5))

# Plot the information sources
sns.set_color_codes('muted')
sns.lineplot(x = df_length.index.values.astype(int),
             y = df_length['length'],
             alpha=0.5).set_title('Nombre de mots par article après nettoyage', u
             .fontsize = 18)
ax.text(x=0.5, y=-0.15, s="Nb : L'axe \"y\" a été coupé à 20 000 mots, le plus
         .grand article faisant 120 311 mots",
         fontsize=11, alpha=0.75, ha='left', va='bottom', transform=ax.transAxes)

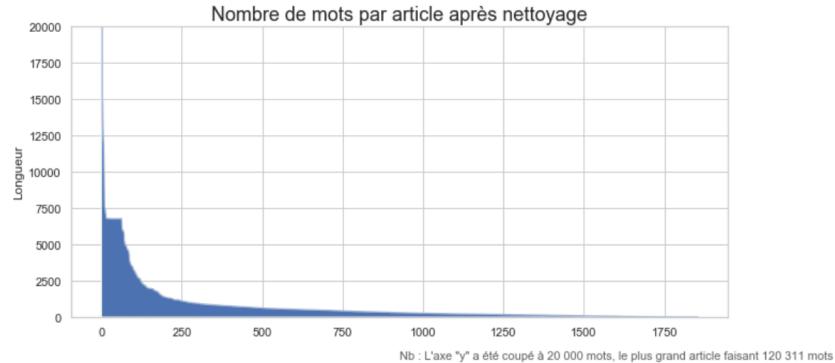
# Add an informative axis label
ax.set(ylabel = 'Longueur',
       xlabel = '')

# Adjust graphics
plt.fill_between(df_length.index.values, df_length.length.values)
plt.ylim(0, 20000)

del(df_length)

```

[36]: (0, 20000)



```
[42]: # Here, we'll remove rows with a word count < 20
df_ml = pd.read_json (r'data/lemmas.json')

df_ml['tokens'] = df_ml['text'].apply(word_tokenize)
df_ml['length'] = df_ml['tokens'].apply(len)
df_ml = df_ml[df_ml['length'] >= 15]
df_ml.drop(['tokens', 'length'], axis=1, inplace=True)
df_ml = df_ml.reset_index()

df_ml.to_json('data/df_machine_learning.json')
```

```
[45]: df = pd.read_json (r'data/df_machine_learning.json')

df_true = df[df['true_false'] == 'VRAI']
df_false = df[df['true_false'] == 'FAUX']

def print_freq_words(df) :
    articles = df.text.str.cat(sep=' ')
    #function to split text into word
    tokens = word_tokenize(articles)
    #create a set of unique words
    vocabulary = set(tokens)
    print('Vocabulary length of all the words used in the articles:', len(vocabulary))
    #calculate each word's frequency
    frequency_dist = nltk.FreqDist(tokens)
    #order each frequency in descending order, retain the first 50 elements
```

```

frequency_words = sorted(frequency_dist,key=frequency_dist.__getitem__,
↪reverse=True)[0:50]
frequency_words = ", ".join(word for word in frequency_words)
true_count = df['true_false'].str.contains('VRAI').sum()

if true_count > 0:
    print('\n50 words most commonly used in true articles:  

↪\n'+frequency_words+'\n')

else :
    print('\n50 words most commonly used in false articles:  

↪\n'+frequency_words+'\n')

print_freq_words(df_true)
print_freq_words(df_false)

```

Vocabulary length of all the words used in the articles: 16745

50 words most commonly used in true articles:  
être, que, qui, plus, c, s, pouvoir, autre, bien, site, france, dire, monde,  
faux, lire, al, celui, aller, voir, article, sans, non, falloir, jour, enfant,  
contre, an, personne, pourcent, français, nouveau, information, compte, savoir,  
vouloir, après, 2017, encore, passer, mettre, rien, seul, premier, vaccin,  
temps, moins, dernier, grand, publier, \_

Vocabulary length of all the words used in the articles: 53206

50 words most commonly used in false articles:  
être, que, qui, faux, plus, s, pouvoir, lire, c, site, explication, autre, bien,  
france, dire, détaillé, agir, al, article, vaccin, aller, an, celui, banque,  
monde, sans, information, non, voir, premier, nouveau, loi, contre, aucun, euro,  
pourcent, vouloir, falloir, jour, personne, moins, français, après, the, enfant,  
compte, détailler, savoir, mettre, politique

