Naïla EL HAOUARI

ENSAE 3ème année

Stage de fin d'études

Année scolaire 2019-2020

Analyse des symptômes autodéclarés du COVID19 sur Twitter

CRI Paris

8 bis rue Charles V

75004 PARIS

Maître de stage : **Marc Santolini**

Du 12/03/2020 au 09/10/2020

Table des matières

Remerciements

Bien que j'aie débuté et effectué une grande partie de mon stage dans les conditions peu optimales qu'étaient celles du confinement, j'ai appris énormément de cette expérience, aussi bien sur des sujets techniques que sur moi-même. Et, si les premiers mois étaient très étranges en raison du manque d'interactions sociales et du fait que je n'ai pu rencontrer la plupart des membres de mon équipe qu'à partir de l'été, je suis extrêmement reconnaissante envers Marc et Sam pour leur aide précieuse, leur enthousiasme, et leur accompagnement tout au long mon stage.

Mes remerciements vont également au reste de l'équipe et aux personnes du CRI avec lesquelles j'ai pu collaborer et interagir le plus, notamment Bastian et Lionel, qui m'ont beaucoup aidée surtout d'un point de vue technique et permis d'être davantage efficace dans mes projets. Je suis également très ravie d'avoir pu échanger avec une grande partie des membres du CRI, et d'avoir pu profiter de leur bonne compagnie au cours de l'été.

Enfin mon stage au CRI m'a permis de découvrir plus directement le champ des sciences sociales computationnelles, de travailler sur des thématiques qui m'intéressent particulièrement, et de réfléchir sur mon orientation et la façon dont je vois mon avenir, ce dont j'avais particulièrement besoin en fin d'études.

1 Introduction

1.1 L'interdisciplinarité au sein du CRI Paris

Le CRI Paris – Centre de recherche interdisciplinaire – est un institut au sein duquel se côtoient des acteurs au parcours divers, dont l'objectif est d'explorer de nouvelles façons d'apprendre, enseigner, et faire de la recherche. Le département de recherche, qui est associé à l'Université de Paris, héberge des équipes travaillant sur des projets à la frontière entre différentes disciplines traditionnelles, principalement autour de la biologie et des mathématiques, mais également en sciences des réseaux, sciences de l'éducation ou science citoyenne (*citizen science*).

Dans le cadre de mon stage, j'ai rejoint l'équipe *Interaction Data Lab*, dirigée par Marc Santolini. En dehors des collaborateurs extérieurs, elle est composée de cinq personnes qui travaillent sur l'étude des interactions sociales et des effets de réseaux, notamment sur la façon dont s'organisent des communautés (compétitions iGEM, collaborations de chercheurs sur arXiv), en adoptant une approche fondée sur les données.

1.2 Projet de stage

J'ai débuté mon stage le 12 mars 2020 à temps partiel, en même temps que la fin de ma troisième année à l'ENSAE ; et j'ai poursuivi à temps plein à partir de mai 2020. Mon sujet initial de stage était l'analyse des réseaux de radicalisation sur Twitter, en se focalisant sur l'étude des réseaux djihadistes. Ce projet est co-encadré par Samuel Fraiberger, computational social scientist à la Banque Mondiale et à New York University, et Hugo Micheron, sociologue à l'ENS Ulm spécialisé sur ces thématiques. Cependant, en raison des retards administratifs (le projet est financé par l'ENS) et de la situation sanitaire, je n'ai pas pu encore travailler sur ce sujet. J'ai alors commencé un autre projet avec Marc et Samuel, sur l'analyse des symptômes auto-déclarés du COVID-19 sur Twitter. Cette étude, dont l'objectif était initialement de travailler sur des données Twitter pour apprendre à les collecter et les manipuler, a fini par être mon sujet principal en raison des résultats convaincants que nous avons obtenus ; nous sommes par ailleurs actuellement en train de rédiger un article dans lequel nous développons notre méthodologie.

Dans le cadre de l'évolution très actuelle de l'épidémie de COVID-19, qui a

provoqué un bouleversement intense de l'organisation de nos sociétés, nous nous demandions initialement si l'on observe une dynamique d'attention précédent l'incidence de la maladie en France. Nous avons travaillé sur le réseau social Twitter à partir de tweets collectés via l'API, provenant d'utilisateurs en Île-de-France, et nous nous sommes intéressés plus spécifiquement aux tweets se référant à des symptômes du COVID-19. Dans un premier temps, avec cette approche fondée sur des mots-clefs, nous avons pu observer une corrélation importante entre le nombre de tweets mentionnant ces symptômes et le nombre de personnes admises aux urgences pour COVID-19. Cependant, en regardant plus attentivement les tweets, notre signal est brouillé par un certain nombre de "faux positifs" : certains tweets mentionnant des symptômes ne se réfèrent en réalité pas aux symptômes de l'auteur du tweet, mais à des mesures d'hygiène générales, ou à des blagues. Pour nettoyer le signal, nous avons exploré diverses solutions, telles que l'ajout d'autres règles ; et avons finalement adopté une approche liée à la science citoyenne (*citizen science*), où l'on demande à des volontaires d'annoter nos tweets pour améliorer à notre projet.

2 Motivation et données

2.1 Recherche sur le sujet

Depuis quelques années et avec l'utilisation massive et croissante des outils numériques, s'est généralisée la collecte et l'étude des données disponibles sur les réseaux sociaux, aussi bien à des fins commerciales, politiques, ou au sein de projets de recherche. Avec la démocratisation des réseaux sociaux, les utilisateurs sont alors parfois très descriptifs et abordent des sujets personnels de façon publique. Une étude de Naaman, Boase et Lai 2010 mettait notamment en évidence la diversité des types de messages sur le réseau social Twitter : les auteurs catégorisent les tweets en diverses catégories, dont les plus fréquentes sont le partage d'informations, le partage d'opinions, des propos et pensées assez aléatoires (sur la météo par exemple), et une catégorie nommée "Me now", regroupant des tweets où l'utilisateur décrit son état actuel ("Je suis fatigué", "je passe un bon moment chez X",...). Ce partage de données personnelles, qui peuvent être parfois aussi précises que l'expression d'opinions sur des sujets politiques ou sociétaux, ou la description de l'état de santé de l'utilisateur, peut alors permettre de prendre connaissance de ces informations de façon assez immédiate.

Ces derniers mois ont vu un bouleversement radical dans l'organisation de nos sociétés. L'évolution rapide de la pandémie de COVID-19 a forcé les gouvernements à prendre des mesures inédites pour ralentir la progression du virus, et accélérer la recherche sur le sujet. Dans le cadre d'épidémies précédentes, l'étude des données d'utilisateurs collectées grâce à leur utilisation d'Internet se sont révélés être potentiellement utiles pour prédire et évaluer l'incidence de maladies. L'un des outils qui s'est fait le plus connaître est celui mis en place par Google, le *Google Flu Trends*, documenté dans l'article de Ginsberg et al. 2009, qui visait à étudier, entre 2008 et 2015, l'incidence de la grippe saisonnière aux Etats-Unis. Les données étaient collectées à partir des recherches du moteur de recherche Google. Si cet outil s'était avéré être relativement précis un certain temps, son efficacité a décru au cours du temps, en raison d'un méthodologie erronée : le principe était d'utiliser comme termes du moteur de recherche les expressions qui matchaient le mieux avec les données médicales, même si celles-ci n'avaient aucun lien avec la grippe, comme documenté dans l'article de Lazer et al. 2014.

Les prédictions de l'évolution de maladies à partir de données de moteurs de recherche peuvent également être biaisées par le fait qu'une augmentation des recherches à propos d'une maladie ou de ses symptômes ne sont pas uniquement causées par une augmentation du nombre de personnes infectées, mais potentiellement par un soudain intérêt pour ce sujet. Les données collectées sur les réseaux sociaux comme Twitter présentent alors l'avantage d'être suffisamment descriptives pour pouvoir repérer les tweets liés uniquement à des personnes malades. Chew et Eysenbach 2010 s'intéressaient alors au contenu des tweets liés à l'épidémie de grippe H1N1 en 2009, et en dressaient une typologie : si certains tweets mentionnent des informations, des opinions sur le sujet, ou des blagues, d'autres se rapportent à du vécu personnel.

Plus récemment, dans le cas de la pandémie de COVID-19, d'autres analyses ont déjà été faites sur les tweets liés à la pandémie et aux symptômes. Sarker et al. 2020 analysent des tweets d'utilisateurs affirmant sur le réseau social avoir été testés positifs au COVID-19, et rapportent les symptômes les plus fréquemment mentionnés. Les auteurs ont annoté manuellement les profils d'utilisateurs et leurs tweets afin de s'assurer de filtrer les faux positifs – des individus qui écrivent "testé positif", alors qu'il n'ont pas été malades eux-mêmes. Une autre étude, menée par Mackey et al. 2020, se concentrait sur les tweets mentionnant directement des symptômes du COVID-19. En utilisant un modèle de *topic analysis* appelé *biterm topic model*, un modèle d'apprentissage non supervisé, ils distinguent différents groupes de tweets mentionnant des symptômes. Ils ont annoté manuellement plusieurs tweets des différents clusters obtenus afin de n'analyser que les tweets se rapportant aux symptômes, à la guérison, ou aux tests liés au COVID-19. Ils distinguent alors plusieurs clusters de tweets : des tweets mentionnant les symptômes de l'utilisateur ou de l'un de ses proches, des tweets d'un utilisateur attendant le résultat de son test, ou encore des tweets mentionnant des symptômes que l'utilisateur aurait eu plusieurs semaines auparavant.

Ces études soulèvent alors la richesse de l'information présente sur les réseaux sociaux, et en particulier Twitter, pour travailler sur un sujet d'étude de la progression de la maladie ; et elles mettent également en évidence la grande diversité des tweets, et la nécessité d'adopter des méthodes de filtrage, qu'elles soient automatiques ou (bien souvent) manuelles, pour être sûrs de n'étudier que les tweets directement liés à du vécu personnel et à une personne rapportant ses propres symptômes.

