Université de Montréal

L'utilité des médias sociaux pour la surveillance épidémiologique : une étude de cas de Twitter pour la surveillance de la maladie de Lyme

Par

Elda Kokoe Elolo Laison

École de Santé Publique de l'Université de Montréal

Département de Médecine Sociale et préventive

Mémoire présenté en vue de l'obtention du grade de Maîtrise en Santé Publique, option Recherche

Décembre, 2022.

Université de Montréal

Unité académique : École de Santé Publique de l'Université de Montréal/Département de Médecine Sociale et préventive

---

Ce mémoire intitulé

L'utilité des médias sociaux pour la surveillance épidémiologique : une étude de cas de Twitter pour la surveillance de la maladie de Lyme

Présenté par

Elda Kokoé Elolo Laison

A été évalué(e) par un jury composé des personnes suivantes

Delphine Bosson-Rieutort

Présidente-rapporteuse

Bouchra Nasri

Directrice de recherche

Jean Noel Nikiema

Membre du jury

Résumé

La maladie de Lyme est la maladie transmise par tiques la plus répandue dans l'hémisphère du Nord. Le système de surveillance des cas humains de la maladie de Lyme est basé sur un système passif des cas par les professionnels de santé qui présente plusieurs failles rendant la surveillance incomplète. Avec l'expansion de l'usage de l'internet et des réseaux sociaux, des chercheurs proposent l'utilisation des données provenant des réseaux sociaux comme outil de surveillance, cette approche est appelée l'infodémiologie. Cette approche a été testée dans plusieurs études avec succès. L'objectif de ce mémoire est de construire une base de données à partir des tweets auto-déclarés, des tweets classifiés et étiquetés comme un cas potentiel de Lyme ou non à l'aide des modèles de classificateurs basés sur des transformateurs comme, BERTweet, DistilBERT et ALBERT. Pour ce faire, un total de 20 000 tweets en anglais en lien avec la maladie de Lyme sans restriction géographique de 2010 à 2022 a été collecté avec la plateforme API twitter. Nous avons procédé au nettoyage la base de données. Ensuite les données nettoyées ont été classifiées en binaire comme cas potentiels ou non de la maladie de Lyme sur la base des symptômes de la maladie comme mots-clés. À l'aide des modèles de classification basés sur les transformateurs, la classification automatique des données est évaluée en premier sans, et ensuite avec des émojis convertis en mots.

Nous avons trouvé que les modèles de classification basés sur les transformateurs performent mieux que les modèles de classification classiques comme TF-IDF, Naive Bayes et autres ; surtout le modèle BERTweet a surpassé tous les modèles évalués avec un score F1 moyen de 89,3%, une précision de 97%, une exactitude de 90% et un rappel de 82,6%. Aussi l'incorporation des émojis dans notre base de données améliore la performance de tous les modèles

d’au moins 5% mais BERTweet a une fois de plus le mieux performé avec une augmentation de tous les paramètres évalués. Les tweets en anglais sont majoritairement en provenance des États-Unis et pour contrecarrer cette prédominance, les futurs travaux devraient collecter des tweets de toutes langues en lien avec la maladie de Lyme surtout parce que les pays européens où la maladie de Lyme sont en émergence ne sont pas des pays anglophones.

Mots-clés : maladie de Lyme, réseaux sociaux, twitter, apprentissage automatique, infodémiologie, BERT, emojis, modèles de classification.

Abstract

Lyme disease is the most common tick-borne disease in the Northern Hemisphere. The surveillance system for human cases of Lyme disease has several flaws which make the surveillance incomplete. Nowadays with the extensive use of internet and social networks, researchers propose the use of data from social networks as a surveillance tool, this approach is called Infodemiology. This approach has been successfully tested in several studies.

The aim of this thesis is to build a database from self-reported tweets, capable of classifying a tweet as a potential case of Lyme or not using BERT transformer-based classifier models.
A total of 20,000 English tweets related to Lyme disease without geographical restriction from 2010 to 2022 were collected with twitter API. Then these data were cleaned and manually classified by binary classification as potential Lyme cases or not using as keywords the symptoms of Lyme disease; Also, emojis have been converted into words and integrated. Using classification models based on BERT transformers, the labeling of data as disease-related or non-disease-related is evaluated first without, and then with emojis.

Transformer-based classification models performed better than conventional classification models, especially the BERTweet model outperformed all evaluated models with an average F1 score of 89.3%, precision of 97%, accuracy of 90%, and recall of 82.6%. Also, the incorporation of emojis in our database improves the performance of all models by at least 5% but BERTweet once again performed best with an increase in all parameters evaluated. Tweets in English are mostly from the United States and to counteract this predominance, future work should collect tweets of all languages related to Lyme disease especially because the European countries where Lyme disease are emerging are not English-speaking countries.

Keywords: Lyme disease, social networks, Twitter, machine learning, Infodemiology, BERT, emojis, classification models.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

CDC: Centre de Contrôle et de prévention des maladies (*Center of Disease Control and Prevention*)

API: Interface de Programmation d'Applications (*Application Programming Interference*)

BERT : Représentations d'encodeurs bidirectionnels à partir de transformateurs (*Bidirectional Encoder Representations from Transformers*)

TF-IDF : Fréquence du terme-fréquence inverse du document (*Term frequency-inverse document Frequency*)

NLP : Traitement du Langage Naturel (*Natural Langage Processing*)

ML : Apprentissage automatique (*Machine Learning*)

RF : Forêt aléatoire (*Random Forest*)

NB : *Naive Bayes*

s.l : sensu lato

Ce mémoire n'aurait pas été possible si Tu ne m'avais pas fortifié, mon Créateur je te dis merci.

Remerciements

Je tiens à remercier sincèrement ma directrice de recherche Bouchra Nasri qui a su me guider dans la concrétisation de mon projet d'études avec beaucoup de patience et de conseils et m'a permis de découvrir les multiples facettes du monde de la recherche. Un grand merci à toute l'équipe de recherche surtout à Hamza qui m'a aidé dans une courte période à l'analyse des données et du côté méthodologique de l'analyse des données

À ma mère, merci de me soutenir dans mes projets de vie bien qu'incertains tu m'apportes toujours ton soutien infaillible.

À Sonia et Sorel, vous êtes plus que d'amis, merci pour votre aide

Enfin, mais pas du moindre, à mon créateur qui m'a soutenu et m'a donné la force nécessaire pour atteindre mon objectif

À vous parents et amis, je vous remercie de l'aide apportée de près comme de loin.

Chapitre 1 – [Introduction]

La maladie de Lyme a été décrite pour la première fois en 1977 par le scientifique Allen Steere après une éclosion de cas d'arthrite chez de jeunes patients dans une ville nommée Lyme, d'où le nom de la maladie, dans l'État de Connecticut aux États-Unis d'Amérique (Steere et al., 1977). La bactérie responsable de l'affection a ensuite été découverte un peu plus tard en 1982, ce qui a permis de mieux comprendre la pathogenèse de la maladie (Burgdorfer et al., 1982; Kahl et al., 1998).

La maladie a ensuite progressé au Canada où elle gagne de plus en plus de terrain (Ogden et al., 2015). Le nombre de cas est également en croissance dans plusieurs pays européens (Jánová, 2019; Mysterud et al., 2017). Cependant, il existe des preuves suggérant que la bactérie responsable de la maladie de Lyme a affecté les humains depuis des milliers d'années puisqu'elle a été identifiée dans des fossiles âgés de plusieurs milliers d'années (Cardenas-de la Garza et al., 2019). Plusieurs chercheurs affirment alors que la maladie a toujours existé mais, qu'elle est en ré-émergence et une des raisons énumérées serait la reforestation (Bonds et al., 2012; A. M. Kilpatrick et al., 2017).

