Exploitation des données des réseaux sociaux pour une analyse de la propagation épidémiologique

par

Askoum Koumtingue

Essai présenté au CeFTI

en vue de l'obtention du grade de maître en technologies de l'information

(Maîtrise en génie logiciel incluant un cheminement de type cours en technologies de l'information)

FACULTÉ DES SCIENCES

UNIVERSITÉ DE SHERBROOKE

Longueuil, Québec, Canada, décembre 2017

Sommaire

Le développement fulgurant de l'Internet au cours de ces 20 dernières années a transformé le monde de diverses manières. L'arrivée du web 2.0 a vu la création d'un nouveau type d'application qui a changé la manière dont les rapports humains fonctionnent. En effet, en permettant la création de ce qu'on appelle aujourd'hui les réseaux sociaux, cette nouvelle donne a conduit à décupler le nombre d'échange et d'interactions humaines. Ainsi, les internautes peuvent interagir (partager, échanger) de façon simple, à la fois avec le contenu et la structure des pages web. Une page web représente un document mis en ligne sur un site et consultable à l'aide d'un navigateur web. L'accès à cette dernière se fait grâce à une adresse web qui peut être entrer manuellement par l'internaute ou en activant un hyperlien.

Les internautes ne sont plus de simples consommateurs, mais participent à la création de contenus, rendant des données disponibles et accessibles en temps réels. L'exploitation de ces donnes basées sur le comportement des utilisateurs, leurs pensées et leurs avis permet de générer des informations stratégiques pouvant servir d'aide à la décision.

Au Canada, compte tenu de l'impact de la grippe sur la population, le gouvernement a confié à l'Agence de la santé publique du Canada (ASPC) la gestion de la grippe saisonnière. Cette dernière a mis en place un programme dénommé Surveillance de l'influenza ayant pour objectif la collecte, l'analyse et l'interprétation continue et systématique des données issues de différentes sources. Dans le souci d'améliorer l'efficacité du système, l'ASPC cherche à élargir les sources d'information de surveillance de la grippe. Les données des réseaux sociaux apparaissent alors comme une opportunité pour améliorer ce système. Nous nous sommes posé la question à savoir dans quelle mesure, l'exploitation de cette masse de données que constitue Twitter, permettrait d'améliorer le système traditionnel de surveillance épidémiologique au Canada. Le présent essai se base sur une analyse comparée des données de Twitter avec celles issues des méthodes dites traditionnelles afin de démontrer le potentiel de l'utilisation ou non des données Twitter dans la surveillance de l'influenza.

Une méthodologie basée sur une approche de recherche type quantitatif corrélationnel a été utilisée. Les cas confirmés en laboratoire sont considérés comme la variable indépendante et le nombre de *tweets* par semaine contenant les mots-clés spécifiques syndromiques de la grippe sont utilisés comme variable dépendante.

L'analyse a permis de conclure à une corrélation entre la courbe de l'épidémie décrite par les données de Twitter et la courbe de l'épidémie décrite par les données de l'ASPC. L'hypothèse selon laquelle les données de Twitter peuvent être utilisées pour le suivi épidémiologique semble être vérifiée. En outre, les données de Twitter annoncent l'épidémie deux semaines avant les données de surveillance traditionnelle. La méthode a été utilisée sur les données d'un autre pays notamment les États-Unis. Elle a conduit à une conclusion semblable. Toutefois, l'étude comporte des limites qui méritent d'être considérées pour des recherches ultérieures.

Remerciements

La réalisation de cet essai n'aurait été possible sans l'intervention de certaines personnes. Qu'elles trouvent ici l'expression de mes plus sincères remerciements pour leurs précieux conseils et leur contribution directe ou indirecte.

À mes sœurs pour m'avoir encouragé et permis d'entreprendre cette formation. Sans elles, je ne serais pas là.

À M. Michel Hebert, directeur de cet essai, pour son aide précieuse, sa disponibilité et pour le temps qu'il m'a consacré.

À Mme Lynn Legault et M. Vincent Echelard d'avoir su me guider vers les bonnes références et de m'avoir accordé un peu de leur temps.

À Mme Nathalie Pigeon d'avoir pris le temps de répondre à mes interrogations sur la grippe saisonnière.

Je tiens également à remercier M. Claude Cardinal pour les conseils d'orientation qu'il nous a prodigués.

Mes remerciements vont aussi à l'endroit de tout le corps professoral du CeFTI.

Qu'il me soit enfin permis de remercier toute ma famille pour leur amour et leur soutien constant, en particulier à ma nièce Ariane, de m'avoir supporté durant toute cette période d'étude.

Merci à vous.

Table des matières

Liste des tableaux

Liste des figures

Glossaire

Application Programming Interface	Interface de programmation par laquelle un logiciel offre des services à d'autres logiciels
Hashtag	Un mot ou groupe de mots précédé du symbole dièse (#)
Micromessage	Un texte qui comporte au maximum 140 caractères, lien Internet compris
Mot-clic	Un mot ou une phrase précédée par le symbole dièse (#), dont le but principal est de mettre l'accent sur un sujet en particulier dans une publication
Réseau social	Ensemble de relations d'un type spécifique
Tweets	Communément appelés micromessages
Twitter	Twitter est un réseau social de microblogage géré par l'entreprise Twitter Inc. Il permet à un utilisateur d'envoyer gratuitement de brefs messages, appelés tweets, sur Internet, par messagerie instantanée ou par SMS
Web 2.0	Désigne l'évolution de l'Internet et ainsi l'apparition des nouvelles applications, des interfaces qui facilitent l'utilisation du web par les internautes

Liste des sigles, des symboles et des acronymes

API : *Application Programming Interface*

ASCP : Agence de la santé publique du Canada

OMS : Organisation mondiale de la santé

PCET : Programme canadien d'épidémiologie de terrain

CDC*: Centers for Disease Control and Prevention*

DCC: *Disaster and Community Crisis Center*

CGU : Contenu généré par les utilisateurs

Introduction

L'informatique occupe de plus en plus une place prépondérante comme un outil d'aide à la décision dans de nombreux domaines, y compris le secteur médical. De ce fait, l'application de l'informatique dans le secteur de la santé couvre un vaste champ allant de la gestion des établissements médicaux à l'e-santé, lequel est défini par l'OMS comme étant « le numérique au service du bien-être de la personne » [1].

L'augmentation des renseignements sur les patients, la complexité croissante et la quantité d'information à traiter, l'optimisation de la posologie des médicaments requièrent à la mise en place des systèmes d'information performants et capables d'aider des professionnels de la santé. Ce genre de système met à la disposition à ces professionnels des méthodes, des techniques et des outils permettant d'améliorer la formalisation des données et des connaissances à des fins d'une meilleure prise en charge du patient. L'informatique de santé permet également d'améliorer les politiques collectives de santé, par le biais d'une meilleure protection de la santé contre les dangers épidémiques et environnementaux. De ce point de vue, le bénéfice attendu de l'informatisation du système de santé est le suivi du dispositif de veille sanitaire grâce à une circulation verticale rapide de l'information ainsi qu'une meilleure connaissance épidémiologique d'une population donnée [2]. En fait, l'exploitation des données massives dans le domaine de la santé est un véritable potentiel qui pourrait supporter les professionnels de la santé sur divers plans en partant de la médecine préventive à la médecine personnalisée. Ces données massives, « big data » ou mégadonnées, sont nées de l'évolution de la technologie. Cette dernière rend possible la collecte systématique, la conservation et l'analyse d'informations de tout genre, notamment les données de communication, de la santé, des contenus web, de la géolocalisation, ainsi que celles des réseaux sociaux.

Les réseaux sociaux sont devenus des outils incontournables pour échanger des informations entre les personnes et ceci à travers le monde entier. Certains internautes l'utilisent pour faire valoir leur opinion sur les sujets d'actualité alors que d'autres pour faire savoir à leurs cercles d'amis ce qui se passe instantanément dans leur vie ou dans leur

entourage. L'ensemble des informations échangées et les comportements sur l'Internet sont enregistrés instantanément et stockés sur les serveurs des compagnies. Toutes ces données, si elles sont extraites et analysées, génèrent des informations stratégiques pour se renseigner sur les indicateurs de développement d'un pays. Elles permettraient de prédire une crise conjoncturelle, de connaitre la situation du marché de l'emploi ou encore de faire de la surveillance épidémiologique.

L'influenza ou communément appelée grippe est une infection virale aigüe respiratoire qui engendre de nombreux cas de décès et d'hospitalisation à travers le monde chaque année. Avoir un système de surveillance de l'influenza assez performant demeure l'une des principales préoccupations des agents de la santé publique. Au Canada, le système de surveillance de la grippe est géré par le programme « surveillance de l'influenza » de l'Agence de la santé publique du Canada. Cette dernière collecte des données provenant des sources des laboratoires de santé publique des différentes provinces et les croisent avec d'autres informations pour établir des rapports hebdomadaires des activités de la grippe.

Cet essai s'inscrit dans ce cadre et vise à déterminer si les données issues des réseaux sociaux peuvent augmenter la disponibilité des informations concernant l'épidémie de la grippe saisonnière au Canada.

Le présent document est organisé en six chapitres comme suit : Le premier chapitre situe le sujet dans son cadre spécifique afin d'informer le lecteur de l'idée directrice et de l'étendue du travail. Le deuxième chapitre, consacré à la revue de la littérature, nous permet de faire une revue sur l'état actuel des recherches relatives à ce sujet. Le troisième chapitre est consacré à la problématique de l'essai et énonce l'hypothèse, notamment l'existence d'une corrélation positive entre les données provenant du réseau social Twitter et celles des sources traditionnelles. Le quatrième chapitre explique et justifie la méthodologie d'analyse et les différentes variables utilisées. Le cinquième chapitre présente l'analyse des résultats obtenus. Enfin, le dernier chapitre présente la discussion des résultats et la conclusion.

Chapitre 1

Mise en contexte

Les technologies de l'information et de la communication ont joué un rôle important dans le domaine de la santé au cours de ces dernières années. Elles ont contribué non seulement à l'amélioration de la qualité de soins de santé, mais également permis aux individus de mieux comprendre les questions liées à leur santé en utilisant les moteurs de recherches. Grâce à l'Internet, les réseaux sociaux et autres outils numériques de communication, les patients ou communément appelés e-patients prennent leur santé en main en cherchant des informations sur leur état de santé et en discutant avec les autres internautes qui partagent les mêmes expériences. Considérés dans un premier temps comme un simple moyen de communication visant à maintenir des liens entre amis, collègues et parents, les réseaux sociaux sont devenus de véritables outils de diffusion d'information. De nombreuses études ont été réalisées sur l'utilisation des réseaux sociaux dans le domaine de la santé. Il s'agit entre autres des études portant sur l'utilisation des réseaux sociaux pour transmettre des informations relatives à la vaccination ou la prévention contre certaines épidémies, le partage des informations sur les recherches de clinique ou les opinions des patients dans un hôpital.

L'entreprise Synthesio, spécialisée dans la réputation des entreprises sur Internet, s'est intéressée à l'analyse du contenu des discussions, des messages diffusés sur les blogues et Facebook. Cette étude [3] affirme que 20 % des discussions sur l 'Internet traitent de la santé, et d'après un sondage réalisé par le site Pew Internet, 60 % de la population se tournerait en priorité vers l'Internet pour rechercher des informations liées à la santé [4].