Dans mon projet de stage, je me suis concentrée sur l'analyse des tweets en Île-de-France. Après avoir débuté une analyse par mots clefs pour identifier les tweets pouvant se rapporter à des symptômes, nous avons finalement reconnu l'importance de l'annotation manuelle pour identifier précisément les tweets se rapportant à des symptômes autodéclarés.

2.2 Récupération des données via l'API Twitter

Les données sur lesquelles j'ai réalisé mes analyses ont été récupérées grâce à l'API Twitter [1]. Avec la création d'un compte développeur, il est possible de collecter gratuitement des tweets publics, dans certaines limites. Nous avons restreint notre analyse à la région parisienne pour plus de facilités. L'objectif était tout d'abord d'identifier des utilisateurs localisés en région parisienne, et de collecter ensuite leur historique.

Les données ont été collectées sur Python avec *tweepy*, un outil Python pour s'identifier sur l'API Twitter, qui nous renvoie les tweets sous format json. Dans un premier temps, nous avons utilisé la requête *Streaming* [2], qui permet d'écouter Twitter et de collecter des tweets émis en temps réel. La limite de cette requête est que l'on ne collecte qu'un petit échantillon du flux total de tweets (environ 1%). Nous avons alors utilisé cette requête pendant plusieurs semaines, entre mars et mai 2020, en utilisant un filtre géographique ciblé sur l'Île-de-France, afin de ne collecter que des tweets émis dans cette région. Cela nous a permis d'identifier 30 651 utilisateurs, localisés en Île-de-France.

Nous avons ensuite collecté, pour chacun de ces utilisateurs, leurs données historiques, c'est-à-dire les tweets qu'ils avaient écrits précédemment, grâce à la requête *Get tweets timelines* [3]. Cette requête permet d'obtenir, pour chaque utilisateur dont on dispose de l'*user id*, les 3 600 derniers tweets. L'une des limites de cette requête est la présence de *rate limits*, qui nous empêchent de collecter autant de tweets que souhaités par jour. Nous avons détourné dans une certaine mesure cette limite en obtenant des *token pairs* de la part d'autres utilisateurs Twitter (des amis ayant un compte Twitter) qui s'abonnent alors sur mon application, ce qui m'a permis

1. La documentation de l'API est disponible ici : https://developer.twitter.com/en/docs
2. https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/introduction
3. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-user$_timeline$

d'effectuer davantage de requêtes[4]. Par la suite, on a régulièrement relancé cette requête pour collecter, au fur et à mesure du temps, les tweets les plus récents de nos utilisateurs localisés en Île-de-France. La dernière requête a été lancée le 1er septembre 2020, date alors de nos tweets les plus récents.

En utilisant ces deux requêtes, on a alors collecté un total de 46 116 543 tweets ; on présente sur la Figure 1 un schéma résumant la structure de la collecte des tweets.

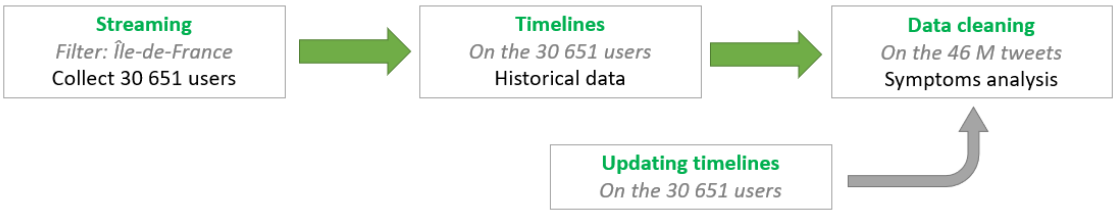


FIGURE 1 – Workflow de la collecte des tweets via l'API Twitter

De cette façon, on ne se concentre pas sur le flux total de Twitter, mais l'on suit les mêmes 30 000 individus au cours du temps. S'ils ont été identifiés en Île-de-France entre mars et mai 2020, on faisait initialement l'hypothèse que ces utilisateurs n'ont pas changé d'endroit, étant donné le confinement en France. Cependant, cette hypothèse peut ne pas être vérifiée après mai 2020 ; d'où la nécessité, pour la suite, de revérifier notre jeu de données.

2.3 Preprocessing et cleaning

En raison du grand volume de données, on a effectué les premières analyses en pyspark. Spark est un outil où les fonctions utilisées sont automatiquement parallélisées, ce qui permet de pouvoir traiter rapidement l'ensemble des données.

Sur les données obtenues, nous avons commencé par préprocesser et nettoyer les données. Dans un premier temps, nous avons fait le choix de ne conserver que les tweets écrits en français, et de retirer les retweets (tweets débutant par "RT") – notre analyse portant sur l'analyse des symptômes, on cherche en priorité à analyser des tweets originaux et traitant du vécu personnel. On choisit également de ne conserver que les tweets après le 1er décembre 2019. Cela réduit la taille de notre échantillon à 12 505 593 tweets.

4. https://developer.twitter.com/en/docs/authentication/oauth-1-0a/obtaining-user-access-tokens

Par la suite, et puisque notre approche se base principalement sur des mots-clefs, nous réécrivons les tweets en minuscules et les anonymisons, en identifiant les url et les mentions des utilisateurs (débutant par "@"). Enfin, puisque le language sur Twitter est souvent familier et équivalent à du language "parlé", on retire tous les accents ("é" devient "e"), et les marques de ponctuation, afin de rendre l'identification de termes plus aisée.

3 Des analyses fondées sur des mots-clefs

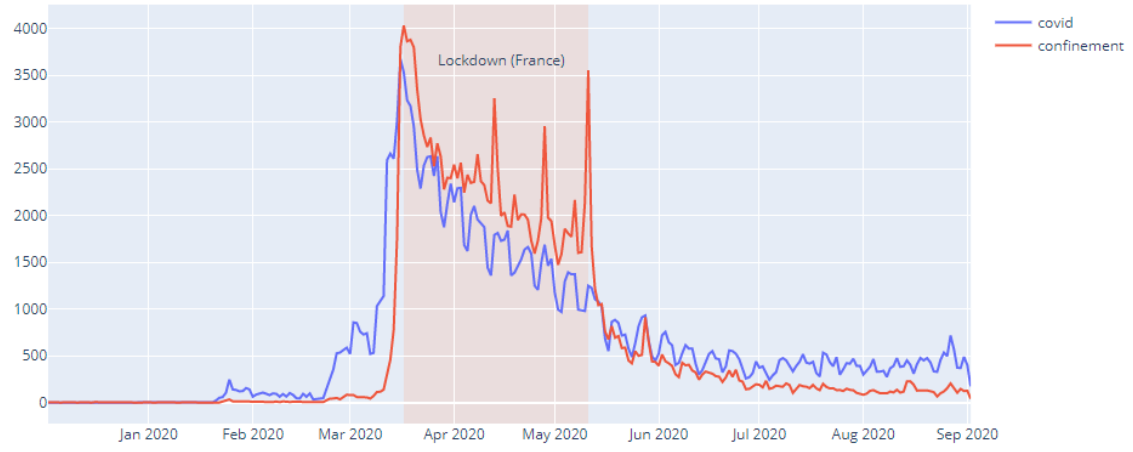
3.1 Analyse des termes liés au COVID

Nous avons effectué nos premières analyses sur les tweets en créant des listes de mots-clefs se rapportant à des expressions associées à l'épidémie. L'objectif est d'étudier dans un premier temps, sans distinguer les différents types de tweets, l'évolution de la mention de ces expressions sur le réseau social.

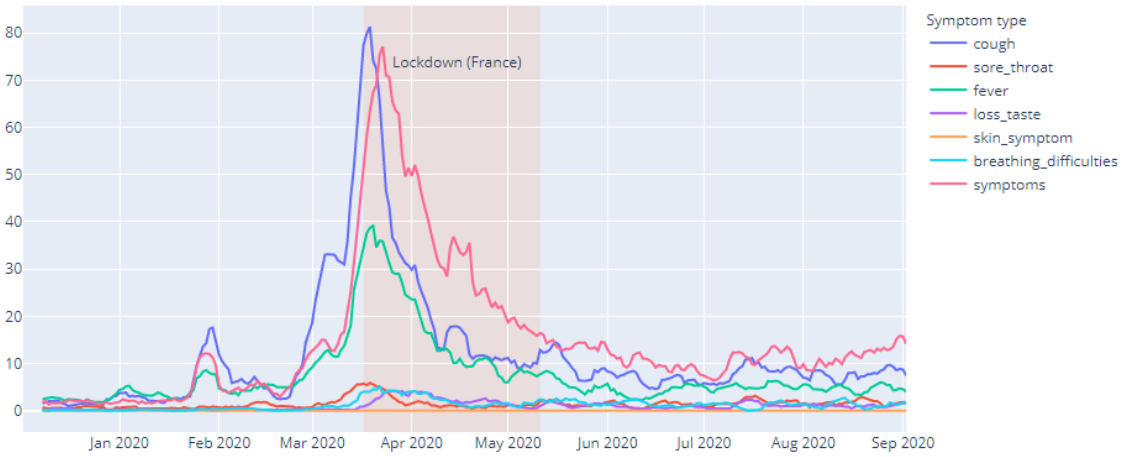
On s'intéresse tout d'abord aux tweets mentionnant directement des termes liés au COVID-19 et à la pandémie. Depuis le 1er décembre 2019, 187 612 tweets mentionnent des expressions liées au COVID-19 (nous avons inclus "covid", "coronavirus" et "corona ") ; et 174 582 tweets mentionnent le confinement ("confin", "quarantaine"). En étudiant l'évolution temporelle de ces mentions, sur la Figure 2a, on observe que les deux courbes sont très corrélées : le coefficient de corrélation de Pearson vaut 0,90. Les premières mentions de COVID apparaissent fin janvier, avec la médiatisation de l'évolution de l'épidémie, et la détection des premiers cas en France. C'est surtout quelques jours avant le début du confinement, le 17 mars 2020, que le nombre de ces mentions explose (jusqu'à 4000 tweets par jour), avant de redescendre et de se stabiliser à partir de juillet 2020. On observe également une saisonnalité de sept jours, car les individus tweetent probablement moins le week-end.