La maladie de Lyme est devenue une maladie à déclaration obligatoire (MADO) en 1991 aux États-Unis, en 2003 dans la province du Québec et ensuite dans le reste du Canada en 2009 (Burgdorfer et al., 1989; Ripoche et al., 2018; Tutt-Guérette et al., 2021). La maladie de Lyme est également à déclarer dans certains pays européens, mais la déclaration obligatoire de la maladie n'est pas uniforme sur tout le continent (Blanchard et al., 2022a; Jánová, 2019). Cependant, l'Union Européenne a récemment encouragé les pays endémiques à rendre la déclaration de la maladie de Lyme obligatoire (Blanchard et al., 2022).

La signalisation des cas se fait principalement par les professionnels de santé et ces cas sont ensuite reportés dans les systèmes de surveillance développés à cet effet. Cette façon d'opérer présente des failles puisque seuls les cas qui recourent à l'attention médicale sont pris en compte. Aussi, les cas qui ne sont pas reportés par erreur de diagnostic sont absents du système de surveillance. Les chercheurs ont proposé plusieurs nouvelles approches pour résoudre ce problème, entre autres la surveillance digitale. Cette approche, qui sera référée dans cette étude comme infodémiologie, propose l'exploitation des données provenant de l'internet à des fins de surveillance (Eysenbach, 2009; Laranjo et al., 2015; Paul & Dredze, 2011). Les données provenant du web, incluant celles des réseaux sociaux, ont prouvé durant les dernières années être utiles pour les interventions et la surveillance en santé publique (Carneiro & Mylonakis, 2009a; Culotta, 2010; Klein et al., 2021; Stevens et al., 2020; Yaya & Ghose, 2018). Par exemple une exploration des discussions sur Twitter et l'occurrence réelle de grippe et de la Coqueluche a démontré une forte association entre les tweets et l'apparition des maladies (Nagel et al., 2013). Aussi selon Marques et ses collaborateurs, les tweets sont capables d'estimer avec succès les cas de dengue en temps réel et jusqu'à 8 semaines dans le futur et donc constituerait un complément utile à la surveillance traditionnelle de la dengue (Marques-Toledo et al., 2017).

Selon une étude systématique, malgré l'avancée de l'exploitation des données du web à des fins de surveillance, il existe un manque quant à la régularisation des méthodes adoptées pour leur analyse (Chen & Wang, 2021). Dans ce mémoire, nous proposons de développer un algorithme avec une base de données en langue anglaise provenant de Twitter en lien avec la maladie de Lyme et d'évaluer leur étiquetage comme cas potentiels ou non de la maladie en utilisant des modèles de classification robustes basés sur des transformateurs BERT à savoir ALBERTA, DistilBERT et BERTweet. Le *chapitre 2* de ce mémoire sera consacré à une revue de littérature sur la maladie de

Lyme à travers le monde, et les facteurs influençant la hausse des cas. Ensuite, dans le *chapitre 3,* nous présenterons un bref aperçu de l'exploitation des données provenant du Web (moteurs de recherche sur internet) particulièrement celles provenant des réseaux sociaux dans le cadre de la surveillance des maladies et leurs particularités. Par la suite, dans le *chapitre 4,* nous expliquerons la méthodologie proposée en élaborant le cadre analytique de cette étude. Au chapitre 5, nous inclurons un manuscrit soumis et actuellement en cours de révision dans une revue scientifique évaluée par des pairs, le *Journal of Medical Internet Research* (JMIR). Cette étude est une extension d'une étude réalisée par notre équipe de recherche où je figure comme co-première auteure et qui est en révision à *BMC Medical Informatics and Decision Making,* intitulée *: Leveraging Machine Learning Approaches for Predicting Potential Lyme Disease Cases and Incidence Rates in United States Using Twitter* par *Srikanth Boligarla, Elda Kokoè Elolo Laison, Jiaxin Fiona Li, Raja Mahadevan, Austen Ng, Yangming Lin, Mamadou Yamar Thioub, Bruce Huang, Mohamed Hamza Ibrahim, Bouchra Nasri.* Nous expliquerons brièvement les résultats de notre étude dans le *chapitre 6* et soulignerons les apports de notre étude à la santé publique de même que les limites associées à la méthodologie. Ce projet de recherche est partiellement financé par la *bourse en Intelligence Artificielle offerte par les ESP (Études Supérieures et Postdoctorales)* au niveau de maîtrise de laquelle j'ai été bénéficiaire durant la session d'hiver 2022.

Chapitre 2 – [Recension des écrits sur la maladie de Lyme]

Les zoonoses constituent un problème de santé grandissant à l'échelle mondiale, représentant plus de 60% des maladies infectieuses et sont influencées par plusieurs facteurs à savoir le réchauffement climatique et la reforestation (Bonds et al., 2012; Jones et al., 2008). Les tiques occupent la première place des vecteurs impliqués dans la transmission des maladies vectorielles aux États-Unis et la deuxième place au niveau mondial (E. Choi et al., 2016). En raison de la complexité du cycle de vie des tiques, les mesures de prévention proposées par les autorités de santé publique contre ces maladies sont difficiles à mettre en œuvre efficacement et donc, leur incidence continue d'augmenter (Clark & Hu, 2008). La densité des tiques est très corrélée avec la densité forestière et la température, ce qui rend l'approche une seule santé, à l'interface homme-animal-environnement, bénéfique dans le cadre de la maladie de Lyme (Curriero et al., 2021).

2.1 La maladie de Lyme comme problème de santé publique et description de son épidémiologie

La maladie de Lyme est reportée comme étant la maladie transmise par tiques la plus répandue dans l'hémisphère du nord et émergente dans certains pays de l'hémisphère du sud (Hussain et al., 2021). La maladie est probablement présente, mais sous-rapportée en Amérique centrale. Elle est également reportée au Chile et en Uruguay (Chomel, 2015).

Aux États-Unis, si le CDC (Control and Prevention of Disease Center) reporte 30 000 cas en moyenne par année, plusieurs études affirment que ce nombre est sous-reporté, ce qui reflète les déficits du système de surveillance actuel (Bobe et al., 2021; Schwartz et al., 2021). La majorité des cas américains de la maladie de Lyme se produit dans le Nord-Est et dans l'Ouest du pays alors

que dans les régions pacifiques et du sud, ils sont peu fréquents (Bacon et al., 2008). En 2018, les États-Unis ont dépensé environ 9.6 millions de dollars américains ($) sur le diagnostic et le traitement de cette zoonose (Geebelen et al., 2019; Mac et al., 2019, 2021; Nam et al., 2022). Dans une étude sur le coût de la maladie publiée en 2006, Zhang et ses collègues ont constaté que le coût annuel de diagnostic était de 1 310 $ par patient précoce, mais de 16 199 $ pour les cas de maladie tardive, soit environ 12 fois de plus, d'où l'importance du diagnostic précoce de la maladie (X. Zhang et al., 2006).