Afin de trouver un échantillon représentatif, le contexte de réalisation de l'essai se définit dans un premier temps en terme logiciel. Le réseau social Twitter est choisi pour l'étude du fait qu'il soit non spécialisé, permettant ainsi aux internautes de se partager des informations de n'importe quel domaine. De plus, Twitter dispose d'un API de recherche qui permet d'effectuer l'analyse des micromessages recueillis. Selon l'Office québécois de la langue française, les *tweets*, appelés aussi micromessages, sont de courts messages au nombre de caractères limités dont le contenu est personnel ou informatif.

Sur le plan géographique, le taux de pénétration du réseau Twitter n'est pas identique au niveau d'un pays ou de par le monde. Il est donc important que l'échantillon de données soit sélectionné dans une région où le taux de pénétration du réseau social avoisine le taux de pénétration d'Internet du pays afin d'avoir un taux représentatif de la population.

Une étude réalisée par l'Institut de la statistique du Canada indique que 70 % des Canadiens ont effectué une recherche en ligne sur les renseignements médicaux. C'est un taux considérable qui pourrait être exploité et servir d'indicateurs par les professionnels de la santé de diverses manières [5].

Le choix de la grippe se justifie par le fait que le Canada vit une grippe saisonnière qui occasionne environ 12 200 hospitalisations et en moyenne 3 500 décès par an selon les données de l'Agence de la santé publique du Canada.

De plus, une certaine forme du virus de l'influenza est à l'origine de pandémies périodiques dans le monde. Un système de surveillance efficace, alimenté par des données abondantes et fiables, permet de prévenir la maladie et de contenir les effets en cas d'épidémie.

L'essai se base sur le cas des maladies confirmées par le laboratoire de la santé publique du Canada au cours de la période de la grippe saisonnière de 2016-2017.

1.1 Les réseaux sociaux

Depuis quelques années, les réseaux sociaux sont au cœur des nouvelles technologies de la communication. Ils ne servent plus seulement à favoriser l'interaction entre parents, amis et relations professionnelles, mais tendent à devenir un canal de communication de plus en plus utilisé pour s'adresser à l'opinion publique.

L'apparition du terme réseau social est très ancienne et remonte au 20e siècle. L'anthropologue John Arundel Barnes [6] fut l'un des premiers à utiliser le terme « réseau social » dans une étude liée au fonctionnement des classes sociales sur une île à l'ouest de la Norvège en 1954. Sur le plan sociologique, Lazega définit le réseau social comme un « ensemble de relations d'un type spécifique (ex. : collaboration, soutien, conseil, contrôle ou encore influence) entre un ensemble d'acteurs » [7].

Sur le plan technologique, Kaplan et Haenlein [8], le définissent comme étant des « Outils et applications du Web 2.0 qui permettent aux individus de se connecter et de créer des contenus dans le but d'échanger avec les amis et collègues ».

À la vue de ces différentes définitions, il est important de mettre l'accent sur l'aspect social et technologique des réseaux sociaux qui ont été relatés par ces différents chercheurs. Ainsi, il peut être défini comme une plateforme du Web 2.0 qui propose des applications de rencontre dans un but amical ou professionnel et qui permet le partage de divers contenus (blogues, vidéos, images, liens, messages privés ou publics).

Sur le plan institutionnel, les entreprises commerciales et les entreprises de services utilisent les réseaux sociaux tel un outil de promotion, mais aussi pour prédire le comportement de leurs clients.

Les réseaux sociaux prennent de plus en plus une place prépondérante dans la vie privée et professionnelle des internautes. En effet, en se basant sur l'activité des internautes dans les réseaux sociaux, des chercheurs de l'Université de Cambridge et Stanford, Youyou, Kosinski et Stillwell [9] ont démontré dans une étude publiée dans le journal de l'Académie des sciences des États-Unis qu'avec des algorithmes, il est possible de déterminer la personnalité d'un individu mieux que son entourage. Cette étude a été menée par un groupe de chercheurs et concerne 86 220 internautes volontaires sur Facebook qui ont complété un test de personnalité basé sur cinq traits de personnalité (« Big Five ») à savoir : extraversion, névrosisme, agréabilité, caractère consciencieux, et ouverture à l'expérience.

D'après les résultats, le modèle basé sur l'analyse des « J'aime » de Facebook se rapproche plus des réponses des volontaires (avec un pointage de 56 %) comparativement aux réponses des amis (49 %) et de la famille (50 %). Les réponses des conjoints sont les plus proches de celles des volontaires avec un pointage de 58 %. Cette étude établit que le contenu publié par un individu sur les réseaux sociaux pourrait en dévoiler plus sur sa personnalité réelle.

Si l'on tient compte du fait que les internautes ont tendance à publier ce qu'ils vivent réellement au quotidien sur les réseaux sociaux, l'étude citée plus haut pourrait démontrer que les informations sur les réseaux sociaux ont un certain niveau de fiabilité.

1.1.1 Catégories de réseaux sociaux

Avant de chercher à catégoriser les réseaux sociaux, il est important de comprendre la nuance qui existe entre ces différents termes : réseau social, média social et Web 2.0.

Le Web 2.0 est une des résultantes de l'explosion de la bulle Internet, ce concept est apparu lors d'une conférence de « brainstorming » en 2004 organisée par Tim O'Reilly et MediaLive International. Selon O'Reilly [10], le Web 2.0 est représenté selon sept principes et peut-être résumé comme étant une tendance dans la manière dont les internautes utilisent l'Internet et où l'on favorise la créativité, le partage et l'interconnexion entre les utilisateurs.

Les médias sociaux sont l'une des conséquences du Web 2.0 et intègrent différentes activités que sont la technologie, l'interaction sociale et la création de contenus. Les médias sociaux sont considérés comme un ensemble d'applications en ligne (les réseaux sociaux, les blogues, sites de partage).

Le réseau social quant à lui est un média qui se définit comme un ensemble d'individus ou d'organisations reliées entre eux par des liens créés grâce aux interactions sociales.

Pour mieux cerner le concept des médias sociaux, Slenger & Coutant [11] ont mené un travail de recherche qui a combiné les entrevues individuelles, l'analyse des profils des utilisateurs et bien d'autres critères. Le but de cette recherche était de proposer une classification des médias sociaux existants et définir les médias sociaux comme des usages Internet :

- dont le contenu est généré par les utilisateurs (le principe de CGU : Contenu généré par les utilisateurs) ;
- qui propose un contenu évolutif ;
- dont l'accès aux plateformes est en général gratuit, mais implique parfois l'exploitation des données des utilisateurs ;
- qui sont les produits de la rencontre de l'usage de la technologie des stratégies économiques et de leurs constructions progressives ;
- dont les outils et plateformes sont très simples d'utilisation de sorte que n'importe quel internaute peut s'en approprier.

Ainsi, le principe central des médias sociaux peut se résumer en trois lettres : CGU pour contenu généré par les utilisateurs (*User Generated Content*).

Avec plus de trois milliards d'utilisateurs à travers le monde en 2017, les réseaux sociaux se multiplient à une vitesse incontrôlable et deviennent de plus en plus spécialisés. Dans son document intitulé « Social Network Sites : Users and Uses », Thelwall [12] a catégorisé les réseaux sociaux selon les types suivants décrits dans le tableau1.1 : la socialisation, le réseautage et le social.

Categories	Types	Exemples	Description
Réseaux sociaux	Profes-sionnels	LinkedIn	Un réseau professionnel international qui permet la mise en relation entre professionnels
		Viadeo	Permet aussi de construire et gérer son réseau professionnel, mais beaucoup plus connu en France
	Généralistes	Facebook	Il permet à l'internaute d'échanger avec sa communauté d'amis sur tout et n'importe quoi. Pour devenir ami avec quelqu'un il faut lui envoyer une demande et ce dernier doit l'accepter
		Twitter	Il permet de partager des messages avec d'autres internautes avec une limitation à 2800 caractères par message, la possibilité de suivre d'autres comptes
		MySpace	Site interactif qui offre à ses abonnés de multiples services combinant blogue, espace personnel, espace communautaire
	Partage Médias	YouTube	YouTube permet de déposer des vidéos, suivre des vidéos et faire des commentaires
		DailyMotion	Un site de partage de vidéo auquel les utilisateurs peuvent télécharger, regarder et partager des vidéos
		Instagram	Il permet aux utilisateurs de créer un compte et de pouvoir éditer, partager des photos, des vidéos et des messages avec son cercle d'amis ou famille
		Flickr	Flickr met à la disposition de son public un espace pour poster photos et vidéos. Il est cependant possible de géolocaliser les endroits où les photos ont été prises

Tableau 1-1 Les types de réseaux sociaux

Traduction libre
Inspirée de : Thelwall M. [12]

1.1.1.1 Le réseau social Twitter

Le réseau social Twitter est un réseau social généraliste informationnel, qui permet aux utilisateurs d'envoyer gratuitement sur Internet de brefs messages appelés micromessages

et met les utilisateurs en relation grâce à leur publication. Un micromessage peut comporter au maximum 280 caractères, lien Internet compris.

Une autre fonctionnalité de Twitter est le « Hashtag » ou encore « mot-clic ». Le mot-clic est défini, selon l'Office québécois de la langue française, comme étant un mot ou une phrase précédée par le symbole dièse (#), dont le but principal est de mettre l'accent sur un sujet en particulier dans une publication. Selon Olivier Ertzscheid [13], le hashtag se défini ainsi : «il s'agit au sein d'un message (un tweet), d'un mot ou d'une concaténation de mots précédés du symbole dièse (#) et qui permet tant de l'indexer, soit pour pouvoir suivre l'ensemble des messages ainsi balisés, soit pour leur ajouter un niveau de sens différent. » Ces hashtags utilisés par les internautes du réseau sont indexés et peuvent être retrouvés sur une page spécifique de la plateforme rassemblant toutes les publications ayant inclus ce même mot-clic.

En date du 30 juin 2016, Twitter comptait plus de 300 millions d'utilisateurs actifs à travers le monde par mois et plus de 500 millions de micromessages publiés, il est donc devenu l'un des réseaux sociaux les plus utilisés du monde. Selon le site « *Kap-Numérique 2017* » 37 % des Canadiens utilisent le réseau social Twitter [14].

1.2 L'environnement de santé publique au Canada

La santé est définie par l'Organisation mondiale de la santé (OMS) comme étant « un état complet de bien-être physique, mental et social, et ne consiste pas seulement en une absence de maladie ou d'infirmité ». Quant à la santé publique, l'OMS la définit comme la « science et l'art de prévenir les maladies, de prolonger la vie et d'améliorer la santé physique et mentale à un niveau individuel et collectif. Le champ d'action de la santé publique inclut tous les systèmes de promotion de la santé, de prévention des maladies, de lutte contre la maladie (médecine et soins) et de réadaptation » [15]. Au Canada, l'agence responsable de cette branche est dénommée « l'Agence de la santé publique du Canada ». Elle aide la population à améliorer sa santé, en menant des actions de : prévention des maladies et blessures, promotion d'une bonne santé physique et mentale et de prestation d'information en soutien à des prises de décision éclairées.