Nous nous intéressons également aux tweets mentionnant des symptômes du COVID-19. On a commencé par créer manuellement une liste de mots-clefs et expressions se rapportant aux symptômes du COVID-19, référencée dans la Table 1. L'objectif est de capter les tweets dans lesquels l'auteur mentionne ses propres symptômes, ou des symptômes d'un proche ; nous avons donc inclus dans cette liste des expressions familières pour désigner certains symptômes ("mal à la gorge", "perdu le goût"). Depuis décembre 2019, 9 917 tweets mentionnent au moins un symptôme.

On représente sur la Figure 2b le nombre tweets mentionnant ces symptômes par jour, moyennés sur une semaine, pour plus de visibilité en raison du grand nombre de courbes. Les expressions les plus mentionnées sont celles liées à la toux, au terme "symptôme", et à la fièvre. Comme pour l'évolution des termes liés au COVID, on observe une augmentation des expressions liées aux symptômes fin janvier, au moment de la médiatisation de l'évolution de l'épidémie ; puis, la mention de ces termes augmente très fortement quelques jours avant le confinement, avant de décroître et de se stabiliser.



(a) COVID et confinement



(b) Symptômes

FIGURE 2 – Evolution de la mention d'expressions liées au COVID

Sont représentés le nombre de tweets par jour mentionnant des termes liés au COVID, en Île-de-France. Pour le nombre de symptômes par jour, les courbes ont été moyennées par semaine pour plus de visibilité.

Symptom	Symptom (FR)	Keywords (FR)
cough	toux	touss, toux
sore throat	maux de gorge	mal a la gorge, mal de gorge, maux de gorge
fever	fièvre	fievre, de la temperature
breathing di□culties	di□cultés respiratoires	di□cultes respiratoires, di□cultes a respirer, di□culte a respirer, mal a respirer
loss of taste	perte du goût et de l'odorat	perte du gout, perte de l odorat, perte de lodorat, perdu le gout, perdu l odorat, perdu lodorat, plus de gout, plus d odeur, plus dodeur
skin symptoms	symptômes cutanés	engelures
symptom	symptôme	symptom

TABLE 1 – Dictionnaire des expressions se rapportant aux symptômes du COVID-19

3.2 Comparaison avec les données médicales

L'objectif de ce projet étant d'étudier dans quelle mesure les tweets mentionnant des symptômes peuvent permettre de prédire l'évolution de l'épidémie, on a choisi d'utiliser des données publiques fournies par Santé Publique France pour faire une comparaison. On a choisi de ne pas utiliser le nombre de cas détectés, étant donné que celui-ci est très dépendant du nombre de tests effectués. On a utilisé les données des urgences hospitalières et de SOS médecins relatives à l'épidémie de COVID-19 [5], plutôt que les données relatives aux hospitalisations [6] car cette base comporte des données plus anciennes que la seconde (depuis début mars 2020). Les données de ces deux bases sont quoi qu'il en soit très corrélées. Les données de Santé Publique France sont actualisées tous les jours, et sont disponibles par département. On se concentre alors sur les données d'Île-de-France.

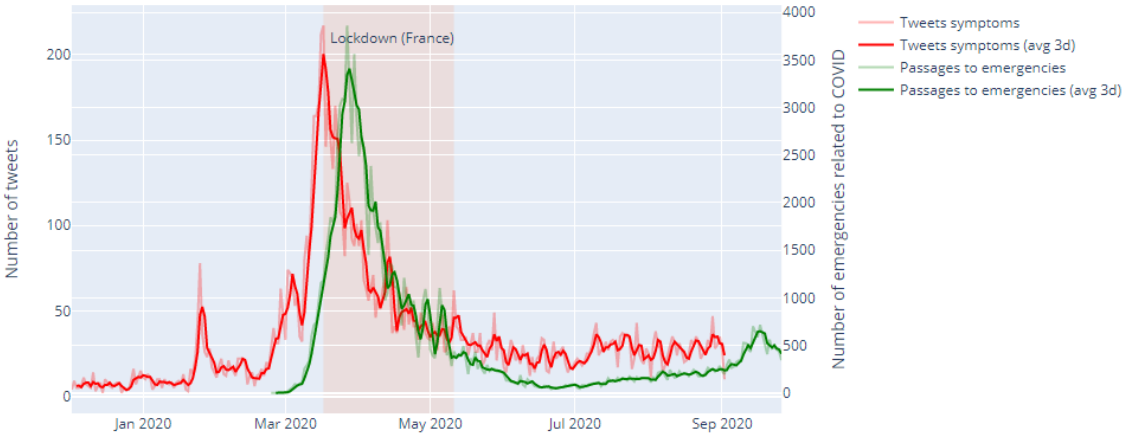
On compare le nombre de tweets mentionnant au moins un symptôme, avec le nombre de passages aux urgences pour suspicion de COVID-19 – c'est-à-dire, des personnes ayant des symptômes et se rendant aux urgences, ce qui permettrait d'es-

5. https://www.data.gouv.fr/fr/datasets/donnees-des-urgences-hospitalieres-et-de-sos-medecins-relatives-a-lepidemie-de-covid-19/

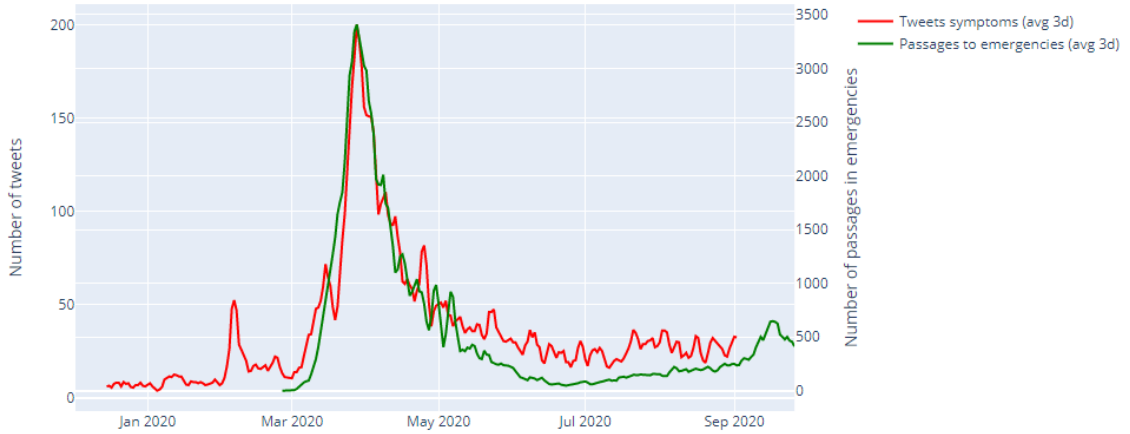
6. https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/

timer l'incidence de la pandémie sur la population. Sur la Figure 3a, on observe que les courbes du nombre de tweets mentionnant des symptômes et du nombre de passages aux urgences sont très similaires : elles augmentent très rapidement quelques jours avant le début du confinement, et diminuent plus progressivement ensuite, avec un écart visiblement de quelques jours – ce qui semble cohérent avec l'idée selon laquelle les individus auraient d'abord des symptômes (et les mentionneraient potentiellement sur les réseaux sociaux), avant de se rendre aux urgences si leur état de santé se dégrade.

On a représenté sur la Figure 9 en annexe la cross-corrélation entre les deux séries temporelles de nombre de tweets mentionnant des symptômes, et de passages aux urgences, moyennés sur 3 jours ; on constate que les deux séries temporelles sont très corrélées, avec un lag d'autour 10 jours. Dès lors, lorsqu'on représente une nouvelle fois ces deux courbes mais en décalant la courbe du nombre de tweets de 11 jours, comme sur la Figure 3b, celles-ci se superposent, du moins lors de la période entre mars et mai : l'augmentation et la diminution des mentions sur Twitter suivent la même évolution, et au même rythme, que l'évolution des passages aux urgences, et le coefficient de corrélation de Pearson entre les deux courbes vaut alors 0,96. Il semblerait alors que les mentions des symptômes sur Twitter puissent permettre de prédire dans une certaine mesure l'évolution de la pandémie.



(a) Nombre de tweets et de passages aux urgences par jour, et moyennés sur 3 jours



(b) Nombre de tweets décalés de 11 jours de passages aux urgences moyennés sur 3 jours

FIGURE 3 – Evolution du nombre de tweets mentionnant des symptômes et des passages aux urgences pour suspicion de COVID-19

Notons cependant quelques limites de notre approche. D'une part, si les deux courbes se superposent entre mars et mai, cette tendance n'est pas toujours respectée. On observe un premier pic dans le nombre de tweets le 24 janvier 2020, ce qui correspond en réalité à la date des premiers cas officiels de COVID-19 en France ; ces tweets mentionnant des symptômes témoigneraient probablement de nouvelles ou recommandations globales, plutôt que de réels cas de contaminations, puisque le nombre de tweets diminue juste ensuite. En effet, en regardant de plus près notre jeu de données et en étudiant manuellement des tweets mentionnant des symptômes, on observe que nombreux sont ceux dans lesquels les auteurs ne mentionnent pas leurs propres symptômes, mais partagent les mesures d'hygiène à adopter, ou font des blagues sur le sujet. De nombreux tweets dans nos données sont alors des "faux-positifs", que l'on souhaiteraient filtrer, et qui brouillent notre analyse. Cela pourrait également expliquer la raison pour laquelle le nombre de tweets ne diminue pas autant que le nombre de passages aux urgences, à partir de juin : s'il représentait réellement un indicateur du nombre de cas de COVID-19, on devrait s'attendre à ce qu'il soit presque aussi faible en juin-juillet qu'en janvier 2020, mais il stagne à un niveau plus élevé.