L'incidence de la maladie au Canada est moindre comparée à son voisin, les États-Unis. Si l'incidence de la maladie de Lyme était relativement faible au début du siècle, le nombre de cas de la maladie de Lyme rapporté par toutes les provinces a toutefois augmenté de 144 en 2009 à 992 en 2016, ce qui représente une augmentation de 0.4 à 2.7 pour 100 000 habitants durant cette période (Gasmi et al., 2017; Ogden et al., 2015). En 2019, les données préliminaires montraient plus de 2 500 cas dans le pays avec près de 9 des 10 cas dans les provinces d'Ontario, au Québec et la Nouvelle-Écosse (Canada, 2022). En Europe occidentale, le nombre de cas est estimé à environ 200 000 en moyenne par année (Marques et al., 2021; Stanek & Strle, 2018). Le taux d'incidence est également en hausse et on dénote d'importantes variations géographiques avec un gradient décroissant de l'Est vers l'Ouest européen; et les plus fortes incidences sont observées dans la partie centrale du continent européen (Jánová, 2019(Mysterud et al., 2017; Rizzoli et al., 2014); Piesman & Gern, 2004a; Stanek & Strle, 2008; Steere, 2001a). La maladie de Lyme est endémique dans les pays européens suivants : Autriche, Slovénie, Allemagne, Danemark, Suisse, Suède, Finlande, Estonie, Lituanie, République Tchèque et la France (Bregnard et al., 2020; Stanek & Strle, 2008).

Bien qu'initialement décrite dans l'hémisphère nord, la maladie de Lyme gagne de plus en plus du terrain. La borréliose encore appelée la maladie de Lyme a également été décrite dans d'autres parties du monde, telles que l'Asie, principalement les zones rurales, et peut-être l'Océanie et l'Amérique du Sud (Bregnard et al., 2020; Brown, 2018; Campbell et al., 1998; Gordillo-Pérez et al., 2003; Ogden et al., 2015). En Asie, une étude réalisée en Inde du Nord a révélé la présence de cas sporadiques dans la région. Toutefois, la prévalence de la maladie demeure incertaine (Vinayaraj et al., 2021). Elle a aussi fait l'objet d'études en Chine, particulièrement dans la province de Hainan, au Japon et en Corée (W. C. Lee et al., 2019; Wen et al., 2021). En Afrique, il n'existe pas de données sur les cas de Lyme. Cependant, selon une étude en Égypte, les tiques de type *Ixodes* y sont présentes sur les chiens sans preuve d'infection humaine (Alkishe et al., 2021; Hussain et al., 2021).

Aux États-Unis, la majorité des cas (environ 90%) sont regroupés dans les États du Nord-Est et du nord du Midwest (Clayton et al., 2015). L'incidence de la maladie par rapport à l'âge est bimodale dans les zones endémiques, plus élevée chez les enfants de 14 ans ou moins et des adultes de plus de 50 ans (Chomel, 2015; Diuk-Wasser et al., 2021; Dumes, 2020). La majorité des cas sont des personnes de sexe biologique masculin en Amérique du Nord et en Asie tandis qu'en Europe, la prédominance est chez les personnes de sexe biologique féminin (Bisanzio, Fernández, et al., 2020; Chomel, 2015). Les cas de Lyme en Europe ont une répartition disproportionnée; la maladie est endémique dans le centre du continent tandis que son incidence est faible dans les régions du Nord et du Sud (Piesman & Gern, 2004). Les cas humains rapportés sont plus fréquents au cours de l'été et de l'automne (Aenishaenslin et al., 2017; H. J. Kilpatrick et al., 2014; Kugeler et al., 2015).

2.2 Facteurs liés au risque d'infection

Le risque de contracter la maladie de Lyme dans la population humaine est influencé d'une part par l'abondance et la distribution des tiques et d'autre part par les activités extérieures qui mettent les humains en contact avec les tiques infectées (Gilbert, 2021; Gray et al., 1994; Kahl et al., 1998; Lindsay et al., 1999). Plusieurs facteurs contribuent à l'établissement et à l'expansion de la population de tiques dans une région donnée à savoir les facteurs climatiques, écologiques et environnementaux, le tout influencé par le changement climatique (Brownstein et al., 2005; H. J. Kilpatrick et al., 2014; Ripoche et al., 2018; Simon et al., 2014; Stone et al., 2017).

Les facteurs écologiques qui peuvent faciliter ou empêcher l'étendue des cas humains de la maladie de Lyme restent jusqu'à ce jour peu compris (Brownstein et al., 2005; Eisen et al., 2012). Il semblerait que la litière composée de feuilles mortes, constitue un facteur protecteur de la survie des tiques (Brownstein et al., 2005; Feria-Arroyo et al., 2014). Les tiques hivernent généralement dans la litière à la surface du sol, qui fournit une excellente couche d'isolation, même lors de températures hivernales extrêmes (Eisen et al., 2016). Il est donc accepté qu'un hiver plus doux diminuerait la mortalité des tiques et donc favoriserait sa dispersion (S. Lin et al., 2019). Aussi un niveau d'humidité élevé aiderait à la survie des tiques, parce que cette dernière dépend de la capacité des tiques à ne pas dessécher en milieu adverse (Dehnert et al., 2012; Gabriele-Rivet et al., 2015). La fragmentation environnementale influence principalement les caractéristiques écologiques de certains réservoirs des tiques, qui ont tendance à se regrouper dans des zones de transmission et donc se propage rapidement dans les zones fragmentées (Brownstein et al., 2005; Kilpatrick et al., 2017; Simon et al., 2014). Le risque d'infection humain dans ces zones fragmentées est plus élevé puisque ces zones sont fréquemment utilisées pour des activités humaines récréatives ou professionnelles (Vourc'h et al., 2016).

La température semble être un des facteurs climatiques primordiaux qui influencent la survie des tiques. En effet la mortalité des tiques diminue drastiquement lorsque les températures demeurent au-dessus de – 10ºC (Allehebi et al., 2022; Eisen et al., 2016; Ogden et al., 2008). Un des exemples, serait le cas de la province du Québec où des températures élevées sont de plus en plus enregistrées dans le sud de la province, ce qui expliquerait la hausse des cas de cette région (Bouchard et al., 2015; Gasmi et al., 2016).

Habiter en milieu rural ou à proximité d'un boisé, ou encore avoir observé des cerfs sur la propriété (Bayles et al., 2013; Magnavita et al., 2022) ont également été identifiés comme des facteurs associés. Dans le même ordre d'idée, les travailleurs extérieurs sont considérés comme un groupe à risque pour la maladie de Lyme. Plusieurs études se sont intéressées au risque professionnel des travailleurs extérieurs et forestiers et ont montré que le risque d'avoir une sérologie positive était plus élevé chez ces travailleurs (Magnavita et al., 2022; Slatculescu et al., 2022; Smith et al., 1988). En résumé en plus des facteurs climatiques et environnementaux, d'autres aspects tels que la biodiversité, le microclimat influencent également le cycle de transmission de la bactérie causante de la maladie de Lyme *B.burgdorferi* (Bouchard et al., 2013; Ogden et al., 2009; Werden et al., 2014).

2.3 Cycle de vie des vecteurs de la maladie de Lyme

Les vecteurs de la maladie de Lyme sont tous des ectoparasites hématophages obligatoires ; c'est-à-dire qu'ils sont des parasites qui vivent sur la surface corporelle d'un être vivant, appelé un hôte, dont ils ont besoin pour leur survie. Elles peuvent toutefois causer des maladies à ces hôtes, comme dans ce cas, la maladie de Lyme (Burgdorfer et al., 1989; Stewart & Rosa, 2018). Les espèces de tiques ont différentes saisons d'activité ; les types *Ixodes ricinus* et *Ixodes persulcatus*

sont actives en début de printemps et mi-été et le type *I. Scapularis* du début d'été au début d'automne (Gilbert, 2021).