L'ASPC a élaboré un programme dénommé Programme canadien d'épidémiologie de terrain (PCET), dont le but est d'augmenter la capacité du Canada en matière de santé publique.
Les objectifs de ce programme sont entre autres de :

- Former des professionnels de la santé publique en épidémiologie appliquée en leur fournissant des compétences et des techniques nécessaires pour régler différents problèmes de santé publique dans des situations réelles;
- Faire des appels à des épidémiologistes de terrain de partout dans le monde afin d'appuyer les organisations de santé publique qui doivent intervenir lors des situations d'urgences ou de catastrophes de santé publique.

La santé publique intervient sur plusieurs facteurs qui ont une influence directe ou indirecte sur l'état de santé de la population. Ces facteurs sont le revenu, le statut social, les réseaux de soutien social, l'éducation, l'emploi et les conditions de travail, l'environnement social et physique, les habitudes personnelles liées à la santé et les capacités d'adaptation, le développement sain de l'enfant, le patrimoine biologique et génétique, les services de santé, le sexe et la culture. Ainsi, pour mieux gérer ces facteurs, l'ASPC s'intéresse de manière particulière au contexte dans lequel évolue la santé publique.

Dans le rapport sur l'état de la santé publique au Canada de 2014 [16] , plusieurs questions ont été soulevées :

- La population canadienne est changeante, et ce changement va avoir une influence sur la santé publique dans l'avenir du fait que (i) la population canadienne est vieillissante, et cette tendance devrait se maintenir pendant plusieurs décennies ; (ii) la croissance de la population canadienne est principalement due à l'immigration et non une croissance naturelle;
- Les ainés sont aux prises à divers problèmes de santé chronique notamment les maladies mentales, les affections neurologiques et les blessures. Des tendances inquiétantes s'observent également de plus en plus chez certaines personnes plus jeunes;
- Les changements démographiques ont donné lieu à des changements sociétaux ayant des conséquences sur la santé, notamment en ce qui concerne le travail, la retraite, les pensions, la famille, les soins et les relations intergénérationnelles.

- Il convient de poursuivre les travaux de recherche et les investissements dans le domaine de la santé publique pour tenir compte des changements démographiques à venir.

Les fonctions de base traditionnelles de la santé publique constituent donc une assise solide pour protéger la population canadienne contre les maladies et risques y compris les risques associés aux changements climatiques.

Chapitre 2

Revue de la littérature

La présente revue de la littérature met l'accent sur la question de recherche en prenant en considération les différents concepts et permet entre autres de se positionner par rapport aux recherches antérieures.

La surveillance épidémiologique est d'une importance capitale dans le système de santé des pays. Il permet l'observation de l'émergence de phénomènes sanitaires, l'analyse et l'évaluation de l'impact des phénomènes dans le temps. Plusieurs méthodes de surveillance d'épidémie existent. Cependant, ces méthodes demeurent coûteuses et ne fournissent pas de retour en temps réel. Les réseaux sociaux, en favorisant l'interaction humaine et les échanges d'information, deviennent ainsi un poste d'observation intéressant de l'état de santé des populations. Ceci explique et justifie l'engouement et l'intérêt dans le monde de la recherche pour les réseaux sociaux.

Ce présent chapitre passe en revue quelques-unes des recherches précédentes qui ont été réalisées sur cet aspect des réseaux sociaux et met l'accent sur le véritable avancement de l'utilisation des réseaux sociaux dans un but de surveillance d'une épidémie.

La recherche a été effectuée sur Google Scholar, un outil accessible de l'Université de Sherbrooke qui permet de rechercher des articles scientifiques dans des bases de données spécialisées, afin de recueillir les données sur l'utilisation des réseaux sociaux en matière de santé.

2.1 L'épidémie de la grippe saisonnière

Selon l'information recueillie sur le site de l'OMS [17], la grippe saisonnière est une infection virale aigüe qui se propage facilement d'une personne à une autre, et peut toucher n'importe quelle tranche d'âge de la population. Elle est fréquente pendant les périodes hivernales dans les zones tempérées et cause de nombreux cas d'hospitalisation et de décès. Il existe

trois types de grippes saisonnières (A, B et C), avec des sous-types différents classés selon leur glycoprotéine de surface. Le virus de type A est le plus dangereux, il a provoqué plusieurs pandémies meurtrières, notamment la grippe espagnole qui a fait plus de 20 millions de décès en 1918.

2.2 L'analyse des réseaux sociaux

Les réseaux sociaux sont décrits comme un ensemble de relations entre différents acteurs qui peuvent être soit des individus, un ensemble de communautés, un groupement d'amis, des associations.

Cette relation peut être définie sous différentes natures notamment le partage des idées ou des conseils et bien d'autres. La principale caractéristique est l'interaction qui existe entre ces différents éléments. Plusieurs recherches innovantes ont été réalisées à ce sujet par des sociologues comme Georg Simmel (1908) [18] et des anthropologues tels que Barnes 1954 [6]. Ces études sont à l'origine d'importants développements de l'analyse des réseaux sociaux. Si Barnes fut l'un des premiers à utiliser le terme « réseau social » et a eu le mérite d'avoir le « droit d'auteur » du terme, l'on pourrait attribuer la paternité de l'analyse des réseaux sociaux en tant que théorie à Simmel [18]. Il la définit comme le fondement de la sociologie, science des structures des relations sociales.

L'analyse des réseaux sociaux est un moyen permettant de visualiser et de modéliser les relations sociales comme des nœuds (individus, organisations) et des liens (relations entre ces nœuds). Elle permet d'observer et de calculer les degrés, la force ou la densité de liens entre les acteurs d'un réseau. De cette façon, l'analyse des réseaux sociaux est fondée sur une approche structurale des relations entre les membres d'un milieu social organisé. Selon Lazega [7], elle permet de décrire les interdépendances entre les acteurs et ainsi avoir une « représentation simplifiée d'un système social complexe ».

D'après les études réalisées par Divjak et Peharda [19], l'analyse des réseaux sociaux (ARS) se définit comme étant la cartographie et la mesure des relations et des flux entre personnes, groupes, organismes, ou tout simplement tout autre entité de traitement d'information. Les nœuds dans le réseau sont les personnes ou les groupes tandis que les liens montrent les

relations ou les flux entre les nœuds. Elle permet d'avoir une analyse visuelle et mathématique des relations humaines.

On pourrait définir l'analyse des réseaux sociaux comme étant un moyen de collecte de données provenant des réseaux sociaux dans le but de générer des informations stratégiques et des indicateurs de développement. L'analyse des réseaux sociaux se popularise dans tous les secteurs d'activités, notamment dans le domaine médical où l'utilisation de ces derniers se fait de plus en plus fréquente. Ainsi, grâce aux réseaux sociaux, on comprend davantage le contexte dans lequel les gens deviennent malades et vivent avec leur maladie.

Les réseaux sociaux peuvent ainsi servir à diffuser les informations sur les sujets liés à la santé notamment les structures de santé, la période vaccinale contre la grippe, les avertissements sur les changements de la température. Ce caractère d'instantané des réseaux sociaux permet aux responsables de santé publique de transmettre des messages à temps réel à la population.

2.2.1 Les réseaux sociaux comme moyen de communication

Sur le plan mondial, les réseaux sociaux sont utilisés de plusieurs manières dans la communication en santé. Vu la disponibilité d'accès des réseaux sociaux, il est devenu plus simple pour les professionnels de la santé d'interagir avec les internautes via certaines plateformes de réseaux sociaux. Les auteurs Campbell & Craid ont relevé le fait que les réseaux sociaux ne sont pas seulement utilisés pour rechercher des informations sur la santé, mais mettent les patients au cœur de l'action [20]. Ceux-ci deviennent l'acteur de leur santé en cherchant à connaitre davantage sur leur état de santé.

Les centres américains de prévention et de la lutte contre les maladies (Centers for Disease Control and Prevention/CDC) soulignent la particularité qui fait des médias sociaux un outil efficace de sensibilisation de la santé et le qualifie en « trois P » (personnalisation, présentation, participation). De ce fait, le CDC utilise les réseaux sociaux pour offrir aux utilisateurs un accès aux informations médicales crédibles dont ils ont besoin [21].

D'après Bertot et ses collaborateurs [22], le principe fondamental des réseaux sociaux repose sur le fait qu'il donne la capacité aux utilisateurs d'être autonomes en leur fournissant

des outils et applications pour se faire entendre, en ce sens, il permet à toute personne disposant d'un accès à l'Internet de diffuser les informations en temps réel. Cette communication à temps réel est un moyen pour les professionnels de la santé d'utiliser ces réseaux sociaux pour transmettre des messages liés à la santé publique.

Selon Dr J. Brian Houston du département de communication au Disaster and Community Crisis Center (DCC), la communication est un moyen d'influence sur les attitudes et les comportements individuels lors des évènements de catastrophes. [23]. Dans son article intitulé « un Cadre d'utilisation des médias sociaux dans la planification des catastrophes, la réponse et la recherche », Houston [23] indique que l'utilisation des technologies de communication telles que les réseaux sociaux offre davantage de possibilités de communication bidirectionnelle pendant les moments de crises. Il l'explique par le fait que la couverture médiatique des catastrophes est limitée et n'implique normalement que des messages créés par une source unique et diffusés à un large public, avec peu de possibilités de réaction et de participation du public. En considérant cette caractéristique de bidirectionnalité des réseaux sociaux, cette action permet d'aller chercher les informations issues des internautes pour en faire de l'analyse.

2.2.2 Les réseaux sociaux et soins aux patients

De nombreux outils des réseaux sociaux ont été mis à la disposition des professionnels de la santé dont le but principal d'améliorer les soins des patients. Ces outils sont utilisés entre autres pour améliorer la collaboration entre les professionnels en suscitant la discussion sur les sujets se rapportant aux recherches de certaines maladies ou pour éduquer les patients.

De récentes recherches ont montré que les médecins développent un intérêt à interagir avec leurs patients en ligne [24] à travers les réseaux sociaux tels que Twitter et Facebook en vue notamment d'améliorer la communication avec leurs patients.

Dans un récent sondage publié par Chauhan et ses collègues, environ 60 % des médecins se sont révélés favorables à l'interaction avec leurs patients à travers les réseaux sociaux. Selon eux, cela permettrait d'assurer l'éducation des patients, les encourager à changer de comportement et à respecter les prescriptions médicales [25].

D’un autre côté, le sondage mené par Farnan et ses collègues dans une clinique de consultation familiale a révélé que 56 % des patients auraient souhaité que leur médecin utilise des réseaux sociaux pour faciliter la prise des rendez-vous, les rappels, les résultats des tests et pour répondre à certaines questions d’ordre générales [26].

2.2.3 Les réseaux sociaux et l’éducation aux patients

Les réseaux sociaux peuvent améliorer l’accès aux informations sur les soins de santé et d’autres ressources d’éducation, vu le nombre, de plus en plus croissant, des internautes à travers le monde. C’est un moyen qui permet aux patients de se joindre aux communautés virtuelles, participer aux recherches, recevoir une aide financière ou morale, se fixer des objectifs, suivre leur progrès et s’encourager mutuellement. De plus, les recherches par Househ ont montré que l’intervention à travers les réseaux sociaux a un effet positif sur le combat pour la perte de poids, l’abandon du tabac et des comportements sexuels à risques [24].

Les patients pourraient utiliser les réseaux sociaux dans le but de sortir de leur isolement et de se mettre en relation avec des gens avec qui ils partagent les mêmes conditions. C’est le cas du site réseau social (www.patientslikeme.com) qui permet aux patients d’accéder à l’information, aux suggestions et au soutien d’autres personnes qui ont la même maladie ou condition physique.