Par ailleurs, on voudrait isoler l'effet seul de la mention des symptômes pour maladie de la dynamique d'attention. Si peu de tweets mentionnent des symptômes en décembre et commencent à en mentionner beaucoup en mars, on peut supposer que cela est dû à la médiatisation de l'évolution de la pandémie et à l'importance

des mesures sanitaires prises ; certaines études font en effet état d'une circulation du virus avant les premiers cas officiels [7], ce qui expliquerait l'augmentation soudaine des hospitalisations. On peut également se demander si la stabilisation du nombre de tweets mentionnant des symptômes à un niveau assez élevé, en juillet, est aussi due à un changement dans les comportements des individus, qui s'exprimeraient plus spontanément sur les réseaux sociaux dès les premiers symptômes.

Enfin, notre analyse se base en réalité sur un nombre assez restreint de tweets, et sur une région très spécifique. Il conviendrait également de travailler sur un jeu de données plus grand. Si le jeu de données que j'ai utilisé lors de mon stage est assez restreint, mon co-superviseur de stage, Samuel Fraiberger, dispose d'un jeu de données assez conséquent : 1,5 million d'utilisateurs en France, et des données dans d'autres pays comme les Etats-Unis ou le Mexique. Ses données appartiennent cependant à la Banque mondiale, il ne peut pas me les partager directement. L'objectif est d'essayer de voir si l'on arrive à tirer des conclusions dans un premier temps sur mon jeu de données, puis d'extrapoler les analyses à plus grande échelle au reste de la France, et dans d'autres pays.

7. https://www.fondation-diaconat.fr/images/Presse/2020/CP-HAS-Imagerie-mdicale-7-mai-2020.pdf

4 Une approche en citizen science

4.1 Motivations

4.1.1 Présence de faux positifs

L'une des principales limites de notre analyse concerne la netteté du signal, qui est brouillé par un grand nombre de faux positifs. En effet, on cherche à se concentrer sur les tweets dont les auteurs témoignent de leurs propres symptômes ou de symptômes de leurs proches ; mais un grand nombre de ces tweets évoquent à l'inverse des règles sanitaires à respecter, des blagues, ou encore des individus mentionnant des symptômes qu'ils ont eu quelques mois plus tôt.

Initialement, nous avons tenté de régler ce problème en incluant davantage de règles pour filtrer nos tweets automatiquement : on retirait de notre base les tweets qui contenaient un lien url, et on ne conservait que ceux qui contenaient un pronom ("je", "ma", "mon", etc.). Cependant, ces règles ne permettent pas d'améliorer de façon significative notre jeu de données et de filtrer ces faux positifs. La solution finalement retenue a été d'adopter une approche de *citizen science*, en demandant à des individus bénévoles et intéressés par aider notre projet d'annoter volontairement des tweets, afin de retirer les faux positifs.

4.1.2 Recherche sur le sujet

Le recours au *crowdsourcing*, ou *citizen science*, consiste à demander à des volontaires, au travers d'une interface, d'annoter des données, bien souvent pour des projets pour lesquels il n'existe pas de jeux de données déjà labellisés ou suffisamment proches. L'un des outils les plus connus est celui d'Amazon Mechanical Turk [8], où la plateforme d'annotation est déjà pré-construite, et les annotateurs y sont payés à la tâche. Ce type de mécanismes est alors souvent utilisé pour labelliser des données Twitter : Finin et al. 2010 y ont notamment recours pour leur projet de *named entity recognition* sur des données de Twitter, car les modèles de NER déjà existants étaient entraînés sur des textes bien plus longs, et donc très différents de la structure des tweets qui eux ne contiennent que 140 caractères.

Il convient de noter les controverses liées à l'utilisation de ce types de ressources. Les annotateurs sont en général des individus payés à la tâche dans le cas des

8. https://www.mturk.com/

outils tels que Amazon Mechanical Turk, ou travaillent gratuitement dans le cas de plateformes de *crowdsourcing* originales. Un grand nombre des critiques à l'égard de ces platerformes, telles que celles émises par Fort, Adda et Cohen 2011, mettent en évidence les conditions de travail de ces "microtravailleurs" : leur rémunération est souvent inférieure au montant du salaire minimum, et ils n'ont pas les mêmes accès au système de protection sociale que des salariés. S'il est également possible de créer soi-même une plateforme de *crowdsourcing* et de demander à des volontaires – non rémunérés – d'aider à exécuter des tâches, ces types de plateformes présentent également des enjeux éthiques, du fait de la non-rétribution des volontaires et du manque de crédit parfois accordé aux contributeurs.

Une alternative à ces approches de *crowdsourcing* traditionnelles est la création de projets de *citizen science* dans lesquels le travail effectué par les volontaires est rendu accessible dès le début, ce qui leur permet de participer directement au projet. Wikidata, documenté dans l'article de Vrandečić 2012, peut être considéré comme un exemple de ce type de projets, qui permet la réutilisation des données.

Dans le cadre de notre projet et pour labelliser nos données, nous avons choisi de construire nous-mêmes une plateforme de *crowdsourcing*, avec l'aide d'un développeur du CRI. Au travers de cette plateforme, nous donnons la possibilité aux participants de télécharger s'ils le souhaitent les données – anonymisées –, afin de prendre part s'ils le souhaitent au projet ; et nous représentons graphiquement les résultats des annotations, afin de montrer aux volontaires l'impact de leur participation.

4.2 Plateforme de crowdsourcing

4.2.1 Présentation de la plateforme

Avec l'aide de Bastian Greshake Tzovaras, un chercheur au CRI, nous avons mis en place une plateforme de *crowdsourcing* permettant à des individus volontaires de labelliser nos données. Cette plateforme, accessible aussi bien sur smarphone ou sur ordinateur, a été lancée le 3 août 2020, et est accessible sur ce lien : `https://covid-twitter.thecommons.science/` On représente en Figure 4 une capture d'écran de la plateforme d'annotation.

L'utilisateur arrive sur une plateforme d'annotation, sur laquelle on lui demande de répondre si le tweet présenté contient des symptômes du COVID-19 autodécla-

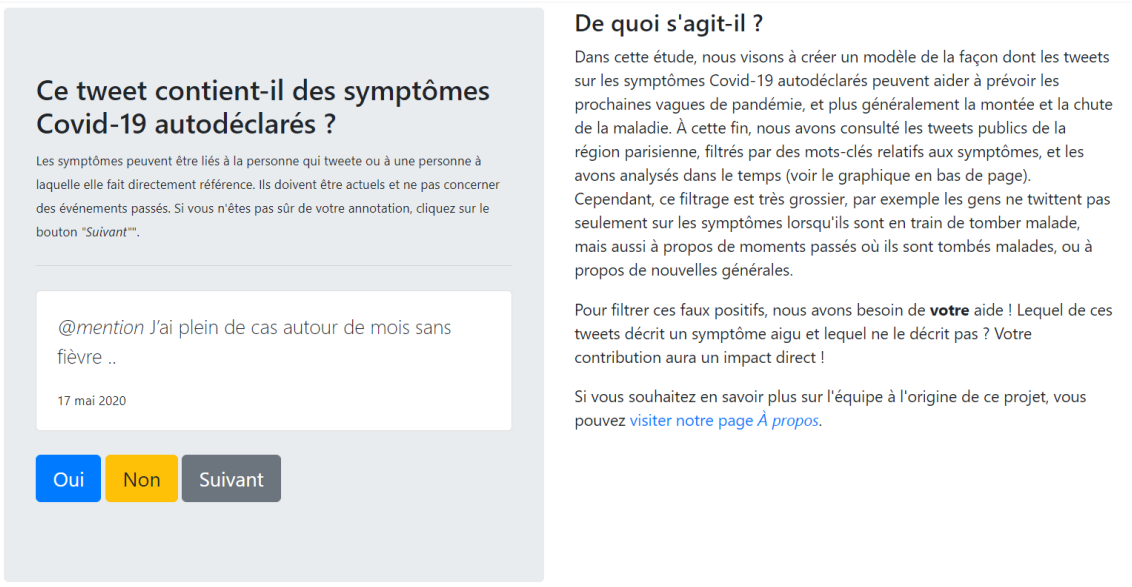


FIGURE 4 – Présentation de la plateforme d'annotation des tweets

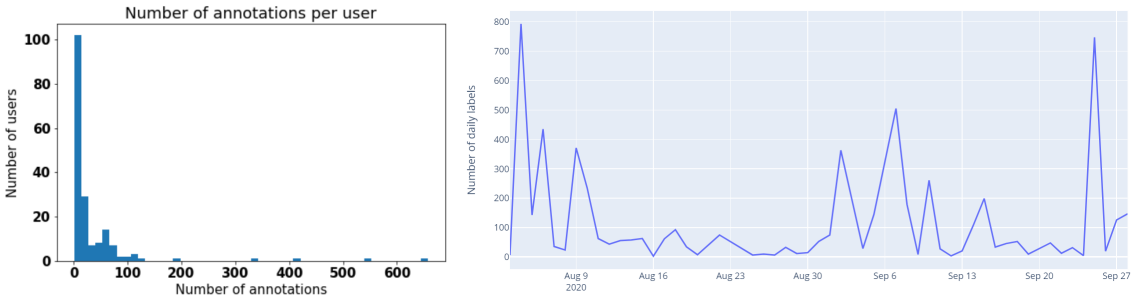
rés, en précisant que l'on cherche à identifier les tweets mentionnant des symptômes actuels et non passés, lié à l'auteur du tweet ou à l'un de ses proches. Les tweets présentés sont ceux de notre base de données – 9 917 tweets mentionnant des symptômes ; nous les avons anonymisés en retirant les noms des utilisateurs mentionnés, et les liens url. L'on explique également sur la plateforme l'intérêt pour nous de ces annotations en décrivant notre projet, et il est possible de télécharger les tweets et les annotations afin de participer au projet.