Le cycle biologique actif des tiques passe par 3 stades de développement: larvaire, nymphal et le stade d'adulte (Stewart & Rosa, 2018). Étant des hématophages, un repas de sang est nécessaire pour se muer et passer d'un stade au suivant et pour permettre aux femelles adultes de pondre leurs œufs avec succès (Stanek et al., 2012). Plusieurs mois sont nécessaires pour que les tiques passent d'un stade à un autre (Clark & Hu, 2008; Ostfeld & Brunner, 2015). Le cycle biologique actif des tiques dure en moyenne deux ans; toutefois il peut varier jusqu'à quatre ans (Piesman & Gern, 2004b; Stanek et al., 2012). La phase larvaire commence par l'éclosion des œufs vers la fin du printemps (mai à juin) et se termine par leur passage en larve active en quête d'hôtes pendant la période de fin juillet à septembre (Eisen & Eisen, 2018). Une fois rencontrés, les larves se nourrissent sur leur hôte. Les nymphes sont plus abondantes dans la saison hivernale et se retrouvent sous les litières sur le sol. Au printemps suivant (mai à juillet), si les conditions climatiques et environnementales le permettent, les nymphes commencent à chercher des hôtes sur lesquelles ils se nourrissent pendant une période de 3-4 jours (De Silva & Fikrig, 1995). Les nymphes, une fois engorgées de sang, tombent et commencent leur processus de maturation qui se culmine vers la fin de l'automne où elles se muent en adultes (Kurokawa et al., 2020). Les adultes ensuite s'accouplent et le cycle recommence. Les cas humains sont davantage reportés de mai à septembre, coïncidant avec l'activité nymphale et les activités extérieures (Eisen et al., 2012). Les réservoirs de ces vecteurs sont généralement de petits mammifères, principalement les souris à pattes blanches (*Leucopus peromyscus*) (Beckmann et al., 2019; Bouchard et al., 2013; Ratti et al., 2021). Les oiseaux migrateurs sont plus que des réservoirs. Ils jouent un rôle de distributeur et sont importants dans le maintien de la population des tiques dans de nouvelles régions (Dumas et al.,

2022; Werden et al., 2014). L'être humain qui n'intervient pas dans le cycle de vie naturel de cette maladie est ainsi considéré comme un hôte accidentel (Bouchard et al., 2011; Ogden et al., 2008). Le choix des hôtes varie en fonction du stade de vie des tiques (Bregnard et al., 2020; Simon et al., 2014).

Les vecteurs responsables de transmettre la bactérie de la maladie de Lyme aux humains sont les tiques de type *Ixodes* (Steere, 2001). Le type de tiques diffère dépendamment du lieu géographique. Les vecteurs de ses 3 espèces pathogènes à l'humain sont (1) *Ixodes scapularis,* encore appelés les tiques de cerfs ou les tiques à pattes noires en Amérique du Nord (Nord-Est et Nord Central des États-Unis ; dans la partie centrale et Est du Canada) ; (2) *Ixodes Pacificus* (Ouest des Etats-Unis et en Colombie Britannique) *; (3) Ixodes Ricinus* encore nommé la tique à chèvre européenne en Europe et en Asie. On retrouve les tiques de l'espèce *Ixodes persulcatus* (taïga tick) majoritairement en Asie (Belongia, 2002; Gasmi et al., 2016, 2017; Gern et al., 1997; Ji et al., 2022; Lindsay et al., 1999). La figure 1 est une illustration de la distribution spatiale des différentes espèces de tiques vecteurs de la bactérie causante de la maladie de Lyme dans le monde.

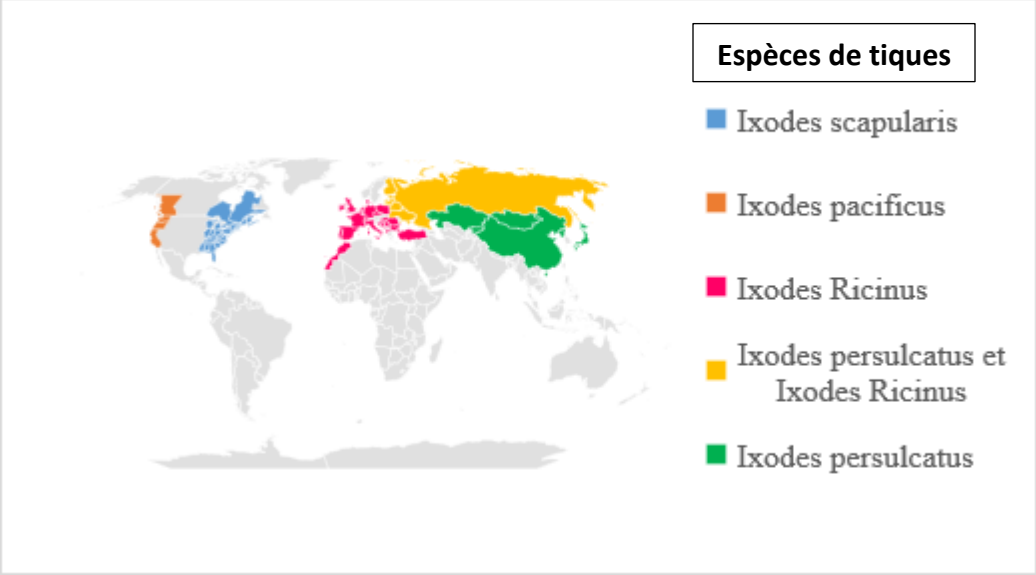

Figure 1 Distribution spatiale des différentes espèces de tiques vecteurs de la maladie de Lyme dans le monde.

2.4 Étiologie de la maladie de Lyme

La bactérie qui est responsable de causer la maladie de Lyme chez l'humain fait partie du genre *Borrelia*, appartient à la famille des Spirochaetaceae, à l'ordre des Spirochètes (Burgdorfer et al., 1989).

Ce complexe bactérien *Borrelia burgdorferi sensu lato* (s.l.), est constitué d'une diversité d'espèces, dont au moins 34 espèces de spirochètes desquelles 9 sont potentiellement capables d'infecter les humains (*B. afzelii, B. bavariensi, B. bissetti, B. burgdorferi sensu stricto* communément appelé *B. burgdorferi, B. garinii, B. burtenbachii, B. lusitaniae, B. spielmanii, B. valaisianna*) (Burgdorfer et al., 1982; Piesman & Gern, 2004; Stone et al., 2017). Néanmoins, d'un point de vue clinique, trois d'entre elles sont les plus significatives quant à leur potentiel d'infection chez l'être humain et elles sont : *B. afzelii, B. garinii, B. burgdorferi* (Cook, 2015; Lou & Wu, 2017). Leur répartition géographique est diversifiée. En Europe et en Asie, le *B. afzelii et B. garinii* sont actifs, tandis qu'en Amérique du Nord, seul l'espèce de *B. burgdorferi* est présente (Bockenstedt & Wormser, 2014; Ogden et al., 2015).

Il existe une différence de transmission de la bactérie causante. Il a été constaté que *I. pacificus* avait un succès d'attachement à l'hôte et des poids d'engorgement significativement plus élevés, mais une efficacité de transmission des agents pathogènes significativement inférieure par rapport à *I. scapularis* (Cook, 2015). Ceci pourrait expliquer la faible incidence de la maladie de Lyme dans les régions où cette espèce de tiques réside, notamment dans l'Ouest des États-Unis et en Colombie-Britannique au Canada (Couper et al., 2021). Le tableau 1 illustre la répartition géographique des espèces de la bactérie causante de la maladie de Lyme et celles des tiques qui les transportent.