2.2.4 Utilisation des réseaux sociaux au Canada

L’explosion de la bulle Internet en 2001 est à l’origine du concept Web 2.0 qui désigne la nouvelle étape de l’évolution d’Internet. Cette évolution a permis la création des nouveaux outils de communication et de travail parmi lesquels figurent les réseaux sociaux.

D’après « *We are social* »[27], sur les 3,81 milliards d’internautes à travers le monde 2,91 milliards sont actifs sur les réseaux sociaux soit plus 2/3 des internautes. Utilisés par le quart de la population mondiale, les réseaux sociaux sont devenus non seulement un moyen populaire et rapide pour faire circuler l’information à travers le monde, mais aussi un outil qui permet d’avoir une tendance sur un sujet donné ou une actualité. Le plus important est qu’il est devenu un outil de prédiction pour des sujets sérieux tels les maladies.

L'utilisation des réseaux sociaux est très répandue au Canada. Selon les chiffres produits en 2017 par « *kap-numerique* » [14], sur les 36 millions de personnes représentant la population canadienne, 33 millions utilisent régulièrement l'Internet et 23 millions , soit 58 % de la population , sont des utilisateurs actifs sur les réseaux sociaux.

L'Agence de la santé publique du Canada (ASPC) et d'autres agences de santé des provinces du Canada utilisent également les réseaux sociaux tels que Facebook, YouTube. Pour diffuser différents messages liés à la santé aux Canadiens. Par exemple, l'ASPC [28] dispose d'une page Facebook dans laquelle elle publie des messages et des vidéos sur une base régulière pour attirer l'attention des internautes sur des sujets de santé publique.

2.2.5 Analyse des réseaux sociaux en santé publique

La surveillance épidémiologique requiert des données à temps réel pour prendre rapidement les décisions qui s'imposent avant de stopper une épidémie ou en réduire l'impact. Plusieurs travaux ont été réalisés quant à l'utilisation des réseaux sociaux dans le domaine de la santé.

L'une des plus récentes utilisations des données massives à des fins de surveillance épidémiologique a été constatée lors de la récente épidémie d'Ebola qui a frappé certains pays de l'Afrique de l'Ouest entre 2014 et 2016. L'entreprise en démarrage, HealthMap de Boston spécialisée dans l'utilisation des données massives pour détecter l'apparition d'épidémies et le suivi de leur propagation en temps réel, a réussi à observer la propagation du virus en Guinée grâce à l'analyse des informations en provenance de Twitter [29].

Il faut noter que déjà en 2014, des chercheurs de l'Université de Rochester ont réussi à cartographier en temps réel la propagation de la grippe. Ils ont été capables de prédire avec une précision de plus de 90 % le prochain abonné de Twitter qui sera contaminé par le virus de la grippe, en analysant le fil Twitter de 630 000 abonnés sur une période d'un mois. Ces fils de Twitter ont été recensés sur les réseaux sociaux Facebook et Twitter, et contenaient les mots-clics fièvre, mouchoirs et bien d'autres indicateurs précieux qui, pour les épidémiologistes, ont été considérés comme des signaux d'alarme [30].

Pervaiz [31] et ses collègues ont décrit dans leur étude l'utilisation d'une application dénommée « Google Flu Trends ». Celle-ci a permis de recueillir environ 50 millions de

requêtes des internautes les plus courants faites aux États-Unis et contenant les symptômes de la grippe et les compare au taux de la grippe reporté par le CDC.

Plusieurs organisations de santé utilisent le contenu des réseaux sociaux tels que Twitter en les combinant avec d'autres sources pour donner une réponse rapide aux catastrophes et ainsi assurer la surveillance des menaces et la protection de la santé de la population. Nombreux sont les organismes de santé qui ont une présence active sur les réseaux sociaux pour suivre les micromessages qui pourraient indiquer la présence d'une épidémie et pour partager les mises à jour sur certains incidents [32].

Fogelson [33] a documenté l'utilisation des micromessages lors des catastrophes naturelles pour identifier les zones ayant le plus besoin d'aide à partir des tweets des abonnés. Par ailleurs, les blogues, les pages des internautes sont surveillés par les hôpitaux pour obtenir une information sur les évènements potentiels de blessures massives. Obtenir des informations à temps réel dans ces genres de situations, permet une grande agilité et une meilleure réponse aux catastrophes et aux urgences de santé publique.

L'utilisation des réseaux sociaux peut également influencer les objectifs de la santé publique. En effet, l'être humain a tendance à suivre et reproduire les faits et habitudes de son entourage ou son cercle d'amis. Le cas pratique est celui où Facebook a permis aux utilisateurs de publier leur statut de donneur d'organe sur leur profil. Les résultats ont été surprenants. Selon « Donate Life America », le nombre de donneurs d'organes a été multiplié par plus de 21.

Cette propension à stimuler le changement a conduit le professeur Andrew Cameron [34] (professeur à l'Université Johns Hopkins et instigateur de cette idée) à affirmer que : « Ces dernières années, les réseaux sociaux ont montré qu'ils n'étaient pas seulement des espaces destinés à partager ce que vous avez mangé ou à poster des images de chats mignons, indique-t-il. Ils peuvent être des moteurs de changement social. »

Des études antérieures ont démontré qu'il est possible d'utiliser le réseau social Twitter pour la détection des maladies. Aslam et ses collègues [35] notamment ont collecté des micromessages contenant le mot-clé « grippe » ou les noms des symptômes de la grippe (toux, fièvre, maux de gorge) sur un rayon de 27 km environ et couvrant 11 villes américaines. Ils ont trouvé une forte corrélation entre ces micromessages et des cas

confirmés de grippe par les statistiques du centre de contrôle et de prévention des maladies (CDC). Culotta [36] a trouvé des résultats similaires à partir d'une analyse de plus de 500 000 messages postés sur Twitter sur une période de 10 semaines.

Cette revue de la littérature a permis de montrer l'importance accordée par les autorités canadiennes à la surveillance de l'épidémie de la grippe. Elle a permis de montrer aussi qu'au Canada, des sources de données en dehors du système classique de surveillance sont utilisées pour la surveillance de la grippe. D'autre part, la majorité des études ci-dessus ont mis en évidence l'influence positive des réseaux sociaux sur l'amélioration des soins de santé. Toutefois, le nombre des études réalisées sur l'utilisation du réseau social Twitter au Canada pour la surveillance des épidémies demeure très limité.

Chapitre 3

Problématique

L'arrivée du Web 2.0 a vu la création d'un nouveau type d'application qui a changé la manière dont les rapports humains fonctionnaient. Avec cette technologie, on assiste aujourd'hui au développement des objets connectés, l'apparition des nouvelles applications notamment les réseaux sociaux.

En effet, aujourd'hui, les réseaux sociaux sont au cœur des nouvelles technologies de la communication du fait qu'ils permettent la mise en relation des individus de divers horizons, des personnes ayant les mêmes affinités de se regrouper, et de partager des informations et des idées. Les internautes ne sont plus des simples consommateurs, mais participent au développement des contenus. Dans cette optique, les patients ont tendance à prendre le contrôle de leur santé en main. Ils créent des groupes de discussion sur des sujets liés à leur santé, en partageant leur expérience, leur combat face à certaines maladies.

Une importante quantité d'information traitant des sujets très variés circulent sur les réseaux sociaux notamment des informations relatives à la santé.

La surveillance de l'état de santé de la population est au cœur de la science de la santé publique et constitue l'un des objectifs de la politique de santé publique au Canada. Elle a recours à diverses sources des données pour fournir des informations en temps réel aux décideurs. À l'heure actuelle au Canada, la surveillance de l'influenza est faite grâce au croisement de plusieurs données qui permettent une analyse complète de la survenue et de l'expansion de l'épidémie de la grippe dans tout le Canada. Cette activité de surveillance réunit plusieurs acteurs de différents secteurs : le ministère de la Santé et l'Agence canadienne de la santé publique, les laboratoires d'hôpitaux, les réseaux de surveillance nationale, les médecins du réseau sentinelle, les agents des pharmacies. Lors de la pandémie de la grippe de 2009 dans la province du Québec, cette surveillance s'est étendue aux établissements scolaires. Ces derniers contribuent d'une manière indirecte à la surveillance en fournissant des données sur le nombre d'élèves absents du fait de la grippe,

les parents absents du travail pour des raisons de maladies de leurs enfants. L'étude réalisée par Kom Mogto et ses collègues [37] a montré qu'il existe une corrélation entre l'absentéisme et les autres indicateurs notamment les cas d'hospitalisation. Selon Schmidt, Pebody, et Mangtani [38], ce type de surveillance est peu coûteux à mettre en place, mais demeure cependant très limité du fait de l'incohérence des données causée par l'indisponibilité des enfants pendant les périodes de congés ou les weekends.

L'ASPC reconnait que les méthodes traditionnelles de collecte de données de surveillances de l'influenza souffrent de quelques insuffisances qui pourraient être corrigées par méthodes permettant l'élargissement ses sources d'information. Au nombre de ces insuffisances, l'on peut citer :

- Le décalage entre le moment où les individus sont infectés et le moment où les données sont reçues et analysées;
- L'incohérence des données causée par l'indisponibilité des enfants pendant les périodes de congés ou les weekends;
- Le nombre de cas d'hospitalisation ne reflète pas en général la population, car la plupart du temps, ces rapports ont tendance à représenter les personnes âgées et les enfants très jeunes dus à leur santé fragile.

L'utilisation des données de réseaux sociaux apparait comme une opportunité pour combler certaines limites du système traditionnel. De manière spécifique, le caractère d'instantané des informations issues des réseaux sociaux présente l'avantage d'un suivi et de l'analyse de l'évolution d'une épidémie en temps réel. Cette dernière pourrait en faciliter la détection et servir de base pour un plan de riposte rapide. L'autre utilité des réseaux sociaux est de pouvoir tirer profit des nombreuses informations produites par les internautes pour améliorer la qualité de services des soins de santé.

C'est dans ce cadre que s'inscrit la présente essai qui vise à répondre à la problématique suivante : est-ce que l'on pourrait utiliser les techniques d'analyse des données massives sur les données extraites des réseaux sociaux pour faire de l'analyse épidémiologique?

Pour atteindre ce but, plusieurs objectifs intermédiaires sont requis :

- Identifier les noms des symptômes qui peuvent servir de mots-clics;
- Analyser les micromessages contenant ces mots-clés;

- Mesurer le temps mis pour obtenir les résultats des analyses avec l'API Twitter;
- Montrer qu'il est possible, sur le plan de la démarche (choix des symptômes servant de mots-clés), de temps et de programmes, d'extraire et d'organiser les données de Twitter pour une analyse;
- Vérifier l'existence d'une corrélation positive entre les méthodes d'analyse Twitter et les cas confirmés de grippe en rétrospective dans les laboratoires.

La figure 3-1 illustre le cadre conceptuel de cette recherche et fait état des méthodes d'analyse des données Twitter et des méthodes traditionnelles. Il n'est cependant pas nécessaire de prouver laquelle des deux méthodes est la meilleure, mais simplement de démontrer que les méthodes d'analyse du réseau Twitter permettent d'avoir plus rapidement les informations liées à une épidémie et ainsi de pouvoir préparer une réponse rapide.