```
[46]: # Finally, we can create a WordCloud to see the words mostly used for true and
↪false articles
def generate_wordcloud(df) :
    articles = df.text.str.cat(sep=' ')
    tokens = word_tokenize(articles)
    frequency_dist = nltk.FreqDist(tokens)
    wordcloud = WordCloud().generate_from_frequencies(frequency_dist)
    plt.imshow(wordcloud)
    true_count = df['true_false'].str.contains('VRAI').sum()
    if true_count > 0:
        plt.title('WordCloud des articles vrais')
    else :
```

```

plt.title('WordCloud des articles faux')
plt.axis("off")
plt.show()

generate_wordcloud(df_true)
generate_wordcloud(df_false)

```



# Annexe 5 – Application de l’algorithme de Machine Learning

## 5 - Machine Learning

May 24, 2020

```
[3]: import pandas as pd
import nbprogress
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

from imblearn.over_sampling import SMOTE
from imblearn.pipeline import make_pipeline as make_pipeline_imb
from imblearn.under_sampling import NearMiss
from imblearn.metrics import classification_report_imbalanced
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, ↵
    TfidfTransformer
from sklearn.feature_selection import SelectKBest, chi2
from sklearn.metrics import classification_report, adjusted_rand_score, f1_score
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import make_pipeline, Pipeline
from sklearn.svm import LinearSVC
from sklearn.exceptions import ConvergenceWarning
from warnings import simplefilter

[5]: # Import the data
df = pd.read_json ('r'data/df_machine_learning.json')

# Select each independent Pandas series to a variable
Xtest = df["text"]
ytest = df["true_false"]

# Ignore a filter which will appear due to the loop below
simplefilter("ignore", category=ConvergenceWarning)

# Split training and test datasets
X_train, X_test, y_train, y_test = train_test_split(Xtest, ytest,test_size=0. ↵
    2,random_state=123)
```

```
[3]: # We create empty lists which we'll use later for a data visualization
iteration = []
f1_model = []

# We want to select the best K features for our model, so we run a loop
for i in nbprogress.log(range(10000), every = 100):
    # We are only interested in running this loop every 100 iteration (runtime/
    ↴100)
    if i % 100 == 0 and i != 0:
        # We create the pipeline with the tf-idf, features selection and
        ↴machine learning model
        pipeline = Pipeline([("vect", TfidfVectorizer(ngram_range = (1, 1))),
                            ("chi", SelectKBest(chi2, k = i)),
                            ("clf", LinearSVC())])
        # We apply this pipeline to our training data
        model = pipeline.fit(X_train, y_train)
        # We append the iterator value and the f1-score to our dataframe
        iteration.append(i)
        f1_model.append(f1_score(y_test, model.predict(X_test), average='macro'))

# We make a df out of our iterators
k_candidate = pd.DataFrame(
{
    'iteration': iteration,
    'f1_score': f1_model
})
)

# Then plot our k candidates
sns.set(style = 'whitegrid')

# Initialize the matplotlib figure
f, ax = plt.subplots(figsize = (11, 5))

# Plot the information sources
sns.set_color_codes('muted')
sns.lineplot(x = k_candidate['iteration'],
             y = k_candidate['f1_score'],
             alpha=0.5).set_title('Score Macro F1 en fonction du nombre de
             ↴features filtrés', fontsize = 18)