Espèce bactérienne	Types de tiques	Régions géographiques
Borrelia burgdorferi	*Ixodes scapularis*	Nord-Est et Nord central aux États-Unis; Canada (Sud et Est)
	Ixodes pacificus	Ouest des États-Unis et en Colombie Britannique (Canada)
	Ixodes ricinus	Europe
Borrelia garinii	*Ixodes ricinus*	Europe
	Ixodes persulcatus	Asie
Borrelia afzelii	*Ixodes ricinus*	Europe
	Ixodes persulcatus	Asie

Tableau 1. – Répartition géographique des tiques et des espèces de bactérie à travers les zones endémiques à la maladie de Lyme. Adapté de (Radolf et al., 2021)

2.5 Manifestations cliniques

La borréliose de Lyme a une vaste symptomatologie qui dépend principalement de la souche bactérienne présente dans la région (Hanincova et al., 2013; Kurtenbach et al., 2006; Ogden et al., 2015). Les symptômes de la borréliose de Lyme sont classiquement organisés en trois (3) phases à savoir *la phase localisée précoce*, la *phase disséminée précoce* et la *phase disséminée tardive* (Bockenstedt & Wormser, 2014).

La phase localisée précoce est marquée par une symptomatologie indolente avec une infection localisée accompagnée de symptômes généraux tels que la fièvre, les maux de tête, la fatigue (Cardenas-de la Garza et al., 2019). Elle est caractérisée par un érythème migrant, qui est une lésion maculopapuleuse érythémateuse de plus de 5 cm de diamètre avec ou sans décoloration centrale qui s'étend de manière annulaire et centrifuge (Wormser et al., 2006). La fréquence de l'érythème migratoire varie en fonction des études. Si elle est le signe caractéristique de cette phase, elle n'est toujours pas présente ou peut se présenter de façon atypique (Steere et al., 2003). L'érythème migratoire se développe le plus souvent avec les piqûres par nymphes (Rebman et al., 2021).

La phase disséminée précoce peut être concomitante ou succéder à la première phase et elle correspond à la dissémination sanguine de la bactérie chez l'hôte infecté du site de piqûre vers les organes distants. Elle est caractérisée par l'apparition de deux ou plus érythème migratoire, d'un lymphocytome cutané avec des symptômes neurologiques (neuroborréliose), cardiaques (cardite), articulaires (arthrite de Lyme) ou oculaires (Cardenas-de la Garza et al., 2019). À cette étape, les manifestations cliniques prédominantes dépendent de la région géographique et de l'agent étiologique de l'infection. En Europe et en Asie où le *B. garinii* et le *B. afzelii* sont présents, les manifestations neurologiques sont les plus fréquentes suivies des manifestations cutanées, tandis qu'en Amérique du Nord où le *B. burgdorferi* est présente, l'arthrite de Lyme est la forme la plus commune (Carriveau et al., 2019; Wormser et al., 2006).

La phase disséminée tardive est due à une infection par *B. burgdorferi* de multiples organes qui peut persister durant une période courte ou à long terme (Wormser et al., 2006). Les manifestations articulaires et cutanées sont présentes durant cette phase, mais la particularité de ces manifestations cutanées est qu'elle ne disparaît pas spontanément (Stanek et al., 2012).

Une forme chronique de PTLDS (*Post-Treatment Lyme Disease Syndrome*) est toujours en discussion dans la communauté scientifique. Nombreux sont les patients qui rapportent des symptômes persistants sur une durée de six mois ou plus comme la fatigue et la dépression entre autres et ce, malgré un traitement précoce et effectif contre la maladie de Lyme (Batheja et al., 2013). Une des raisons de cette incertitude est l'inaptitude des tests sérologiques disponibles de détecter une infection active ainsi que la pathogénie de la maladie qui demeure inconnue (Aucott, 2015).

Malgré les traitements antibiotiques adéquats, 10% des cas finissent par avoir des complications et développer une forme tardive (Stark et al., 2022).

2.6 Diagnostic

La symptomatologie clinique du patient avec la présence de l'érythème migrante avec une historique de piqûre de tique est suffisante pour faire un diagnostic de la maladie de Lyme dans une zone endémique (Lantos et al., 2021; Stanek & Strle, 2018). Selon Eisen et collaborateurs, 30-50% des cas de la maladie de Lyme ne se rappellent pas avoir été mordu par une tique, ce qui complique le diagnostic (Eisen & Eisen, 2018). Dans la présence du doute, il existe des tests sérologiques pour confirmer les cas de Lyme. Les tests sérologiques qui sont utilisés pour le diagnostic correspondent à une approche à deux niveaux, en utilisant test ELISA (Enzyme-Linked Immunosorbent Assay) suivie d'un Western blot ou un immunoblot de IgG et IgM si les résultats d'ELISA sont positifs ou erronés (Aguero-Rosenfeld et al., 2005; Borchers et al., 2015; Stanek & Strle, 2018). Les tests sérologiques sont insensibles dans les deux premières semaines d'où l'importance de reconnaître l'érythème migrant et de prendre en considération l'historique de piqûres et des activités réalisées (Beck et al., 2021). Les résultats de ces deux tests et leur interprétation vont dépendre en partie du stade de la maladie de Lyme ; au début de l'infection, la

sensibilité des tests d'anticorps est faible (30-40%), et passe à 70-95% dans le stade de dissémination de l'infection et enfin devient excellente au stade tardif de la maladie de Lyme (Borchers et al., 2015; Maksimyan et al., 2021). En Europe, la diversité bactérienne dans l'incidence de la maladie de Lyme rend l'interprétation des tests sérologiques difficile à normaliser contrairement en Amérique du Nord (Marques et al., 2021).

2.7 Traitement

La maladie de Lyme est traitable avec un traitement antibiotique de courte durée dans la phase initiale de la maladie (C. T. Nguyen et al., 2022; Sanchez, 2015).

En 1990, un vaccin a été développé aux États-Unis, avec sa sécurité et son efficacité prouvées, mais a été retiré en 2002. Ce vaccin montrait une bonne efficacité (78% après trois doses), mais a été retiré du marché après qu'une forte attention médiatique ait affecté négativement la réputation du vaccin en exposant des doutes sur la présence d'effets indésirables importants (Nigrovic & Thompson, 2007). D'autres causes énumérées qui auraient également contribué au retrait du vaccin sont : la réticence des individus à se faire vacciner étant donné la nécessité de plusieurs rappels pour obtenir une bonne efficacité, un coût élevé, l'exclusion des enfants et la présence de certains effets indésirables dont des douleurs musculosquelettiques (Hanson & Edelman, 2003; Hayes & Piesman, 2003; Nigrovic & Thompson, 2007). À date aucun vaccin contre la maladie de Lyme n'est disponible contre les cas humains.

Il existe une prophylaxie orale de 100-200 mg de doxycycline à la suite d'une morsure de tique chez l'adulte; toutefois, la prophylaxie n'est conseillée que dans les zones endémiques de tique dans les 72 heures qui suivent la piqûre de tique (Wormser et al., 2006; Zhou et al., 2021).