Les cas confirmés en laboratoire sont considérés comme la variable indépendante et le nombre de micromessages par semaine contenant les mots-clés décrivant les symptômes de la grippe sont utilisés comme variable dépendante.

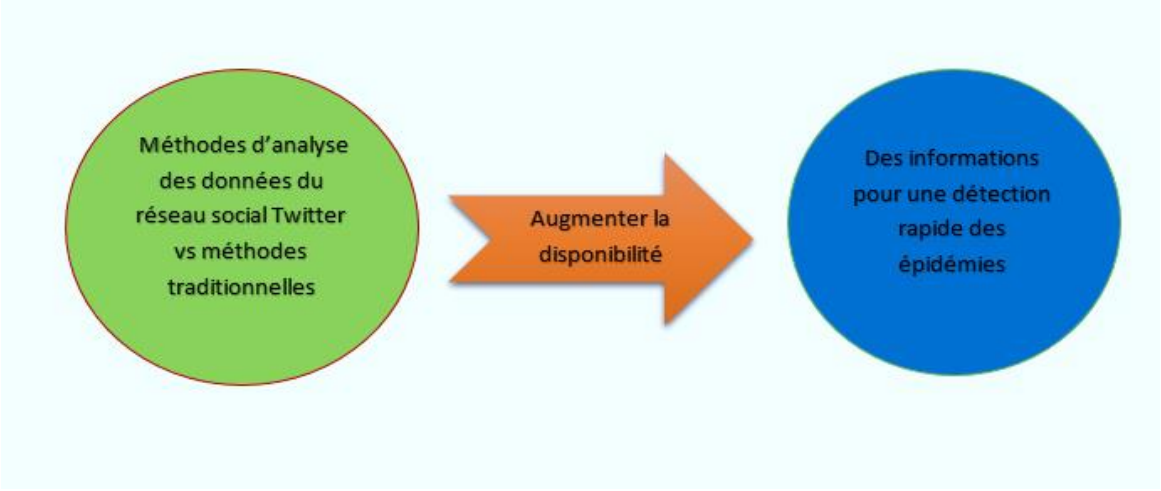

Figure 3-1 Cadre conceptuel de l'essai

3.1.1 Objectifs et hypothèses

Les réseaux sociaux sont devenus de nos jours des plateformes de plus en plus utilisées pour le partage de l'information entre familles, amis, relations, etc. Ces informations partagées par les internautes, se rapporte en général à leurs expériences quotidiennes, c'est-à-dire leur vécu, ce qui se passe autour d'eux, mais aussi ce qui se passe dans les médias. En cas d'épidémie, ces internautes ont tendance à partager l'information non seulement pour alerter les autres abonnés, mais aussi pour exprimer leur frustration des services de santé.

L'analyse du contenu des communications relatives à ces phénomènes permet d'établir une corrélation avec le nombre de cas de malades dans la région et suivre ainsi l'évolution de la situation.

Afin de mieux explorer le sujet, la question de recherche s'intitule : « Les données provenant de Twitter et portant sur une épidémie peuvent-elles être corrélées positivement aux données épidémiologiques provenant des sources traditionnelles? »

L'hypothèse s'annonce comme :

Les données massives issues du réseau social Twitter concernant une épidémie peuvent être corrélées positivement aux données épidémiologiques provenant des sources traditionnelles. La recherche est de type quantitatif, corrélationnel.

Cette question de recherche comporte plusieurs volets à savoir :

- L'utilisation des données issues des réseaux sociaux notamment Twitter permet de faire de la surveillance épidémiologique de la grippe.
    - Elle permet d'anticiper plus rapidement la survenue de l'épidémie de la grippe au sein de la population par rapport les systèmes existants;
    - Elle permet de suivre en temps réel son évolution, le moment où elle atteint son maximum et la fin de l'épidémie.
- La disponibilité et l'accès à des données en temps réel permettront l'extraction, l'analyse et l'information de l'opinion publique sur les mesures de santé publique à prendre en une journée.

3.1.2 Limites

Les technologies des données massives offrent de multiples opportunités dans le domaine médical (nouvelles applications médicales et thérapeutiques) qui révolutionne les méthodes de travail de la santé publique.

Plusieurs méthodes sont utilisées pour faire de la surveillance épidémiologique. Par exemple, « Asthmapolis » collecte des données avec sa technologie mobile incorporée à l'inhalateur des malades grâce au GPS. L'analyse de ces données permet de faire la corrélation des habitudes des malades et leur environnement, et d'en savoir plus sur les agents catalyseurs de l'asthme. Cette étude ne porte pas sur l'utilisation des capteurs ni l'envoi des formulaires de demandes d'informations, mais plutôt sur l'utilisation des données des réseaux sociaux et fait une comparaison avec les données existantes des cas réels d'épidémies.

Chapitre 4

Démarche méthodologique

4.1 Approche proposée

Pour répondre à la problématique de surveillance de la grippe au sein de la population canadienne décrite dans le précédent chapitre, la méthodologie utilisée est basée sur une approche de recherche quantitative expérimentale.

Elle consiste à démontrer qu'il existe une corrélation entre les données portant sur la grippe extraite de la base de données du réseau social Twitter et les données publiées par ASPC (cas confirmés en laboratoire). La variable dépendante est le nombre de cas de grippe hebdomadaire issu du réseau social. La variable indépendante est le nombre de cas de grippes confirmés issus des données de surveillance de l'Agence canadienne de santé publique.

Pour ce faire, il faut dans un premier temps, définir la période de la grippe saisonnière au Canada et dresser un graphe de référence. Ensuite, sélectionner les mots-clés à utiliser pour l'analyse et enfin, procéder à la collecte de données, à l'analyse et la présentation sous forme de tableaux et de représentation graphique.

4.2 Population cible

La population ciblée par l'étude est l'ensemble des utilisateurs qui disposent d'un compte sur le réseau social Twitter et qui a publié des messages contenant les mots-clics liés à la grippe. Ces messages doivent être publiés à partir du Canada, peu importe le sexe, et l'âge de l'utilisateur.

4.3 Échantillon

L'échantillon de cette étude est l'ensemble des messages et mots-clics du réseau social Twitter qui sont relatifs aux symptômes de la grippe saisonnière dont la date de publication se situe dans une période donnée.

4.4 Déroulement de l'essai

L'objectif de cet essai est d'analyser la corrélation existante entre les données issues des méthodes de surveillance traditionnelles et celles de Twitter et de comparer par la suite le temps mis par ces méthodes pour rendre disponibles les informations sur une épidémie. Le diagramme de la figure 4.1 explique les différentes étapes qui ont été suivies pour le déroulement de l'essai.

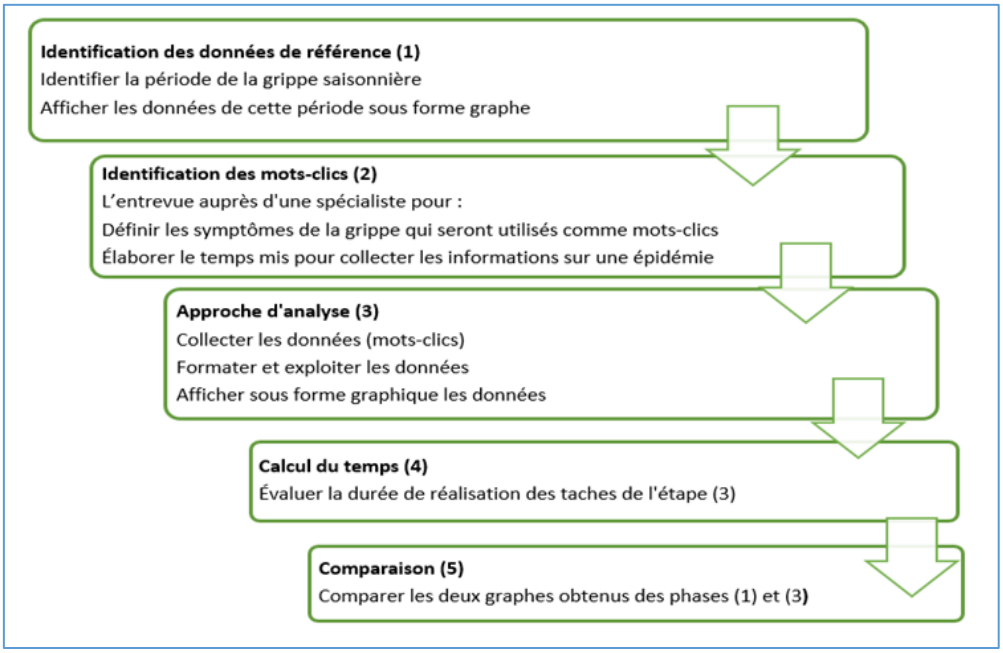

Figure 4-1 Ordonnancement des tâches de l'essai

4.4.1 Identification des données de référence

Les données de référence de cette recherche sont obtenues à partir du portail de l'Agence de santé publique du Canada [39]. Ce portail permet d'obtenir des informations sur la surveillance de l'influenza, d'accéder aux rapports hebdomadaires sur l'activité grippale au Canada et bien d'autres informations. Dans le cadre de cet essai, l'étape d'identification permet de choisir une période grippale et les semaines épidémiologiques associées à un rapport d'influenza.

Les données de l'Agence canadienne proviennent des rapports d'analyse des laboratoires, des diagnostics des cabinets médicaux, il s'agit donc des cas confirmés de grippe. Elles sont agrégées chaque semaine et publiées la semaine suivante en d'autre termes décalés sur une semaine. Les données publiées cette semaine sont en fait le nombre de cas diagnostiqués et confirmés de la semaine dernière.

4.4.2 Identification des mots-clics

Cette phase a pour objectif de définir et de déterminer les symptômes de la grippe saisonnière pouvant servir de mot-clic. Une entrevue avec une spécialiste a permis d'obtenir les mots-clics à utiliser pour la phase d'analyse. La grille de collecte de discussion est présentée en annexe A.

L'entrevue avec les spécialistes de la santé publique constitue une étape importante dans l'essai dans la mesure où il est essentiel de s'assurer d'avoir fait de bons choix de mots couramment utilisés par les utilisateurs de Twitter pour exprimer leur mal-être.

4.4.3 Approche d'analyse des données du Twitter

Cette phase qui porte sur l'utilisation de Twitter illustrée sur la figure 4.2, est réalisée en trois étapes distinctes : (i) la collecte des données, (ii) le formatage et l'exploitation des données (iii) la réalisation des courbes d'évolution pour une bonne observation des similitudes avec les données recueillies auprès des agences de la santé publique.

Figure 4-2 Démarche méthodologique

4.4.3.1 Collecte de données

Twitter est le site de contenu le plus fréquemment mis à jour ; environ 500 millions de messages sont publiés quotidiennement sur sa plateforme.

- La majorité des messages publiés par les abonnés de Twitter sont publics, ainsi leur collecte et leur traitement se font plus facilement grâce à l'API Twitter;
- Il est conçu pour encourager les utilisateurs à partager leurs expériences quotidiennes;
- Les micromessages sont généralement plus longs et plus descriptifs que les mots utilisés sur les moteurs de recherche.

Twitter dispose de plusieurs API qui permettent d'interroger la base de données Twitter, mais aussi d'en faire des plateformes. Elles retournent des centaines de variables par requête sous format brut (json) qu'il faut transformer pour les rendre exploitable.