Chaque utilisateur a un identifiant unique UUID, lié à son adresse IP et à son support électronique. Lorsqu'un utilisateur arrive sur la page pour la première fois, les trois premiers tweets qu'un utilisateur se voit proposer sont des tweets pour lesquels on dispose déjà d'annotations ; par la suite, les tweets sont proposés de façon aléatoire. En analysant les premiers résultats, nous avions constaté qu'un très grand nombre des annotations avaient été réalisées par des annotateurs sans UUID, c'est-à-dire des individus qui n'ont pas accepté les cookies ; cela nous a laissé penser à des bots. Pour les analyses, nous avons alors décidé de ne prendre en compte que les annotations réalisées par des participants ayant un UUID.

4.2.2 Résultats sur les annotateurs

Le 28 septembre 2020, soit un peu moins de deux mois après le lancement de la plateforme, nous avions 6 203 annotations, réalisées par 180 participants. On représente dans la Figure 5 le nombre d'annotations par utilisateur, et l'évolution

des annotations au cours du temps. L'on observe que la plupart des participants annotent assez peu de tweets, même si quelques individus en ont annoté plus de 500 – nous soupçonnons certaines personnes de notre équipe ou du CRI. Lorsqu'on étudie l'évolution du nombre d'annotations au cours du temps, les pics de participation correspondent principalement aux moments où l'on a lancé l'application, le 3 août 2020 ; et aux partages de la plateforme sur les réseaux sociaux, comme le 25 septembre 2020.



(a) Répartition du nombre d'annotations par utilisateur

(b) Evolution de nombre d'annotations au cours du temps

FIGURE 5 – Présentation des résultats sur les annotateurs

4.2.3 Résultats sur les annotations

Sur les 6 203 annotations, 3 848 tweets ont été labellisés au moins une fois, étant donné que plusieurs tweets sont annotés plusieurs fois. Pour corriger notre signal, on considère comme "vrais positifs" les tweets qui ont été annotés au moins 50% du temps comme positifs, c'est-à-dire, moins de 50% comme "non" ou "passer" ; cela ne concerne désormais plus que 969 tweets, ce qui signifie que près de 75% de nos tweets seraient en réalité des faux positifs.

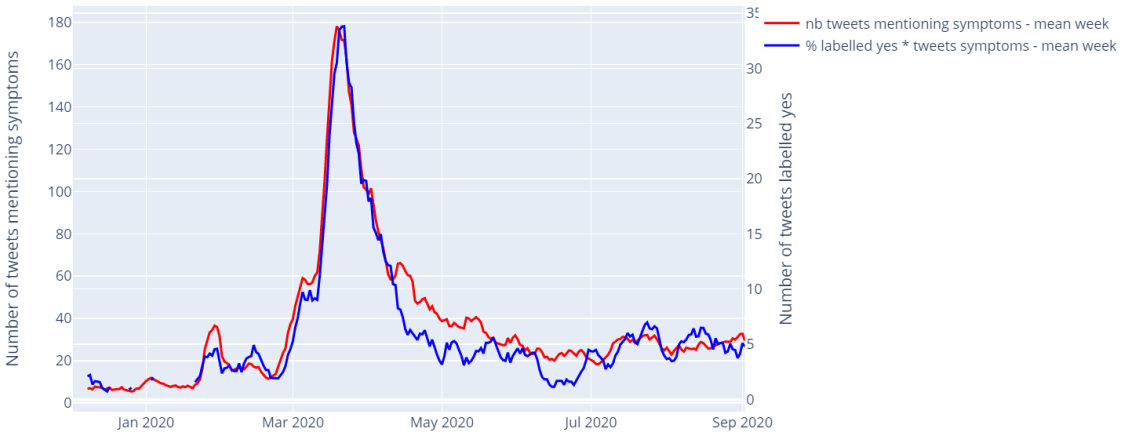
Bien qu'il soit peut-être encore trop tôt pour réaliser des analyses sur le sujet, car dans l'idéal nous préférerions disposer d'au moins deux annotations concordantes pour chaque tweet, on a commencé à travailler sur les tweets déjà annotés. On représente sur les Figures 6 la courbe corrigée. Pour construire cette série temporelle (Y_d), nous n'avons pas simplement représenté le nombre de tweets annotés positivement selon notre méthode, car nous avons ajouté progressivement des tweets dans la plateforme, si bien que l'on a mécaniquement moins de tweets annotés fin août. Cette courbe est construite comme suit :

$$Y_d = \text{Positive rate}_d * X_d$$
$$= \frac{\text{Tweets yes}_d}{\text{Nb annotations}_d} X_d$$

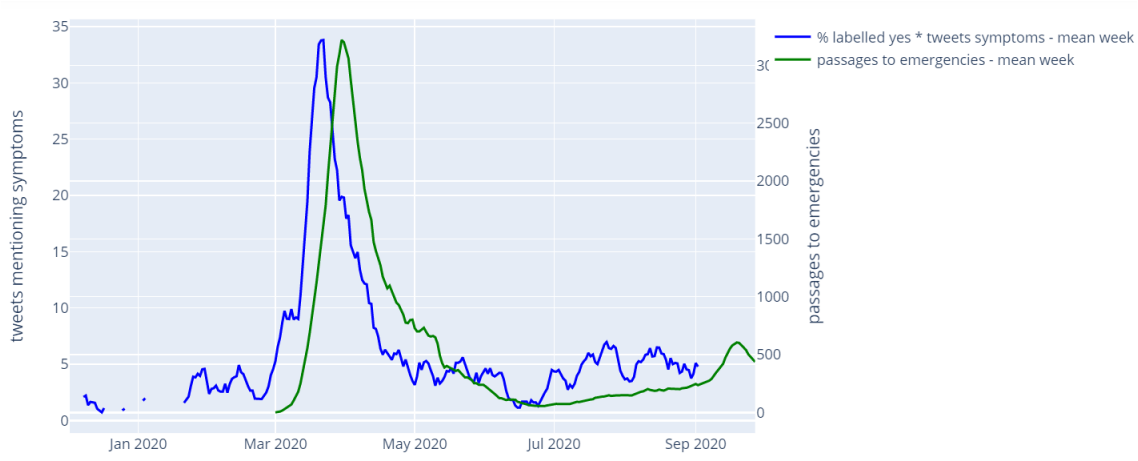
où X_d est le nombre de tweets par jour mentionnant des symptômes, $Positiverate_d = \frac{Tweetsyes_d}{Nbannotations_d}$ le taux de positivité journalier, c'est-à-dire le nombre de tweets labellisés comme "oui" divisé par le nombre d'annotations par jour, et (Y_d) la série temporelle corrigée.

Ainsi, la Figure 6a représente les courbes (X_d) non filtrée (en rouge) et (Y_d) corrigée (en bleu), moyennées sur sept jour pour plus de visibilité. On observe que si les deux courbes sont très similaires, avec un pic en mars 2020 maintenu, elles s'écartent ensuite. En effet, suite à la première vague et la fin du confinement, le nombre de tweets de la courbe corrigée descend plus bas que celui de la courbe initiale, ce qui est plus conforme avec l'évolution des passages aux urgences ; le nombre de tweets positifs remonte ensuite légèrement à partir de juillet 2020.

Lorsque l'on compare la courbe corrigée avec le nombre de passages aux urgences comme dans la Figure 6b, cette courbe corrigée semble être plus appropriée que la courbe précédente des tweets mentionnant des symptômes, bien que le coefficient de corrélation de Pearson ne varie pas entre ces deux séries temporelles (il vaut toujours 0,96, avec une p-value très faible). Cependant, même avec cette courbe-là, il semblerait qu'il y ait plus de tweets mentionnant des symptômes en juillet-août que de cas réels ; cela pourrait venir potentiellement d'un changement dans les manières de tweeter des individus, qui seraient plus spontanés à mentionner leurs symptômes sur les réseaux sociaux.



(a) Comparaison courbe filtrée et non filtrée



(b) Comparaison courbe filtrée et passages aux urgences

FIGURE 6 – Représentation au cours du temps des tweets annotés positifs

4.3 Classification

L'objectif du recours à une plateforme d'annotation des tweets est de disposer d'un jeu d'entraînement labellisé, afin de construire des algorithmes de classification et de pouvoir prédire les tweets non annotés. Mon co-superviseur, Samuel Fraiberger, a également accès à des données Twitter à plus grande échelle, dans d'autres pays ; on souhaiterait également appliquer des algorithmes de classification construits sur cette plateforme à ses données, afin d'éliminer les faux positifs. Bien que l'on ne dispose pas encore de suffisamment de données pour avoir plusieurs annotations concordantes sur chaque tweet, nous avons tout de même commencé à essayer de faire de la classification.

4.3.1 Algorithme de classification

Bag-of-words

Dans une approche de traitement du langage, on adopte une approche BoW (*Bag of Words*), une façon de concevoir le texte en ignorant la grammaire et en se concentrant sur la présence ou non de mots ou expressions dans le document. Nous avons tout d'abord préprocessé les tweets mentionnant des symptômes, en remplaçant le texte en minuscules, et en enlevant les accents et la ponctuation. Nous avons utilisé le package *ekphrasis*[9] en Python, qui est particulièrement utile pour le préprocessing de données de réseaux sociaux : il permet de reconnaître automatiquement les

9. https://github.com/cbaziotis/ekphrasis

utilisateurs, et identifie la présence de *hashtags* (termes ou expressions précédées du symbole "#") en ajoutant le terme *hashtag* dans le tweet ("covid" devient alors par exemple "<hashtag> covid", ce qui permet alors de considérer le terme avec ou sans le hashtag en utilisant des ngrams). Nous avons choisi de ne pas retirer les *stopwords*, termes très fréquents et qui incluent notamment des conjonctions de coordination et des pronoms, justement car la présence de pronoms et de verbes simples sont cruciaux dans notre analyse pour rechercher les symptômes autodéclarés. Nous avons ensuite transformé les textes en *tokens*, c'est-à-dire des mots, en autorisant la présence de *n-grams* avec $n = 4$: on considère alors les groupes comprenant jusqu'à 4 termes. Cela permet de prendre en compte des expressions et portions de phrases.