De multiples stratégies sont proposées pour la protection de la population contre les piqûres de tiques comme la protection personnelle, les stratégies domestiques (modification du paysage et lutte chimique contre les ravageurs), la réduction des cerfs et les vaccins contre la maladie de Lyme, y compris les vaccins ciblant l'immunité contre les tiques. Néanmoins, leurs efficacités ont été évaluées sans consensus concernant la mesure la plus optimale (Aenishaenslin et al., 2017; Atkinson et al., 2014; Eisen et al., 2012).

Chapitre 3 – [L’exploitation des données de l’activité sur Internet comme outil de surveillance épidémiologique]

3.1 Historique des systèmes de surveillance de santé

La surveillance en santé publique se définit comme un système qui permet la collecte, l’analyse et la diffusion continue de données dont le but est de développer des politiques et des programmes en santé publique (Cartter et al., 2018). La surveillance revêt une importance majeure en recherche en santé publique, puisqu'elle nous permet de définir les tendances démographiques et spatio-temporelles d'une maladie donnée (Bacon et al., 2008). La surveillance peut être définit en deux catégories : passive et active. La surveillance active est un processus proactif dans la collecte des données ou des cas tandis que la surveillance passive, initie la notification des cas par les professionnels de santé une fois que ces cas viennent vers eux (Canada, 2022; B. C. K. Choi, 2012). Les systèmes de surveillance traditionnels dans le cadre de la maladie de Lyme se basent sur la surveillance active et passive des tiques, mais également sur la surveillance passive des cas par les professionnels de santé entre autres. La surveillance de la maladie de Lyme tant en Amérique du Nord qu’en Europe, a beaucoup de failles et ne parvient pas à capter les cas en temps réel ni à estimer leur nombre exact (Blanchard et al., 2022). Aussi, ces systèmes utilisent des approches qui sont gourmandes en ressources (Nelson et al., 2015). Donc, les méthodes de surveillance des cas de Lyme bien qu'utiles, sont peu efficaces pour déterminer le fardeau réel de la maladie (Castillo-Salgado, 2010).

Traditionnellement, la surveillance des cas de cette zoonose s’appuie sur des cas confirmés en laboratoire, tels que signalés par un médecin clinicien, un laboratoire ou un service d’urgence

(CDC, 2021). Ces données de surveillance sont assujetties à plusieurs limites, y compris la modification des définitions des cas de surveillance selon les régions, la disponibilité de ressources de santé publique pour la surveillance, les variations dans les pratiques de surveillance et l'ambiguïté entre le lieu de résidence du cas et le lieu exact d'exposition (Curriero et al., 2021; Kugeler et al., 2015b; Mac et al., 2021). Cette limite des systèmes de surveillance affaiblit leur potentiel à saisir le fardeau réel de la maladie de Lyme. Selon une récente revue systématique, tous les systèmes de surveillance traditionnels de la maladie de Lyme sont sujets soit à une sur ou sous déclaration des cas (Blanchard et al., 2022). La surestimation est due d'une part à un mauvais diagnostic des cas de Lyme et d'autre part à l'attribution des cas de Lyme dans des régions où ces cas n'ont pas été contracté (Blanchard et al., 2022b). Aussi selon une étude menée par Forrester et collaborateurs, les cas de Lyme dans les États américains à faible incidence à la maladie de Lyme sans un historique de voyage seraient des cas probables de mauvais diagnostic et donc entraînent une surestimation des cas de Lyme dans ces régions (Forrester et al., 2015). La sous-estimation des cas est aussi documentée dans les écrits. Par exemple, aux États-Unis, la CDC estime les cas de Lyme à environ 30 000 cas par année, mais plusieurs études rapportent un nombre de cas bien supérieur (Geebelen et al., 2019; Kugeler et al., 2022; Nelson et al., 2015; Rosenberg et al., 2018; Schwartz et al., 2021; Tseng et al., 2015). En identifiant des diagnostics de la maladie de Lyme sur la base des codes de facturation spécifiques des prescriptions médicales des cas de Lyme, cette même étude a su démontré que les réclamations des assurances de santé peuvent servir d'outil de surveillance pour identifier les tendances de diagnostic des cas humains de la maladie de Lyme (Schwartz et al., 2021). Cette étude a d'ailleurs révélé des diagnostics en dehors de la saison de la maladie de Lyme. De plus, la majorité des diagnostics erronés ont lieu le plus souvent dans les états à faible incidence ou voisins des états avec une forte endémicité à la maladie de Lyme (Schwartz et al., 2021). L'hétérogénéité dans l'estimation des cas reportés dans le système de surveillance en

Europe et en Amérique du Nord serait dû au manque de consensus dans les méthodes de collecte et des critères diagnostic d'un cas de Lyme d'un milieu à un autre (Blanchard et al., 2022b; Mac et al., 2021).

3.2 Infoveillance, qu'est-ce que c'est ?

L'ère de l'internet améliore non seulement la communication, mais offre de nouveaux horizons à la recherche en santé publique comme l'exploitation des données provenant du web pour des fins de surveillance des maladies (Castillo-Salgado, 2010). Des termes comme « infodémiologie » et « infoveillance » ont commencé à apparaître dans la littérature au début de ce siècle (Eysenbach, 2002, 2009), décrivant initialement une forme de surveillance syndromique qui se réfère à la collecte et à l'analyse systématiques des tendances dans les recherches sur Internet pour prédire les épidémies. L'infoveillance n'est pas une nouvelle méthode pouvant remplacer les méthodes de surveillance classiques mises en place, mais un outil complémentaire permettant l'utilisation de données en temps réel (Barros et al., 2020; McGough et al., 2017). Une surveillance épidémiologique basée sur l'activité en ligne attire de plus en plus l'attention des chercheurs dans le domaine de la santé publique et plusieurs études ont été menées pour tester la performance des données provenant de l'activité en ligne comme outil de surveillance épidémiologique (Chew & Eysenbach, 2010; Chorianopoulos & Talvis, 2016; Culotta, 2010; Mahroum et al., 2018). Les recherches sur Google Trends, l'outil d'exploitation des données développé par Google ont permis de surveiller des éclosions avec plus d'exactitude que les systèmes de surveillance déjà existants (Carneiro & Mylonakis, 2009; Santillana et al., 2015). Santillana *et al.* (2015) dans son étude qui s'est basée sur les données provenant des moteurs de recherche de Google et de Twitter pour prédire le cas d'Influenza, ont su démontré la capacité de ces données à prédire jusqu'à 4 semaines à l'avance, les cas de grippe avec plus de précision que le CDC (Santillana et al., 2015).

3.2.1 Utilité des réseaux sociaux comme source de données pour la surveillance des maladies et leurs particularités :

Les sites de réseautage sociaux communément appelés les *réseaux sociaux* sont devenus dans les deux dernières décennies un phénomène mondial. Selon une estimation, le nombre total d'utilisateurs des différents réseaux sociaux inclusivement devrait atteindre plus de 4,4 milliards de personnes d'ici 2025, ce qui représente plus de la moitié de la population mondiale actuelle (Chiang, 2021). Plusieurs réseaux sociaux sont très utilisés tels que Facebook, Reddit, Instagram, Twitter et autres. Les réseaux sociaux peuvent être définis comme des plateformes Web qui confèrent à leurs utilisateurs la possibilité de créer leur profil pour pouvoir créer une connexion avec d'autres utilisateurs et discuter de plusieurs sujets d'intérêt collectif dont les problèmes en lien avec la santé (Adebisi et al., 2021; Laranjo et al., 2015). Leur principal avantage pour la surveillance des maladies est la rapidité à l'accès de l'information, grâce à la disponibilité de données géolocalisées en temps réel (Eysenbach, 2009). En effet, étant donné les retards inhérents à la collecte et à la gestion de données associés aux nombreux systèmes de surveillance traditionnels des maladies, les données provenant des réseaux sociaux offrent donc une opportunité aux chercheurs de réduire le temps de collecte de ces données (Stoové & Pedrana, 2014). Dans le cadre de ce projet, Twitter a été l'objet de notre étude.