La collecte de données issues des comptes des abonnés du réseau social Twitter se fait à partir de l'API Twitter. Ce sont des publications des commentaires des partages des micromessages des abonnés de Twitter qui contiennent les mots clics cités précédemment.

L'API Twitter permet à l'application de se connecter sur la base de données Twitter et d'extraire les informations recherchées en fonction des mots-clés qui sont entrés pour effectuer la recherche. Cette collecte tient compte des paramètres suivants :

- L’analyse des micromessages qui contiennent les mots-clics;
- L’endroit de la publication des micromessages doit se faire à l’intérieur du Canada;
- La période de publication des micromessages doit se situer dans l’intervalle de la période de la grippe saisonnière.

4.4.3.2 Formatage et exploitation de données

Les données collectées sur Twitter sont généralement des données brutes non structurées. Les API retournent les résultats d’une requête faite sur les bases de données Twitter dans un format brut (json). Cette étape, dans un premier temps, permet d’uniformiser les données et les rend manipulables dans une interface de visualisation (Excel) afin de procéder à l’exploitation de ces données.

4.4.3.3 Présentation des résultats

La présentation des résultats se fait sous une forme plus compréhensive et visuelle (représentation graphique). La présentation graphique organise visuellement les éléments d’informations et met l’accent sur le lien entre les éléments afin d’établir la comparaison. Cette méthode permet entre autres de représenter les éléments sous une forme claire et précise dans un espace limité, de démontrer les changements qui existent et les moments où ces changements sont produits ce qui permet de voir clairement les tendances.

4.4.4 Calcul du temps

Dans cette phase, l’objectif sera de mesurer le temps entre le début d’une épidémie et son dépistage. Il s’agit d’estimer le temps mis pour réaliser les différentes phases de l’étude : la collecte de données, le formatage et l’exploitation de données et la présentation sous forme graphique des résultats.

4.4.5 Comparaison

La comparaison concerne le temps effectué pour réaliser la démarche méthodologique (méthode d’analyse Twitter) et le temps que prend le programme de surveillance de l’influenza pour rendre disponibles les informations de l’épidémie de la grippe saisonnière

(méthodes traditionnelles). Cette différence permet d'affirmer ou non si la méthode d'analyse des données du réseau social Twitter permet d'augmenter la disponibilité des informations pour une détection rapide des épidémies.

4.5 Résultats attendus

Les résultats attendus sont présentés sous une forme graphique dans un espace de temps limité. Les deux courbes (celle de la méthode traditionnelle et Twitter) seront représentées sur le même graphe, afin de permettre de visualiser les périodes de similitudes ou de différences et de décrire la corrélation existante entre ces deux méthodes.

Le chapitre suivant analyse amplement les résultats et démontre si oui ou non il existe une corrélation entre ces méthodes d'analyses. Cette analyse est suivie de la présentation de la mesure de la durée d'exécution de ces méthodes pour rendre disponibles les informations de l'épidémie dans un meilleur délai.

Chapitre 5

Analyse des résultats

Dans ce chapitre, nous allons présenter les résultats permettant de confirmer ou d'infirmer l'hypothèse de recherche. L'hypothèse de départ est que les données massives issues (micromessages des abonnés) du réseau social Twitter concernant une épidémie peuvent être corrélées positivement aux données épidémiologiques provenant des sources traditionnelles. Dans un premier temps, nous allons présenter les résultats obtenus, ensuite vérifier l'hypothèse et enfin démontrer la validité de la conclusion.

5.1 Résultats obtenus

5.1.1 Identification des données

a) Les données de référence

Au Canada, la saison de la grippe s'étend généralement de novembre à avril. Les données de référence sont les cas de grippe publiés par l'Agence de santé publique du Canada. Les données sont le nombre de cas de grippes confirmés par les laboratoires, les cliniques et médecins du réseau de surveillance de l'Agence canadienne de santé publique. Elles portent sur la période du 18 décembre 2016 au 28 janvier 2017 ce qui correspond aux semaines 51 à 54 soit 5 semaines de données.

La figure 5-1, illustre un extrait d'un rapport de l'activité de la grippe de l'année 2016 – 2017. Elle indique que le pic de l'épidémie de la grippe saisonnière de 2016-2017 a été observé durant la semaine 2 avec plus de 3000 tests positifs. Partant de ce fait, il a paru important de choisir la période de prise d'échantillon, deux semaines avant le début du pic de la grippe et deux semaines après, pour avoir une meilleure représentation des données et mieux illustrer ce qui se passe durant cette période.

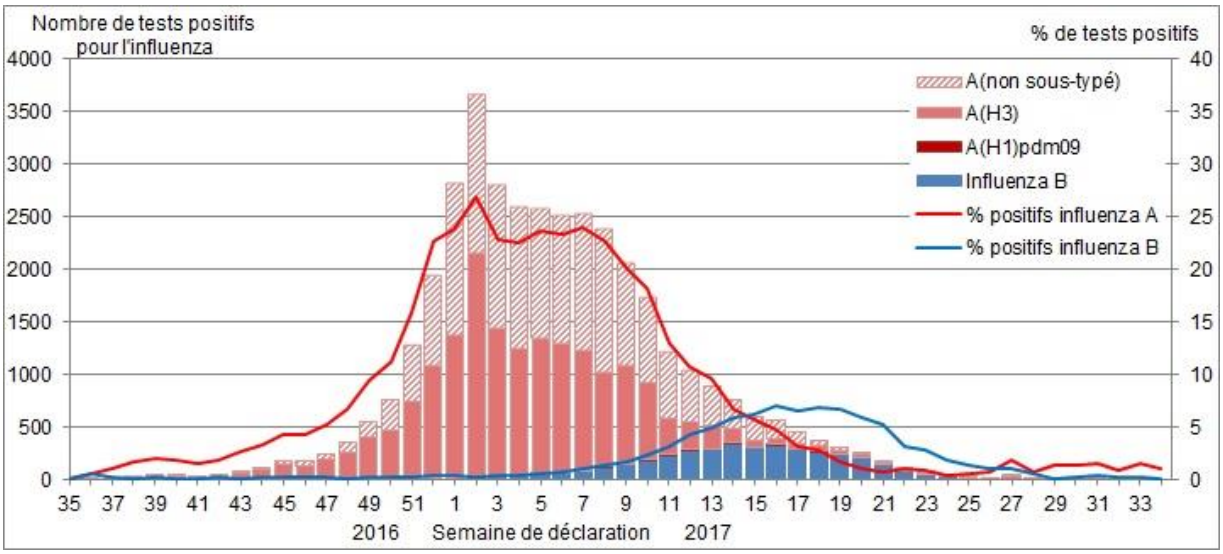

Figure 5-2 Courbe évolutive de la grippe au Canada 2016-2017[1]

b) Les données de Twitter

Les données issues de Twitter sont les micromessages des abonnés émis pendant la période correspondant à la période retenue au Canada et contenant au moins l'un des mots-clics retenus. Une entrevue avec une spécialiste de la santé publique a permis d'obtenir les mots-clics les plus couramment utilisés par les patients pour exprimer les symptômes de la grippe. Il s'agit de : « grippe », « maux de gorge », « fièvre », « j'ai la courbature », « j'ai le nez qui coule », « maux de tête » et « je suis congestionné ».

Compte tenu des coûts liés à l'acquisition des données qui dépend du nombre de jours demandés et du nombre de mots-clics ; il a été retenu dans le cadre de cet essai de se limiter uniquement aux 3 mots-clics les plus importants, à savoir « Flu », « influenza », « grippe ».

Ces mots-clics seront utilisés pour construire une requête qui a été transmise à la compagnie Twitter pour l'extraction des données. Cette requête retourne tous les micromessages du Canada contenant le mot « *grippe* » et tous les micromessages du monde entier contenant les mots « *Flu »* et ceux contenant le mot « *influenza* ». Une fois les données téléchargées et

[1] Tirée de la page web du gouvernement du Canada sur la grippe.

analysées, une requête supplémentaire a été faite pour extraire les données du Canada uniquement.

5.1.2 Analyse comparative des résultats

Au total, 37 337 micromessages de plus 27 pays ont été reçus. Cependant, 98 % des données reçues ne comportaient pas le pays d'émission du tweet. On n'était pas à mesure de déterminer de quel pays ils provenaient. Donc nous avons travaillé uniquement sur les 2 % des micromessages valides ce qui représente 763 micromessages.

Le graphique 5.2 présente la répartition des micromessages par pays d'émission. Le nombre important de micromessages ont été exclus de l'analyse à cause de l'origine inconnue de certains messages, entachent la qualité des données. Toutefois, bien que les données du Canada représentent seulement 1 % des données totales, elles représentent 27 % des données valides.

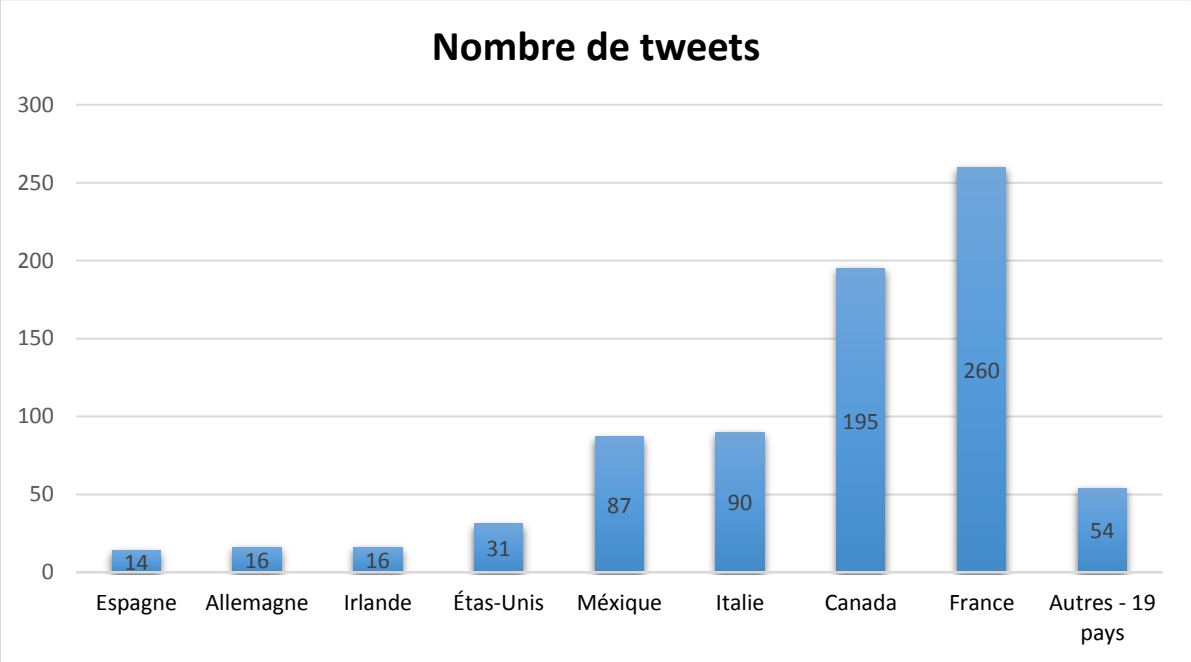

Figure 5-3 Nombres de micromessages par pays d'émission

5.1.2.1 Interprétation du graphe

Le graphique ci-dessous représente le nombre de cas de grippe enregistrés par l'agence de la santé publique du Canada dans le cadre de la surveillance, il est représenté sur le graphique en ligne discontinue et le nombre de notifications de grippes extraites des données de Twitter, en ligne continue.