On ne s'intéresse dans un premier temps qu'aux tweets ayant déjà été annotés. On considère les tweets ayant comme label "positif" ceux ayant été annotés au moins 50% du temps positivement, et négatifs les autres, ayant été annotés au moins une fois.

Feature selection

On a déterminé les *features* les plus importants dans notre analyse en faisant de la *feature selection*. Pour identifier d'abord un ensemble de *tokens* les plus importants dans la prédiction de la présence de symptômes autodéclarés dans un tweet, on utilise un prédicteur linéaire avec la fonction *SelectKBest* du package *scikit-learn*, qui nous renvoie les *features* ayant un effet statistiquement significatif sur le label du tweet. On utilise ensuite un *LASSO* pour affiner cette liste de *features*. Le Lasso est similaire à une régression linéaire où l'on minimise la somme des carrés, avec l'ajout d'un terme de régularisation qui permet d'éviter que les coefficients de la régression prennent des valeurs extrêmes et que notre modèle suraprenne sur les données. Le Lasso est défini comme suit :

$$|y - X\beta|^2 + \lambda_1 \sum_{j=1}^{k} |\beta_j|$$

avec λ_1 le terme de régularisation, qu'on détermine par cross-validation. Avec l'application du Lasso comme méthode de *feature selection*, on peut alors choisir de conserver les termes les plus pertinents pour notre prédiction.

En classant les coefficients du Lasso par ordre décroissant, on peut également observer quels sont les termes les plus importants dans la prédiction de la présence

de symptômes autodéclarés ou non. Les dix premiers termes sont les suivants : "odorat", "gorge", "courbatures", "tous les symptomes", "ai", "je tousse", "de la fievre", "depuis", "suis", "jai". Cela est particulièrement intéressant, car on retrouve à la fois des symptômes assez spécifiques du COVID-19, comme la perte d'odorat ; et également des pronoms et verbes à la première personne du singulier, que l'on peut considérer comme des marqueurs d'expérience vécue.

Modélisation

Nous avons ensuite séparé notre base de données en un échantillon d'entraînement (70% de la base) et un échantillon de test (30%). Nous avons choisi d'utiliser comme modèle de classification la régression logistique, qui est un modèle couramment utilisé dans les cas de classification binaire. A partir des *features* en entrée, ce modèle attribue à chaque observation une probabilité entre 0 et 1 d'appartenir à la classe de tweet mentionnant un symptôme autodéclaré.

Le choix des hyperparamètres – en particulier la valeur du paramètre C, l'inverse du paramètre de régularisation – s'est fait par cross-validation. On a également choisi de fixer le paramètre de *class weight* sur *balanced*, ce qui est utile dans le cas de classes de taille différente comme c'est le cas ici – le nombre de tweets labellisé "oui" représente environ 25% de la base.

4.3.2 Résultats

Résultats sur la base des tweets annotés

Dans cette section, on présente les résultats de notre modèle sur les échantillons d'entraînement et de test. On évalue ces résultats selon plusieurs métriques, définies dans la table 2.

On observe immédiatement que les résultats de la régression logistique sont meilleurs sur le *training set* par rapport à la prédiction sur le jeu de test : l'*accuracy*, est plus élevée sur le training set (0,92) que sur le testing set (0,80), ce qui signifie que le modèle surapprend : il colle probablement trop aux données et a du mal à s'adapter sur de nouvelles données. On observe également cela sur les *ROC curves* de la Figure 7 : le modèle est plus bien précis sur le jeu d'entraînement.

Par ailleurs, notre modèle est bien plus efficace dans la prédiction des "non" que des "oui". En observant la précision pour les "Yes" dans la Table 2, et les matrices de

Métrique	Formule	Label	Training set	Testing set
précision	$\frac{TP}{TP+FP}$	Yes	0.79	0.57
		No	0.98	0.91
recall	$\frac{TP}{TP+FN}$	Yes	0.95	0.75
		No	0.91	0.82
f1 score	$2 * \frac{precision*recall}{precision+recall}$	Yes	0.86	0.65
		No	0.95	0.86
accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$		0.92	0.80

TABLE 2 – Métriques obtenues avec la régression logistique

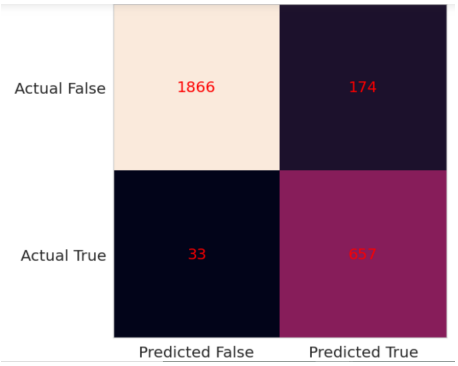
Dans les formules des différentes métriques, TP signifie *"True positive"*, c'est-à-dire le nombre de tweets prédits dans la bonne classe (tweets positifs prédits comme étant positifs, ou tweets négatifs prédits comme étant négatifs) ; TN signifie *"True negative"* ; de même, FP et FN signifient *"False positive"* et *"False negative"*.

confusion de la Figure 7, si peu de vrais "Yes" sont labellisés en "No", de nombreux tweets "No" sont en réalité labellisés comme "Yes". Cela se constate dans la métrique de précision pour les labels "Yes", mais assez peu dans la métrique de "recall" des tweets labellisés "No", en raison du grand nombre de tweets dans cette classe.

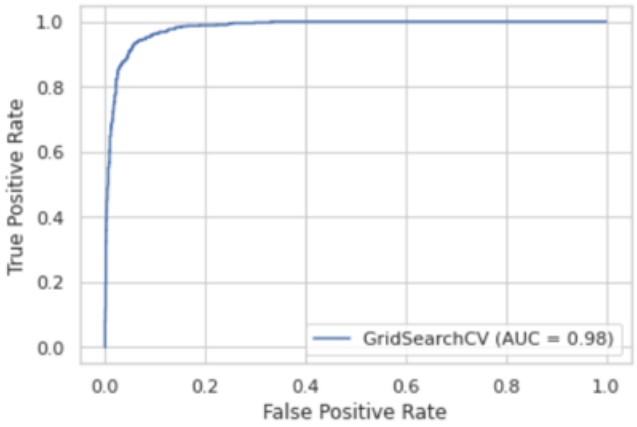
Malgré l'overfitting et le manque de précision pour les tweets labellisés "Yes", notre modèle a tout de même une *accuracy* assez intéressante. Si nos résultats présentés ici ne sont pas définitifs, puisque l'on souhaite avoir plus d'annotations pour améliorer la labellisation des tweets et également davantage de tweets à notre disposition, on a tout de même tenté de construire un modèle de classification de nos tweets.

Prédictions sur l'ensemble des tweets mentionnant des symptômes

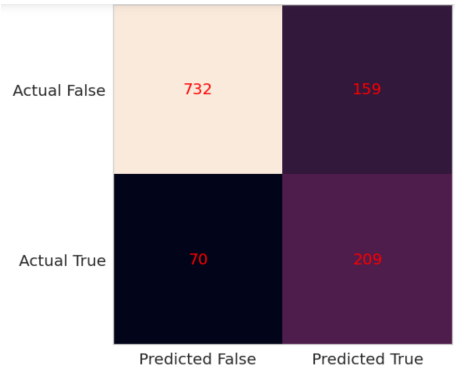
Nous avions construit notre modèle à partir uniquement des tweets annotés, car cela nous permettait d'entraîner et tester notre algorithme sur un jeu de données labellisé. Nous avons alors ensuite utilisé notre modèle de classification pour prédire l'ensemble de nos tweets mentionnant des symptômes, c'est-à-dire également nos tweets qui n'avaient pas été labellisés. On représente dans la Figure 8 la courbe nouvellement construite, moyennée sur 7 jours. On la compare dans un premier temps, dans la Figure 8a, à l'évolution du nombre de tweets mentionnant des symptômes. On constate que les deux courbes sont très similaires, et que le pic du début du