3.2.2 Avantages de twitter sur les autres réseaux sociaux

L'usage de Twitter, un réseau social de microblogage, est très répandu dans le public. En 2020, Twitter comptait à lui seul plus de 500 millions de tweets journaliers. De plus, il a été observé que plus de la moitié des utilisateurs de Twitter utiliseraient également les autres réseaux sociaux tels que Facebook, Instagram et YouTube (Chen & Wang, 2021). Twitter se classe parmi les 5 réseaux sociaux les plus connus parmi la population mondiale (Byrd et al., 2016).

Twitter aurait le potentiel d'être d'une richesse remarquable en données pour la recherche et donc être utilisé comme outil de surveillance des maladies (Arias et al., 2014; Mayor & Bietti, 2021). En 2012, Twitter a développé l'API Twitter (Application Programming Interference), une interface de programmation d'applications de *streaming*, qui permet l'accès gratuit à 1% des tweets aux chercheurs, choisit aléatoirement, facilitant la récupération et l'exploitation des données de ce site (Bour et al., 2021; Sinnenberg et al., 2017). Ces 1% de données accessibles aux chercheurs sont choisis aléatoirement durant une période spécifique (Bour et al., 2021). Un des avantages de Twitter est que les textes publiés (tweets) sont courts (280 caractères au maximum) ce qui rend la récupération et le traitement de ces données plus gérables comparativement aux textes sur Reddit, Facebook, etc. (Jahanbin et al., 2019).

Un autre des avantages de Twitter est la possibilité de détecter l'emplacement (géolocalisation) et la date des messages publiés, information utile pour nous informer sur la distribution spatio-temporelle de l'information recherchée (Tulloch et al., 2019). Plusieurs études l'ont d'ailleurs exploré comme outil de surveillance durant la pandémie de la Covid-19 (Klein et al., 2021; Liew & Lee, 2021; Sarker et al., 2020). Sarker et collaborateurs ont su observé une tendance commune entre les tweets reportant des cas auto-déclarés de COVID et la progression de la pandémie suggérant ainsi une relation positive entre les cas auto-déclarés sur les réseaux sociaux et l'incidence de la Covid-19 (Sarker et al., 2020). Des réseaux sociaux dont les données sont accessibles, la plateforme la plus utilisé dans les recherches scientifiques est Twitter (Bauer & Lizotte, 2021; Benítez-Andrades et al., 2022).

3.3 L'infoveillance dans la surveillance de la maladie de Lyme

Google Trends, Twitter et d'autres données de plateformes de médias sociaux comme YouTube; mais aussi des discussions des différents forums médicaux offrent un outil intéressant

pour surveiller la maladie de Lyme (Basch et al., 2017; Kapitány-Fövény et al., 2019; Kim et al., 2020; Kutera et al., 2021; Pesälä et al., 2017; Sadilek et al., 2020; Scheerer et al., 2020; Seifter et al., 2010; Tulloch et al., 2019; Yiannakoulias et al., 2017). Les résultats de ces études démontrent comment l'analyse en temps réel du contenu des médias sociaux peut compléter d'autres données de surveillance dans un contexte d'éclosion afin de fournir des estimations convenables du fardeau de la maladie de Lyme. Il a été décrit précédemment que les recherches sur le moteur Google avec des termes associés à la maladie de Lyme montrent des modèles similaires dans les variations temporelles et spatiales conformes à la tendance observée dans les données épidémiologiques (Pesälä et al., 2017; Seifter et al., 2010). Dans le cadre de la maladie de Lyme, Twitter n'a été implémenté que dans une étude dans le but de déterminer la correspondance entre l'épidémiologie spatio-temporelle de la maladie de Lyme et les habitudes de tweets sur la maladie en Irlande et aux Royaume-Uni (Tulloch et al., 2019). Il a été démontré que les tweets sur la maladie de Lyme correspondaient à l'épidémiologie de la maladie dans ces pays, mais aussi que les données de Twitter permettaient de localiser les points chauds de la maladie en géolocalisant la provenance des tweets (Tulloch et al., 2019).

Une récente étude réalisée aux États-Unis, en utilisant les données de l'activité de la population en ligne sur plusieurs moteurs de recherche tels que *Google search,* les forums médicaux et autres, a développé un système de surveillance appelé *Lymelight* qui est capable de prédire les cas de Lyme avec 92% de similitude avec les données officielles du CDC. Ce système peut estimer l'incidence réelle de la maladie de Lyme beaucoup plus tôt et plus efficacement que le système officiel de suivi de la maladie de Lyme, mais est également prédire la propagation de la maladie l'année suivante (Sadilek et al., 2020).

En dépit de l'accessibilité des données des réseaux sociaux et de leur utilisation en tant que ressources pour les systèmes de surveillance, l'analyse de ces données nécessite des méthodes spécifiques, et on remarque une non-structuration de la méthodologie adoptée d'où la différence dans les résultats observés (Doan et al., 2019; Tulloch et al., 2019). Dans le cadre de la maladie de Lyme par exemple, on remarque une hétérogénéité dans le choix des mots clés utilisés pour l'analyse des données provenant des réseaux sociaux et des moteurs de recherche de l'internet. On observe dans les écrits recensés qu'il existe une vaste panoplie de mots-clés que les chercheurs associent avec la maladie de Lyme, et les démarches pour choisir ces mots-clés varient selon la région et la source de données et du modèle utilisés (Basch et al., 2017; Kapitány-Fövény et al., 2019; Kim et al., 2020; Kutera et al., 2021; Sadilek et al., 2020; Scheerer et al., 2020; Seifter et al., 2010; Tulloch et al., 2019). En effet, selon Kutera et collaborateurs, seul le mot-clé *« Lyme disease »* est prédicteur de la prévalence de la maladie tandis que pour Kim et ses collaborateurs, « *tick bite »* corrèle avec la maladie de Lyme (Kutera et al., 2020; Kim et al., 2020). D'après Seifter et collaborateurs, « *tick bite* » et « *cough* » était associés avec la distribution des cas de la maladie de Lyme (Seifter et al., 2010). La diversité de mots-clés choisis pourrait expliquer la divergence des résultats des études qui ont tenté de tester ou de prouver la performance de modèles développés à partir de données provenant de l'activité en ligne. Il est nécessaire de mettre au point un consensus sur les mots-clés qui corrèlent le mieux avec la maladie de Lyme dans les discussions sur les réseaux sociaux.