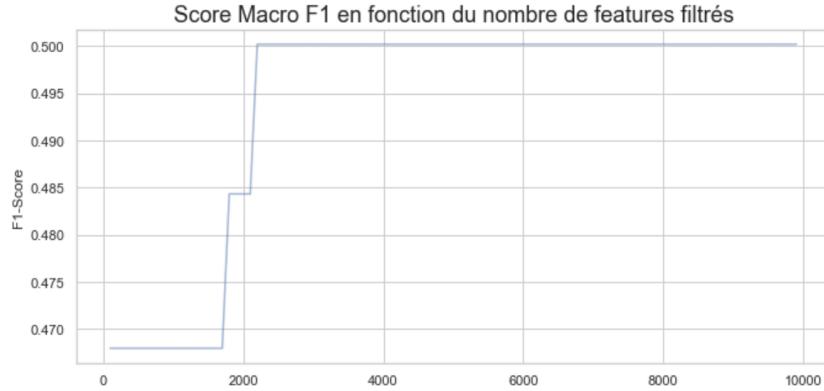
# Add an informative axis label
ax.set(ylabel = 'F1-Score',
       xlabel = '')

VBox(children=(HTML(value=''), IntProgress(value=0, max=10000)))

```

2

```
[3]: [Text(0, 0.5, 'F1-Score'), Text(0.5, 0, '')]
```



```
[28]: # Create a pipeline to get the model at 2000 features
pipeline = Pipeline([('vect', TfidfVectorizer(ngram_range = (1, 1))),
                     ('chi', SelectKBest(chi2, k = 2000)),
                     ('clf', LinearSVC())])

# Fit the model
model = pipeline.fit(X_train, y_train)

# Get our prediction
prediction = model.predict(X_test)

# Print out the report
clsf_report = pd.DataFrame(classification_report(y_true = y_test, y_pred = prediction,
                                                    output_dict=True)).transpose()
clsf_report
```

```
[28]:
```

	precision	recall	f1-score	support
FAUX	0.892128	1.000000	0.942989	306.000000
VRAI	1.000000	0.026316	0.051282	38.000000
accuracy	0.892442	0.892442	0.892442	0.892442
macro avg	0.946064	0.513158	0.497136	344.000000
weighted avg	0.904044	0.892442	0.844487	344.000000

```
[29]: pd.DataFrame(
    confusion_matrix(y_test, prediction, labels=['FAUX', 'VRAI']),
    index=['actual:FAUX', 'actual:VRAI'],
    columns=['predicted:FAUX', 'predicted:VRAI'])
```

```

        )
[29]:      predicted:FAUX  predicted:VRAI
actual:FAUX          306           0
actual:VRAI          37            1

[10]: # We create empty lists which we'll use later for a data visualization
iteration = []
f1_model = []

# We want to select the best K features for our model, so we run a loop
for i in nbprogress.log(np.arange(0.0, 10.0, 0.5)):
    if i != 0:
        # We create the pipeline with the tf-idf, features selection and
        # machine learning model
        pipeline = make_pipeline_imb(TfidfVectorizer(),
                                      SelectKBest(chi2, k = 2000),
                                      SMOTE(random_state = 123),
                                      LinearSVC(random_state = 123, C = i, class_weight =
                                      'balanced'))
        # We apply this pipeline to our training data
        model = pipeline.fit(X_train, y_train)
        # We append the iterator value and the f1-score to our dataframe
        iteration.append(i)
        f1_model.append(f1_score(y_test, model.predict(X_test), average='macro'))

# We make a df out of our iterators
k_candidate = pd.DataFrame(
    [
        {
            'iteration': iteration,
            'f1_score': f1_model
        }
    ]
)