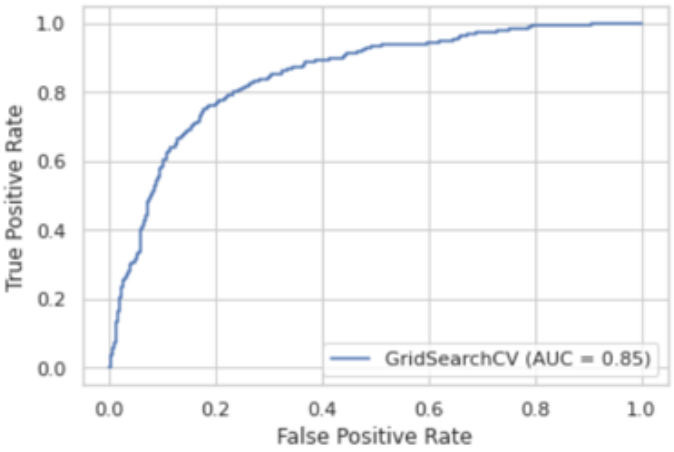
(a) Matrice de confusion - training set



(b) ROC curve - training set



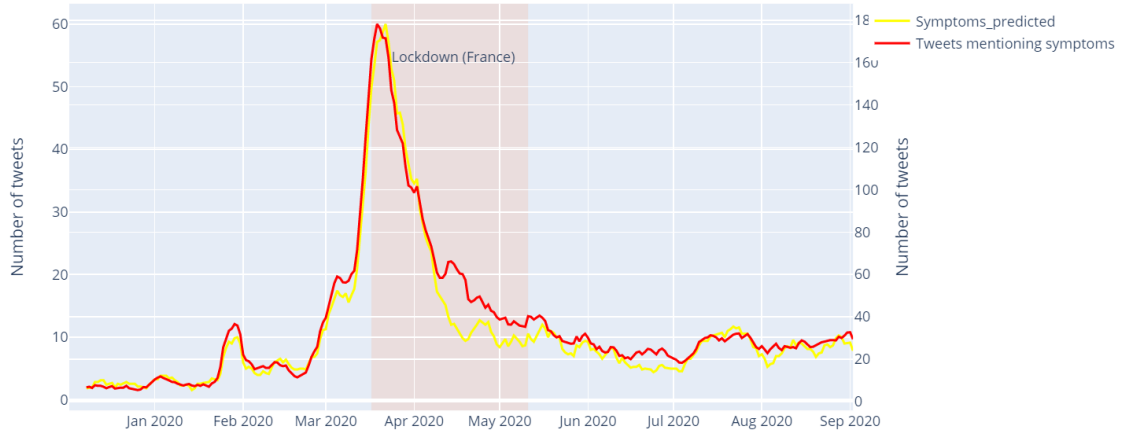
(c) Matrice de confusion - testing set



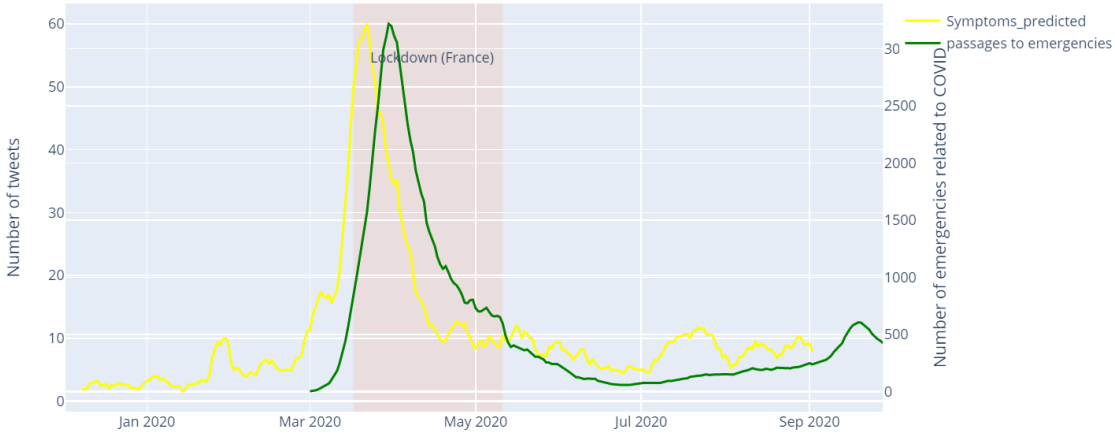
(d) ROC curve - testing set

FIGURE 7 – Présentation des résultats de la régression logistique sur les jeux d'entraînement et de test

confinement en France est maintenu, ce qui est rassurant sur la qualité de prédiction de notre modèle. Par ailleurs, la courbe filtrée descend légèrement plus bas que la courbe non filtrée en mai 2020, ce qui est plus conforme avec l'évolution réelle des cas des COVID, et qui laisserait penser que l'on a bien réussi à filtrer les tweets mentionnant des faux positifs. En effet, lorsqu'on compare cette nouvelle courbe avec le nombre de passages aux urgences comme dans la Figure 8b, on observe que les deux courbes semblent être plus proches que ne l'était la courbe non filtrée initialement (Figure 3) : le coefficient de corrélation entre les deux séries temporelles (avec un décalage de 11 jours) vaut 0.97, avec une p-value très faible. Cela peut alors être encourageant pour notre démarche de *crowdsourcing* pour l'annotation de tweets.



(a) Evolution du tweets mentionnant des symptômes et des tweets prédits par la régression logistique - moyennés sur 7 jours



(b) Evolution du nombre de passages aux urgences et des tweets prédits par la régression logistique - moyennés sur 7 jours

FIGURE 8 – Evolution du nombre de tweets prédits par le classifier comme contenant des symptômes autodéclarés

4.4 Discussion

Après la présentation de tels résultats et l'introduction d'un algorithme de classification à ce stade de notre projet, il convient de discuter de nos résultats. Tout d'abord, bien que nos résultats soient encourageants, nos analyses ne se basent ici que sur un nombre très réduit de tweets : rappelons que sur les 9 917 tweets mentionnant des symptômes que nous avons à notre disposition, seulement 969 sont considérés comme positifs ; il serait très audacieux de s'estimer capables de prédire l'évolution d'une telle épidémie à partir uniquement de ce type d'échantillon. La première limite est alors le faible nombre de tweets à notre disposition, ce que nous

essayerons de pallier en travaillant sur des jeux de données plus grands, notamment ceux que mon co-superviseur Samuel Fraiberger a à sa disposition.

La seconde limite est celle du nombre d'annotations, qui est aujourd'hui encore trop faible pour s'assurer du type de label. Nous avions fait le choix – par défaut et pour obtenir des premiers résultats – de définir comme positif un tweet labellisé au moins 50% du temps positif. Cependant, un choix plus rigoureux serait de ne valider uniquement les labels des tweets ayant été labellisés par plusieurs annotateurs de façon concordante ; dans leur cas de labellisation de tweets via une plateforme de *crowdsourcing*, Burnap et Williams 2015 requièrent notamment un minimum de 4 annotateurs par tweet.

En ce qui concerne le choix du modèle utilisé, nous avons débuté par une classification avec régression logistique, mais il ne s'agissait que d'un début. On souhaiterait utiliser également des modèles plus adaptés au format spécial de données que nous possédons, et utiliser par exemple des *word embeddings* ou des modèles BERT, spécifiques aux textes. Par ailleurs, nos observations sur les données laissent penser que l'utilisation de Twitter au cours du temps est variable : si l'on observe davantage de tweets mentionnant des symptômes après le confinement, c'est également parce que les individus parlent désormais davantage de symptômes sur les réseaux, et en font également notamment des blagues. Il conviendrait alors également de songer à construire des modèles évolutifs en fonction du temps.

Enfin, on peut soulever un dernier questionnement quant à l'utilisation d'une plateforme de *citizen science* créée par nous-mêmes. Il semblerait que peu nombreux sont les annotateurs qui labellisent beaucoup de tweets, et nos annotations sont drivées par un petit nombre d'utilisateurs – probablement des personnes de l'équipe ou des collaborateurs. Si le choix de créer notre propre plateforme était guidé par des raisons économiques et pratiques, afin d'avoir plus de souplesse dans sa configuration, l'un des intérêts non négligeable des plateformes de crowdsourcing est la mise à disposition directe d'une communauté d'individus qui peuvent annoter les données.

5 Conclusion

En pleine pandémie, j'ai eu l'occasion de travailler, lors de mon stage de fin d'études, sur un projet très actuel. Je me suis intéressée à la façon dont les messages publiés sur le réseau social Twitter sont liés à l'évolution de la situation sanitaire en France. Une première partie de mon travail consistait à récupérer des tweets d'utilisateurs géolocalisés en Île-de-France via l'API Twitter, et à travailler sur des bases de données de grand volume. J'ai ensuite construit manuellement une liste de mots-clefs caractérisant des façons d'exprimer des symptômes liés au COVID-19, et ai représenté graphiquement leur évolution. En raison de la présence de faux-positifs dans la base, nous avons lancé une plateforme de *crowdsourcing* permettant à des volontaires d'annoter nos tweets. J'ai enfin commencé à construire un algorithme de machine learning pour classifier nos tweets selon la présence ou non de symptômes autodéclarés.

Pour la suite de ce projet, il nous faudra réfléchir à d'autres algorithmes de classification qui seraient plus adaptés à des données textuelles, et également travailler sur un jeu de données plus grand, en utilisant par exemple les données de Samuel. Travailler sur un tel projet est par ailleurs extrêmement intéressant dans la mesure où il suscite la curiosité d'autres membres de la communauté scientifique, et nous offrirait potentiellement des opportunités de collaboration avec d'autres chercheurs.

Mon stage au CRI m'a beaucoup appris en termes de compétences techniques. Il s'agissait pour moi de la première fois que je travaillais sur des bases de données d'un tel volume, qui requièrent l'utilisation de technologies plus avancées que de simples pandas dataframes, comme pyspark. J'ai également beaucoup découvert en termes de *citizen science* en raison de la tournure qu'a finalement pris notre projet.

Enfin, ce stage de fin d'études a été pour moi une façon de confirmer mon intérêt pour la recherche et les sciences sociales computationnelles. J'ai prévu de rester l'année prochaine au CRI en tant qu'ingénieure de recherche, afin de continuer ce projet et de pouvoir travailler sur l'étude des réseaux djihadistes, projet pour lequel j'avais été initialement recrutée.

Références

[1] Pete Burnap et Matthew L. Williams. “Cyber Hate Speech on Twitter : An Application of Machine Classification and Statistical Modeling for Policy and Decision Making”. In : *Policy & Internet* 7.2 (avr. 2015), p. 223-242. doi : 10.1002/poi3.85. url : https://doi.org/10.1002/poi3.85.

[2] Cynthia Chew et Gunther Eysenbach. “Pandemics in the Age of Twitter : Content Analysis of Tweets during the 2009 H1N1 Outbreak”. In : *PLoS ONE* 5.11 (nov. 2010). Sous la dir. de Margaret Sampson, e14118. doi : 10.1371/journal.pone.0014118. url : https://doi.org/10.1371/journal.pone.0014118.

[3] Tim Finin et al. “Annotating Named Entities in Twitter Data with Crowdsourcing”. In : *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. Los Angeles : Association for Computational Linguistics, juin 2010, p. 80-88. url : https://www.aclweb.org/anthology/W10-0713.

[4] Karën Fort, Gilles Adda et K. Bretonnel Cohen. “Amazon Mechanical Turk : Gold Mine or Coal Mine ?” In : *Computational Linguistics* 37.2 (juin 2011), p. 413-420. doi : 10.1162/coli_a_00057. url : https://doi.org/10.1162/coli_a_00057.

[5] Jeremy Ginsberg et al. “Detecting influenza epidemics using search engine query data”. In : *Nature* 457.7232 (fév. 2009), p. 1012-1014. doi : 10.1038/nature07634. url : https://doi.org/10.1038/nature07634.