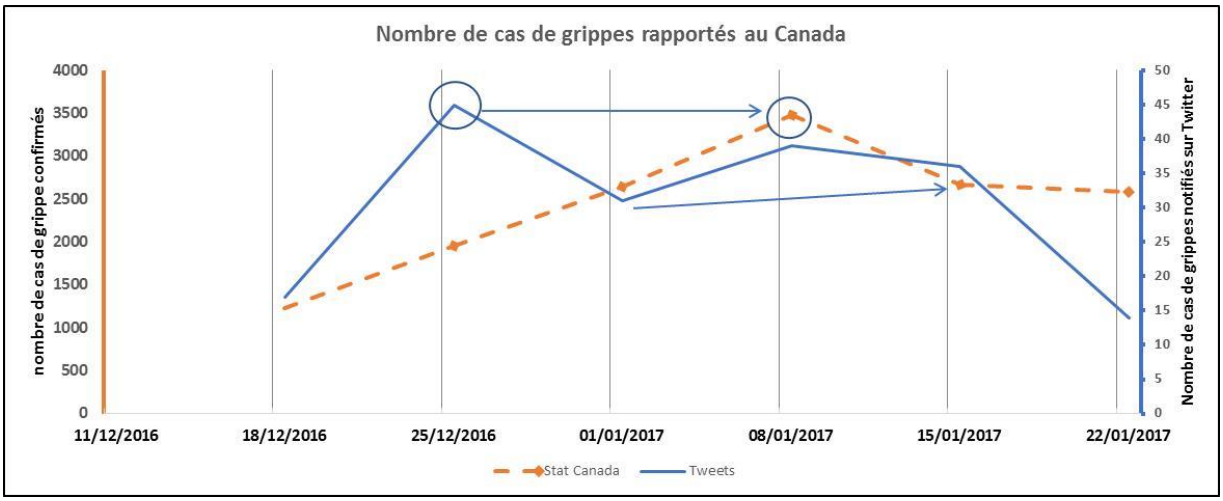

Figure 5-4 Nombre de cas de grippe rapportés au Canada

Sur ce graphique, il y a trois faits majeurs qui méritent d'être soulevés :

- Les deux courbes ont la même forme : un début d'épidémie, suivi d'une période de croissance jusqu'à atteindre un maximum puis elles amorcent une baisse avant de stabiliser;
- Le pic observé sur la courbe des données de Twitter est observé sur la courbe des données de statistique Canada, mais avec un décalage deux semaines;
- Les données de Twitter connaissent une chute brutale dans la période du 31 décembre au 1er janvier, cette même chute sera observée deux semaines plus tard sur la courbe de statistique Canada.

En conclusion, les deux courbes sont corrélées avec un décalage de deux semaines.

5.1.2.2 Estimation du temps de travail

Le tableau 5-1 présente les différentes étapes qui ont conduit à la réalisation de l'analyse et le temps estimé de travail. Il en ressort que le travail qui a demandé le plus de temps est la partie qui correspond à l'acquisition des données.

- Durant la phase de collecte, les données étant payantes et protégées, il y a une démarche administrative à suivre. Cette phase s'étend sur une longue période (la signature de contrat entre les deux parties, le virement bancaire et enfin l'envoi du lien pour s'acquitter des données en écrivant sur un terminal une commande bash).
- Le formatage et l'exploitation de données : l'application IntelliJ_IDEA, environnement de développement intégré (IDE) Java, a permis de télécharger les données dans un système de gestion de base de données MySQL, à l'aide d'un script. La codification s'est faite en une soirée, mais une fois que le script est écrit, il peut servir pour d'autres utilisations, il suffira donc de modifier quelques lignes de code.
- La phase présentation visuelle des résultats s'est faite en utilisant l'outil Excel de Microsoft Office. MySQL permet d'exécuter une requête et d'exporter le résultat dans un format Excel. Cette phase a pris le temps d'écrire une requête et de l'exporter sous Excel.

Numéro	Tâches	Durée
1	Collecte de données	48 heures
2	Formatage et exploitation de données	1 h
3	Présentation visuelle des résultats	1 h
	DURÉE TOTALE	50 h

Tableau 5-1 Les types de réseaux sociaux

5.2 Retour sur les hypothèses

L'hypothèse de départ stipule que l'utilisation des données de Twitter permet d'anticiper plus rapidement sur le début de l'épidémie de grippe, de faire le suivi de son évolution en temps réel.

Pour vérifier cette hypothèse, c'est-à-dire la confirmer ou la rejeter, une comparaison a été faite entre la courbe de données de surveillance de l'épidémie de la grippe de l'ASPC et la courbe réalisée à partir des données de Twitter.

Selon les résultats observés, les données du réseau social Twitter sont corrélées aux données des sources traditionnelles. Ainsi, l'utilisation des données issues des réseaux sociaux notamment Twitter permet de faire de la surveillance épidémiologique de la grippe.

Elles permettent :

- D'anticiper plus rapidement la survenue de l'épidémie de la grippe au sein de la population que le système de surveillance existant;
- De suivre en temps réel son évolution, le moment où elle atteint son maximum et la fin de l'épidémie.

La disponibilité et l'accès à temps réel des données permettront de pouvoir extraire et analyser en une journée et informer l'opinion publique sur les mesures de santé publique à prendre. Mis à part les formalités d'ordre administratives et financières pour l'acquisition des données, le traitement et l'analyse des données de Twitter peuvent se faire comme l'analyse de toute autre donnée. Elle peut se faire en une journée. Toutefois, il existe un coût associé à l'acquisition des données de Twitter ce qui n'est pas le cas pour les autres sources de données de la surveillance classique.

5.3 Démonstration de la validité des résultats

Les résultats similaires pour les données d'un autre pays appuient la validité de la méthode. La même analyse a été effectuée sur les données des États-Unis. Les données de

surveillance traditionnelle sont extraites sur le site du CDC (Center for Disease Control) pour la même période. Les résultats sont présentés sur la figure 5-5.

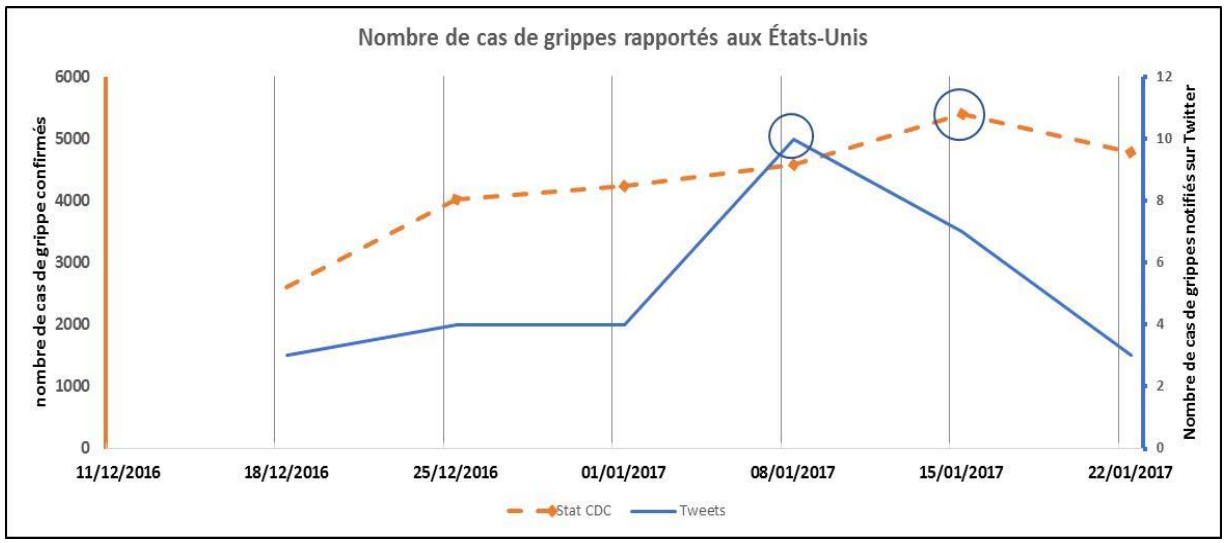

Figure 5-6 Cas de grippe reportés aux É.-U.

Les mêmes faits observés sur le graphe du Canada se reproduisent sur le graphe de l'analyse des données des É.-U. Les courbes ont la même forme, un début d'épidémie suivi d'une période de croissance jusqu'à atteindre un pic puis une baisse. Le maximum et le minimum se produisaient sur la courbe de Twitter deux semaines avant de se produire sur la courbe du CDC. Ces constats semblent concorder avec les conclusions tirées après l'analyse des données du Canada.

Conclusion

Notre étude a montré que sur la saison grippale 2016-2017, l'évolution des cas de grippe confirmés en laboratoire (sources traditionnelles) suit celle déterminée par l'analyse des métadonnées du réseau social Twitter.

L'estimation du temps de travail a montré qu'en dehors des formalités d'ordre administratives et financières pour l'acquisition des données Twitter, l'analyse de ces données se fait dans un délai réduit (entre 24 à 48 heures).

Aux vues des points précédents l'hypothèse selon laquelle une corrélation existe entre les données épidémiologiques de sources traditionnelles et la fréquence d'apparition des mots-clics sur le réseau social Twitter est vérifiée.

Ainsi les données issues du réseau social Twitter pourraient servir à des fins de surveillance épidémiologique de la grippe. Elles permettent (i) d'anticiper plus rapidement la survenue de l'épidémie de la grippe au sein de la population par rapport aux systèmes existants ; (ii) de suivre en temps réel son évolution, le moment où elle atteint son maximum et la fin de l'épidémie.

Il existe cependant un coût additionnel associé à l'acquisition des données de Twitter comparé aux sources de données de la surveillance classique.

Comme toutes les études, cette étude présente des limites. Le cout élevé d'acquisition des données de Twitter a restreint notre étude à quatre semaines et limité par le nombre de mots-clics. En outre, plus de 90 % de nos données étaient inexploitables car les pays de provenance des micromessages n'étaient pas identifiables.

Enfin, à la différence des méthodes traditionnelles qui identifient des cas avérés, les données de Twitter sont des opinions d'utilisateurs, souvent non professionnels du domaine médical. Les données de Twitter ne remplacent donc pas le système traditionnel, mais peuvent cependant être complémentaires.

Pour les recherches ultérieures, nous recommandons d'aller au-delà des limites de notre étude. Il conviendrait d'ajouter des critères supplémentaires pour l'identification des provenances de micromessages :

- Utilisation des adresses IP des abonnés;
- Utilisation des coordonnées géographiques des émetteurs des micromessages.

En outre, nous suggérons aussi d'augmenter le nombre de mots-clics et de considérer une période plus étendue pour mieux cerner les changements dans le temps.

Liste des références

[1] World Health Organization, « eHealth at WHO », *World Health Organization*, 2016. [En ligne]. Disponible à: http://www.who.int/ehealth/about/en/. [Consulté le: 03-févr-2017].

[2] J.-J. JEGOU, « Rapport d'information fait au nom de la commission des Finances, du contrôle budgétaire et des comptes économiques de la Nation sur l'informatisation dans le secteur de la santé », *Ump-Senat.Fr*, 2006. [En ligne]. Disponible à: http://www.ump-senat.fr/IMG/pdf/audiovisuel_public_-_l_heure_du_bilan.pdf. [Consulté le: 10-janv-2017].