# Then plot our k candidates
sns.set(style = 'whitegrid')

# Initialize the matplotlib figure
f, ax = plt.subplots(figsize = (11, 5))

# Plot the information sources
sns.set_color_codes('muted')
sns.lineplot(x = k_candidate['iteration'],
             y = k_candidate['f1_score'],
             alpha=0.5).set_title('Score F1 en fonction de l\'hyperparamètre
             C', fontsize = 18)

```

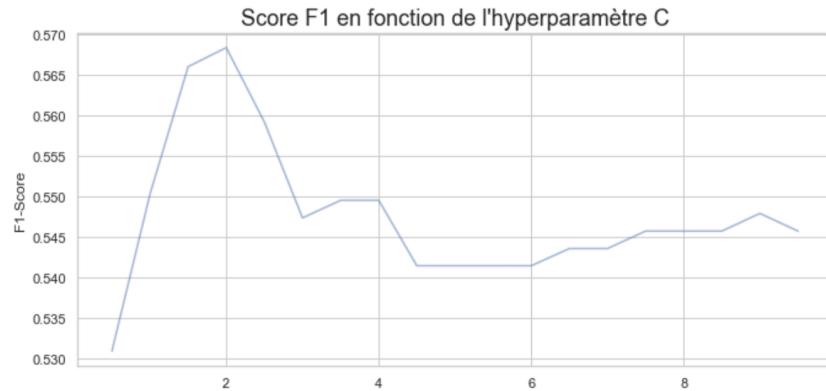
```

# Add an informative axis label
ax.set(ylabel = 'F1-Score',
        xlabel = '')

VBox(children=(HTML(value=''), IntProgress(value=0, max=20)))

```

[10]: [Text(0, 0.5, 'F1-Score'), Text(0.5, 0, '')]



```

[8]: # Create a pipeline to get the model at 2000 features with the SMOTE process
pipe = make_pipeline_imb(TfidfVectorizer(),
                         SelectKBest(chi2, k = 2000),
                         SMOTE(random_state = 123),
                         LinearSVC(random_state = 123, C = 1, class_weight = u
                         ↵'balanced'))

# Fit the model
model = pipe.fit(X_train, y_train)

# Get our prediction
prediction = model.predict(X_test)

# Print out the report
clsf_report = pd.DataFrame(classification_report(y_true = y_test, y_pred = u
                         ↵prediction, output_dict=True)).transpose()
clsf_report

```

[8]:

	precision	recall	f1-score	support
FAUX	0.902357	0.875817	0.888889	306.000000

```
VRAI      0.191489  0.236842  0.211765  38.000000
accuracy   0.805233  0.805233  0.805233   0.805233
macro avg   0.546923  0.556330  0.550327  344.000000
weighted avg  0.823831  0.805233  0.814090  344.000000
```

```
[9]: pd.DataFrame(
    confusion_matrix(y_test, prediction, labels=['FAUX', 'VRAI']),
    index=['actual:FAUX', 'actual:VRAI'],
    columns=['predicted:FAUX', 'predicted:VRAI']
)
```

	predicted:FAUX	predicted:VRAI
actual:FAUX	268	38
actual:VRAI	29	9

```
[32]: # Create a pipeline to get the model at 2000 features with the SMOTE process
# and the Naive Bayes algorithm
pipe = make_pipeline_imb(TfidfVectorizer(),
                         SelectKBest(chi2, k = 2000),
                         SMOTE(random_state = 123),
                         MultinomialNB())

# Fit the model
model = pipe.fit(X_train, y_train)

# Get our prediction
prediction = model.predict(X_test)

# Print out the report
clf_report = pd.DataFrame(classification_report(y_true = y_test, y_pred = prediction,
                                                 output_dict=True)).transpose()
clf_report
```

	precision	recall	f1-score	support
FAUX	0.906863	0.604575	0.725490	306.000000
VRAI	0.135714	0.500000	0.213483	38.000000
accuracy	0.593023	0.593023	0.593023	0.593023
macro avg	0.521289	0.552288	0.469487	344.000000
weighted avg	0.821678	0.593023	0.668931	344.000000

```
[33]: pd.DataFrame(
    confusion_matrix(y_test, prediction, labels=['FAUX', 'VRAI']),
    index=['actual:FAUX', 'actual:VRAI'],
    columns=['predicted:FAUX', 'predicted:VRAI']
)
```

[33] :                   predicted:FAUX   predicted:VRAI  
actual:FAUX               185               121  
actual:VRAI               19               19

<https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2>           <https://towardsdatascience.com/class-imbalance-a-classification-headache-1939297ff4a4>       <https://stackoverflow.com/questions/55740220/macro-vs-micro-vs-weighted-vs-samples-f1-score>