[6] D. Lazer et al. “The Parable of Google Flu : Traps in Big Data Analysis”. In : *Science* 343.6176 (mar. 2014), p. 1203-1205. doi : 10.1126/science.1248506. url : https://doi.org/10.1126/science.1248506.

[7] Tim Mackey et al. “Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter : Retrospective Big Data Infoveillance Study”. In : *JMIR Public Health and Surveillance* 6.2 (juin 2020), e19509. doi : 10.2196/19509. url : https://doi.org/10.2196/19509.

[8] Mor NAAMAN, Jeffrey BOASE et Chih-Hui LAI. “Is it really about me ?” In : *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*. ACM Press, 2010. DOI : 10.1145/1718918.1718953. URL : https://doi.org/10.1145/1718918.1718953.

[9] Abeed SARKER et al. “Self-reported COVID-19 symptoms on Twitter : An analysis and a research resource”. In : (avr. 2020). DOI : 10.1101/2020.04.16.20067421. URL : https://doi.org/10.1101/2020.04.16.20067421.

[10] Denny VRANDEČIĆ. “Wikidata”. In : *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*. ACM Press, 2012. DOI : 10.1145/2187980.2188242. URL : https://doi.org/10.1145/2187980.2188242.

Annexes

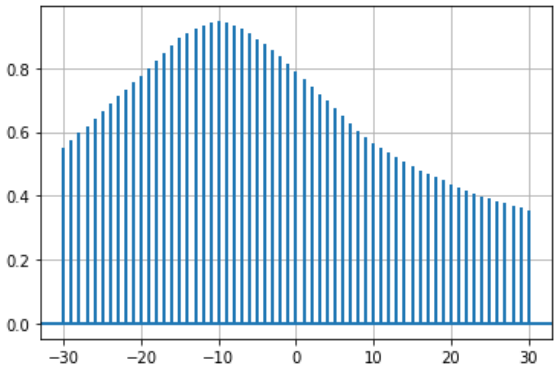


FIGURE 9 – Cross-corrélation entre les séries temporelles du nombre de tweets mentionnant des symptômes et du nombre de passages aux urgences par jour

Note de synthèse

Français

J'ai réalisé mon stage de fin d'études au CRI Paris entre mars et octobre 2020, au sein de l'équipe Interaction Data Lab, dirigée par Marc Santolini ; Samuel Fraiberger (World Bank & New York University) était mon co-superviseur durant ce stage. L'équipe Interaction Data Lab étudie principalement les interactions sociales et l'organisation de communautés, en adoptant une approche fondée sur les données. J'ai été initialement recrutée pour travailler sur un projet visant à étudier l'évolution des réseaux djihadistes sur Twitter, co-supervisé par Hugo Micheron, sociologue à l'ENS ; cependant, en raison de retards administratifs, j'ai finalement commencé à travailler sur un autre projet à partir également de données Twitter.

L'objectif de ce projet est l'étude de symptômes autodéclarés du COVID-19 sur Twitter, c'est-à-dire, la détection et l'analyse de tweets dans lesquels l'auteur mentionne ses propres symptômes ou des symptômes de l'un de ses proches, avec une focalisation sur la région Île-de-France. La première partie de mon étude a été de comprendre le fonctionnement de l'API Twitter, et de collecter ces données : j'ai tout d'abord utilisé la requête Streaming avec un filtre géographique pour collecter des tweets d'utilisateurs géolocalisés en Île-de-France. J'ai ensuite collecté, pour chacun des 30 000 d'utilisateurs identifiés en Île-de-France, leurs données historiques – c'est-à-dire jusqu'aux 3 600 tweets les plus récents – de sorte à former une base de données de plus de 40 millions de tweets.

Deux approches ont été adoptées pour la reconnaissance de symptômes autodéclarés du COVID-19 dans les tweets. Une première se basait sur la construction d'un dictionnaire de mots-clefs contenant des expressions se rapportant à des symptômes du COVID-19. Nous avons établi manuellement une liste des expressions utilisées pour décrire des symptômes du COVID-19, et également de façon familière (exemple : "avoir de la température" pour le symptôme "fièvre") car Twitter se rapproche plus du langage parlé. Au total, 9 917 tweets de notre base de données mentionnent au moins un symptôme. Nous avons ensuite représenté graphiquement l'évolution du nombre de ces tweets par jour, et comparé avec l'évolution du nombre de passages aux urgences pour suspicion de COVID-19 en Île-de-France, données publiques de Santé Publique France. On a observé une forte corrélation entre ces

données, avec un lag de 11 jours : le nombre de passages aux urgences varie avec la même intensité que le nombre de mentions de symptômes sur Twitter, mais 11 jours plus tard. La corrélation entre les deux courbes diminue cependant après mai 2020 : le nombre de tweets ne diminue pas autant que le nombre de passages aux urgences.

Cependant, en observant les tweets individuellement, on a identifié la présence de nombreux "faux positifs" dans notre base de données, c'est-à-dire des tweets mentionnant des symptômes, mais ne concernant pas l'auteur du tweet. Il s'agit souvent de tweets rappelant les mesures d'hygiène à suivre, donnant des nouvelles générales, ou faisant des blagues sur le sujet. Pour filtrer ces faux positifs, ajouter de simples règles, telles que la présence de pronoms dans les tweets, ne suffisait pas. Nous avons alors mis en place, avec l'aide d'un collaborateur du CRI, une plateforme de science citoyenne (*citizen science*), permettant à des volontaires d'annoter bénévolement nos tweets préalablement anonymiser, et de visualiser l'impact de leur contribution.

Cette plateforme a été lancée le 3 août 2020, et fin septembre, on comptait plus de 6000 annotations concernant 3850 tweets. Bien qu'il soit encore trop tôt pour tirer de réelles conclusions sur ces annotations, nous avons commencé à les analyser : nous avons considéré comme tweets "positifs" (mentionnant les symptômes de l'auteur du tweet ou de l'un de ses proches) ceux qui avaient été annotés au moins 50% du temps positivement. Puis, nous avons construit un algorithme de classification, avec régression logistique, sur l'ensemble des tweets déjà ayant été annotés. Si notre modèle n'est pas encore parfait, surapprend (l'accuracy vaut 0.92 sur le training set et 0.80 sur le testing set), et n'est peut-être pas le plus adapté car on aurait pu employer des modèles plus spécifiques au traitement de textes, il permet de nous donner une première estimation de l'évolution des tweets mentionnant des symptômes autodéclarés.

Dans la suite de ce projet, on cherchera alors à utiliser d'autres types de modèles plus adaptés au texte, et des modèles évolutifs au cours du temps, car la façon de tweeter des individus, et en particulier à propos de termes liés au COVID-19 est très changeante. Nous chercherons également à travailler sur des bases de données plus importantes, et localisées dans d'autres régions de France et pays.

English

I did my end-of-studies internship at CRI Paris between March and October 2020, within the Interaction Data Lab, led by Marc Santolini ; Samuel Fraiberger (World Bank & New York University) co-supervised me during this internship. The Interaction Data Lab team mainly studies social interactions and the organization of communities, using a data-driven approach. I was initially recruited to work on a project to study the evolution of jihadist networks on Twitter, co-supervised by Hugo Micheron, sociologist at the ENS. However, due to administrative delays, I eventually started working on another project analyzing Twitter data as well.

The goal of this project is to study COVID-19 self-reported symptoms on Twitter, i.e., to detection and analyze tweets in which the author mentions his or one of his relatives' symptoms, with a focus on the Paris region (Île-de-France). The first part of my study was to understand how the Twitter API works, and to collect these data : I first used the Streaming query with a geographic filter to collect tweets from users geolocated in the Paris region. I then collected, for each of the 30,000 users identified in the Paris region, their historical data – i.e. up to 3,600 most recent tweets – so as to build a database of more than 40 million tweets.

We adopted two different approaches to recognize COVID-19 self-reported symptoms in tweets. The first was based on the construction of a keywords dictionary containing expressions related to COVID-19 symptoms. We manually compiled a list of familiar expressions used to describe COVID-19 symptoms (example : "avoir de la température" for the symptom "fever") because Twitter is closer to spoken language. A total of 9,917 tweets in our database mention at least one symptom. We then graphically represented the evolution of the number of these tweets per day, and compared it with the evolution of the number of passages to emergencies for suspicion of COVID-19 in Ile-de-France, using public data from Santé Publique France. We observed a strong correlation between these time series, with a lag of 11 days : the number of passages to emergencies varied with the same intensity as the number of symptom mentions on Twitter, but 11 days later. The correlation between the two curves decreases after May 2020 though : the number of tweets does not decrease as much as the number of passages to emergencies.

By observing more cautiously the tweets, we have identified the presence of many "false positives" in our database, which are tweets mentioning symptoms, but not

related to the author of the tweet. These are often tweets reminding people of hygiene measures to follow, giving general news, or making jokes on the subject. We first tried adding simple rules, such as the presence of pronouns in the tweets, to filter out these false positives ; but it did not work well. We then set up, with the help of a CRI collaborator, a citizen science platform allowing volunteers to annotate our anonymized tweets, and to visualize the impact of their contribution.

Our platform was launched on August 3, 2020 ; by the end of September, we had more than 6,000 annotations of 3,850 tweets (because some tweets were labelled several times). Although it is still too early to draw real conclusions on these annotations, we have started to analyze them. We considered as "positive" tweets (mentioning the symptoms of the author of the tweet or one of his relatives) those that had been annotated at least 50% of the time positively. Then, we built a classification algorithm, using logistic regression, on all the tweets that had already been annotated. Our model is not yet perfect : it overfits (accuracy is 0.92 on the training set and 0.80 on the testing set), and it is not specific to textual data ; however, it allows us to give us a first idea of the evolution of tweets mentioning self-reported symptoms.

To continue this project, we will then try to use other kins of models that are more adapted to textual data, and to use models that evolve over time : indeed, the way individuals tweet is very changing, and particularly tweets related to COVID-19. We will also seek to work on larger databases, located in other regions of France and countries.