[3] Synthesio, « Les medias sociaux au service de la santé », 2010. [En ligne]. Disponible à: https://www.synthesio.com/wp-content/uploads/2011/01/Les-médias-sociaux-au-service-de-la-santé-2.pdf. [Consulté le: 20-mars-2017].

[4] S. Fox, « The Social Life of Health Information, 2011 », *Pew Internet Am. Life Proj.*, p. 1‑33, 2011.

[5] Statistic Canada, « Canadian Internet Use Survey , 2012 », *Statistique Canada*, 2012. [En ligne]. Disponible à: https://www.statcan.gc.ca/daily-quotidien/100510/dq100510a-eng.htm. [Consulté le: 10-août-2017].

[6] J. A. Barnes, « Class and Committees in a Norwegian Island Parish », *Hum. Relations*, vol. 7, n° 1, p. 39‑58, 1954.

[7] E. Lazega, « Analyse de réseaux et sociologie des organisations », *Rev. française Sociol.*, vol. 35, n° 2, p. 293‑320, 1994.

[8] A. M. Kaplan et M. Haenlein, « The challenges and opportunities of Social Media », *Bus. Horiz.*, vol. 53, n° 1, p. 59‑68, 2010.

[9] W. Youyou, M. Kosinski, et D. Stillwell, « Computer-based personality judgments are more accurate than those made by humans », *Proc. Natl. Acad. Sci.*, vol. 112, n° 4, p. 1036‑1040, 2015.

[10] T. O'Reilly, « What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software », *Commun. Strateg.*, vol. 1, n° First Quarter, p. 17, 2007.

[11] T. Slenger et A. Coutant, « Médias sociaux : clarification et cartographie Pour une approche sociotechnique », *Décisions Mark.*, p. 107‑117, 2013.

[12] M. Thelwall, « Social Network Sites : Users and Uses », *Advances*, vol. 76, n° 9, p. 19‑73, 2009.

[13] Olivier Ertzscheid, « Culture documentaire et folksonomie : l'indexation à l'ère industrielle et collaborative », *Doc. - Sci. l'information*, vol. 47, n° 1, p. 42‑45, 2010.

[14] Kap – Tactiques numériques, « Les chiffres du numérique au Canada en 2017 », 2017. [En ligne]. Disponible à: http://www.kap-numerique.com/chiffres-numerique-canada-2017/. [Consulté le: 20-févr-2017].

[15] « Larousse », *Dictionnaire*, 2016. [En ligne]. Disponible à: http://www.larousse.fr/encyclopedie/divers/santé_publique/90008#HDheskttLXPZ35yl.99. [Consulté le: 26-févr-2017].

[16] Administrateur en chef de la santé publique, « Rapport sur l'état de la santé publique au Canada 2013 », 2013.

[17] OMS, « Grippe (saisonnière) », *Aide memoire*, 2016. [En ligne]. Disponible à: http://www.who.int/mediacentre/factsheets/fs211/fr/. [Consulté le: 03-févr-2017].

[18] M. Forsé, « Les réseaux sociaux chez Simmel: les fondements d'un modèle individualiste et structural », *Deroche-Gurcel L., Watier P.(sous la dir.), La Sociol. Simmel (1908), Eléments actuels modélisation Soc. PUF, coll. Sociol.*, p. 63‑109, 2002.

[19] B. Divjak et P. Peharda, « Social network analysis of study environment », *J. Inf. Organ. Sci.*, vol. 34, n° 1, p. 67‑80, 2010.

[20] Brittany C. Campbell et Clay M. Craig, « Health Professions Students Academic and Personal Motivations for Using Social Media », *Int. J. Commun. Heal.*, vol. 3, p. 8, 2014.

[21] CDC, « The Health Communicator's Social Media Toolkit », *2014*. [En ligne]. Disponible à: https://www.cdc.gov/socialmedia/tools/guidelines/socialmediatoolkit.html.

[22] J. C. Bertot, P. T. Jaeger, et J. M. Grimes, « Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies », *Gov. Inf. Q.*, vol. 27, n° 3, p. 264‑271, 2010.

[23] J. B. Houston *et al.*, « Social media and disasters: A functional framework for social media use in disaster planning, response, and research », *Disasters*, vol. 39, n° 1, p. 1‑22, 2015.

[24] M. Househ, « The use of social media in healthcare: Organizational, clinical, and patient perspectives », *Studies in Health Technology and Informatics*, vol. 183. p. 244‑248, 2013.

[25] B. Chauhan, R. George, et J. Coffin, « Social media and you: what every physician needs to know. », *J. Med. Pract. Manag.*, vol. 28, n° 3, p. 206‑9, 2012.

[26] J. M. Farnan, L. S. Sulmasy, B. K. Worster, H. J. Chaudhry, J. A. Rhyne, et V. M. Arora, « Online medical professionalism: Patient and public relationships: Policy statement from the American College of physicians and the federation of State Medical Boards », *Ann. Intern. Med.*, vol. 158, n° 8, p. 620‑627, 2013.

[27] T. Coëffé, « Le Blog du Modérateur », *Chiffres Internet – 2017*, 2017. [En ligne]. Disponible à: https://www.blogdumoderateur.com/chiffres-internet/. [Consulté le: 26-juin-2017].

[28] Agence de la santé publique du Canada, « Page facebook de l'ASPC ». [En ligne]. Disponible à: https://www.facebook.com/Agence-de-la-santé-publique-du-Canada-14498271095/. [Consulté le: 10-avr-2017].

[29] L. Gilpin, « techrepublic », *How an algorithm detected the Ebola outbreak a week early, and what it could do next*, 2014. [En ligne]. Disponible à: https://www.techrepublic.com/article/how-an-algorithm-detected-the-ebola-outbreak-a-week-early-and-what-it-could-do-next/. [Consulté le: 01-juin-2017].

[30] L. Gamaury, « Twitter : un outil fiable à 90% pour évaluer la propagation de la grippe », *La rédaction*, 2012. [En ligne]. Disponible à: http://www.terrafemina.com/culture/culture-web/articles/16266-twitter-un-outil-fiable-a-90-pour-evaluer-la-propagation-de-la-grippe.html. [Consulté le: 10-juin-2017].

[31] F. Pervaiz, M. Pervaiz, N. A. Rehman, et U. Saif, « FluBreaks: Early epidemic detection from google flu trends », *J. Med. Internet Res.*, vol. 14, n° 5, 2012.

[32] D. R. George, L. S. Rovniak, et J. L. Kraschnewski, « Dangers and opportunities for social media in medicine. », *Clin. Obstet. Gynecol.*, vol. 56, n° 3, p. 453‑62, 2013.

[33] N. S. Fogelson, Z. a Rubin, et K. A. Ault, « Beyond likes and tweets: an in-depth look at the physician social media landscape. », *Clin. Obstet. Gynecol.*, vol. 56, n° 3, p. 495‑508, 2013.

[34] A. M. Cameron *et al.*, « Social media and organ donor registration: The Facebook effect », *Am. J. Transplant.*, vol. 13, n° 8, p. 2059‑2065, 2013.

[35] A. A. Aslam *et al.*, « The reliability of tweets as a supplementary method of seasonal influenza surveillance », *J. Med. Internet Res.*, vol. 16, n° 11, 2014.

[36] A. Culotta, « Towards detecting influenza epidemics by analyzing Twitter messages », *1st Work. Soc. Media Anal.*, n° May, p. 115‑122, 2010.

[37] C. A. Kom Mogto *et al.*, « School absenteeism as an adjunct surveillance indicator: experience during the second wave of the 2009 H1N1 pandemic in Quebec, Canada. », *PLoS One*, vol. 7, n° 3, 2012.

[38] W. P. Schmidt, R. Pebody, et P. Mangtani, « School absence data for influenza surveillance: A pilot study in the United Kingdom », *Eurosurveillance*, vol. 15, n° 3, p. 1‑6, 2010.

[39] Gouvernement du Canada, « Surveillance de l'influenza », *Rapports hebdomadaires sur l'influenza, saison 2016–2017*, 2017. [En ligne]. Disponible à: https://www.canada.ca/fr/sante-publique/services/maladies/grippe-influenza/surveillance-influenza/rapports-hebdomadaires-saison-2016-2017.html.

[Consulté le: 01-mai-2017].

Bibliographie

Hayward, A. C., Fragaszy, E. B., Bermingham, A., Wang, L., Copas, A., Edmunds, W. J., … Zambon, M. (2014). Comparative community burden and severity of seasonal and pandemic influenza: Results of the Flu Watch cohort study. The Lancet Respiratory Medicine, 2 (6), 445–454. https://doi.org/10.1016/S2213-2600 (14)70034-7

Dizon, D. S., Graham, D., Thompson, M. A., Johnson, L. J., Johnston, C., Fisch, M. J., & Miller, R. (2012). Practical guidance: the use of social media in oncology practice. J Oncol Pract, 8 (5), e114-24. https://doi.org/10.1200/jop.2012.000610

Aslam, A. A., Tsou, M. H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., … Lindsay, S. (2014). The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of Medical Internet Research*, *16* (11). https://doi.org/10.2196/jmir.3532

Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T., & Lipsitch, M. (2010). Absolute humidity and the seasonal onset of influenza in the continental United States. PLoS Biology, 8 (2). https://doi.org/10.1371/journal.pbio.1000316

Achrekar H, Gandhe A, Lazarus R, Ssu-Hsin Yu, Liu B. Predicting Flu Trends using Twitter data. In IEEE; 2011 [cité 18 déc. 2017]. p. 702- 7. Disponible sur : http://ieeexplore.ieee.org/document/5928903/

Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. avr 2015; 35 (2):137 - 44.

Han Hu, Yonggang Wen, Tat-Seng Chua, Xuelong Li. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access. 2014; 2:652- 87.

Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 14 mars 2014; 343 (6176):1203 - 5.

Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems [Internet]. déc 2014 [cité 18 déc 2017]; 2 (1). Disponible sur : http://link.springer.com/10.1186/2047-2501-2-3

Thiébaut R, Hejblum B, Richert L. L'analyse des « Big Data » en recherche clinique. Revue d'Épidémiologie et de Santé publique. févr 2014 ; 62 (1) : 1-4.

Annexe A

Grille du questionnaire avec l’experte

L’entrevue avec les spécialistes de la santé publique constitue une étape importante dans l’essai. Il est donc essentiel de s’assurer d’avoir fait de bons choix de mots couramment utilisés par les utilisateurs de Twitter pour exprimer leur mal-être.

Les questions suivantes lui ont été posées.

Questionnaire de l’entrevue
Les questions posées permettent de bien cerner les bons mots-clics et de pouvoir ainsi valider cette phase d’identification : • Quels sont les symptômes de la grippe saisonnière ? • Quels sont les mots utilisés par les patients pour décrire les symptômes de la maladie lors des consultations chez les médecins ? Combien de temps l’agence de la santé publique met-elle pour rendre disponibles les données de l’épidémie de la grippe d’une saison donnée ?

Annexe B

Requête pour l’acquisition des données via la plateforme Twitter

Les données reçues de Twitter portent sur la période de 18/12/2016 au 22/01/2017.

La requête est basée sur une combinaison avec des opérateurs logiques de 4 mots-clés : Canada, Influenza, Flu, grippe.