3.4 Objectifs de l'étude :

Notre étude vise à développer un système qui permet d'identifier des tweets associés à la maladie de Lyme comme étant ou non des cas potentiels de Lyme. L'objectif de cette étude étant ainsi d'établir une base de données comme outil complémentaire de surveillance de la maladie de

Lyme capable de détecter des cas potentiels qui aurait échappé au système de surveillance de base en temps réel. Cette étude sera effectuée sur un échantillon de tweets en anglais collecté à travers le monde entre 2010 et 2022 et se basant sur une sélection large des mots clés reliés à la maladie. Nous utiliserons des algorithmes de classification robustes tels que les modèles de classification basés sur les transformateurs BERT, tout en incluant des emojis comme un outil d'enrichissement. Les retombées de cette étude sur la santé publique sont (1) de constituer une base de données des tweets classés et étiquetés qui servira à l'entraînement des modèles pour la classification automatique des tweets futurs liés à la maladie de Lyme et la prédiction des zones à risque et des zones d'émergence ou réémergence de la maladie; (2) d'étudier la distribution des tweets actuels reliés à la maladie de Lyme dans le monde. Finalement, cette étude a aussi des retombées méthodologiques qui sont (1) d'analyser la performance des modèles de classification robustes basés sur des transformateurs BERT (BERTweet, DistilBERT et ALBERT) pour prédire les cas potentiels de la maladie de Lyme en les comparant avec plusieurs classificateurs classiques tels que (TF-IDF, Naive Bayes et autres); (2) d'évaluer l'effet de l'intégration d'emojis en tant que fonctionnalités d'enrichissement pour améliorer les performances du classificateur à base de transformateur.

Chapitre 4- [Cadre méthodologique de notre étude]

4.1 Cadre analytique pour notre étude

En dépit de l'abondance des données provenant des réseaux sociaux, l'analyse exploratoire de ces données constitue un travail complexe nécessitant des techniques d'apprentissage automatique pour extraire l'information utile. Les textes générés sur les réseaux sociaux, sont truffés d'erreurs orthographiques, de mots dissociés du reste du texte (bruit), de termes linguistiques non standard (Stoové & Pedrana, 2014). Il est complexe pour les humains de déterminer et de filtrer les discussions sur les réseaux sociaux tels que Twitter comme des potentiels cas ou non de la maladie de Lyme en raison de leur grand volume (Pollett et al., 2017). Le traitement du Langage Naturel (NLP) est une technique d'apprentissage automatique qui permet l'analyse des types des textes ainsi que d'extraire de l'information pertinente dans les réseaux sociaux (Elnaggar et al., 2022; Gavrielov-Yusim et al., 2019; Richard & Reddy, 2021). Ces techniques visent à développer des algorithmes qui sont capables par la suite de reconnaitre des mots et classifier un texte comme en lien ou non à un contexte donné (Elnaggar et al., 2022). Ces techniques explorent les données textuelles et représentent les relations existantes entre les mots d'un texte donné sous forme de *jetons* encore appelés *tokens* (fragments de texte ou mots qui se représentent sous forme vectorielle) (Conway et al., 2019; Doan et al., 2019). Les modèles de NLP peuvent être utilisées dans le cas d'un apprentissage supervisé, semi-supervisé ou non supervisé. L'apprentissage supervisé nécessite une annotation manuelle de la base de données dans le but de prédire des nouvelles observations contrairement à l'apprentissage non-supervisé qui classifie les données dans des classes selon leurs similarités sans passer par un processus d'annotation manuelle. La première approche nécessite un travail laborieux et exhaustif surtout lors du traitement d'un grand volume de données (Conway et al., 2019; Doan et al., 2019). En revanche, la seconde peut être très

peu fiable et peut présenter des scores de classification très faibles avec des données bruitées et non-structurées comme celle provenant des réseaux sociaux (Liew & Lee, 2021). Un juste milieu est l'utilisation de l'apprentissage semi-supervisé qui nécessite une annotation préalable faite manuellement sur une partie de ces données dans le but d'entrainer le modèle de classification pour s'assurer de leur capacité à classifier correctement ces données non annotées (Goodman-Meza et al., 2022). Le but de l'annotation manuel des tweets nettoyés est d'obtenir une classification juste et plus précise afin d'entraîner le modèle de classification et d'obtenir des résultats de prédiction avec plus d'exactitude et de précision (Conway et al., 2019). Les algorithmes qui seront employées dans le cadre de ce mémoire se baseront sur la vectorisation des mots c'est-à-dire la représentation des mots ou un fragment de texte qui constitue un texte sous forme de vecteurs pour capter les informations contextuelles (Jiang et al., 2018a; Shibayama et al., 2021). Nous commençons par la description des techniques classiques de NLP: (1) TF-IDF; (2) Naive Bayes; (3) k-Nearest-Neighbors; (4) Support vector machine; (5) Logistic regression; (6) Quadratic Discriminant Analysis; (7) AdaBoosting; (8) Random Forest.

(1) *TF-IDF* : est une technique qui se base sur la fréquence de terme (TF) pour le traitement des données, une méthode facile à implémenter et avec peu d'exigences. Elle est la plus utilisée dans l'analyse des données provenant de twitter (Mackey et al., 2020). Elle consiste à marquer les mots et indiquer leur fréquence dans un texte (Nasser et al., 2021); (2) *Naive bayes* (NB) : est une technique d'apprentissage automatique qui analyse des données textuelles en considérant que tous les mots ont la même importance et ne tient pas compte de la position du mot dans le texte (Costa et al., 2013). Cette méthode est basée sur les approches bayésiennes, nécessitant l'application du théorème de Bayes basée sur une hypothèse de simplification dite naïve (Byrd et al., 2016; John & Langley, 1995); (3) *K-Nearest Neighbors* (KNN) est une technique d'apprentissage automatique

qui permet de développer un modèle de classification qui prédit la classe à laquelle appartient un mot en examinant les classes des mots qui entourent ce mot donné. La classification est calculée à partir d'un vote à la majorité simple des k-voisins les plus proches de chaque point. La difficulté de ce modèle est le choix optimal des voisins à considérer (Pedregosa et al., 2011); (4) *Support vector machine* (SVM), une technique courante de classification ou de régression qui se base des projections sur des hyperplans d'un ensemble de données permettant dans le cadre de classification, de mettre les données dans deux ou plusieurs groupes (Ojo et al., 2022); (5) *Random forest (RF),* un type de classificateur qui s'adapte à différents classificateurs d'arbre de décision en parallèle sur plusieurs sous-échantillons ce qui augmente considérablement sa précision. La forêt aléatoire est un algorithme d'apprentissage automatique couramment utilisé, qui combine les résultats de plusieurs arbres de décision (structure hiérarchique de décisions binaires) pour obtenir un seul résultat (Tang et al., 2021); (6) *Adaptive Boosting* (AdaBoost), un algorithme de classification adaptative qui se base sur une approche itérative visant l'amélioration de l'apprentissage sur la base des erreurs des autres modèles (Conway et al., 2019; Mahroum et al., 2018); (7) *Quadratic Discriminant Analysis* (QDA), une technique qui développe des fonctions de classification avec des variables indépendantes non linéaires en supposant que la classe de ces variables suit une distribution gaussienne (Hasan et al., 2021). (8) Logistic regression (LR) une type de régression qui se base sur la classification d'une variable catégorielle en fonction des variables étudiées (Shariatnia et al., 2022).

Les modèles de classification classiques mentionnées plus haut, représentent les mots individuels sous forme de vecteurs sans prendre en compte le contexte dans lequel le mot est utilisé donc, elles ne comprennent pas le contexte des mots et leur développement nécessite un grand ensemble de données annotées manuellement pour chaque application spécifique (D. Choi et al.,

2020). Aussi ils ne prennent pas en compte les représentations sous-jacentes de ces mots ou fragment de texte. Étant donné la nature des données des réseaux sociaux qui ne sont pas toujours structurées et qui sont remplis de « bruit »; ces modèles ont tendance à ne pas performer d'une façon optimale (Jiang et al., 2018a).

Les modèles de classification basés sur les Transformers BERT ont la particularité de capturer les informations sémantiques contextuelles et donc de comprendre les différents contextes d'un mot utilisé. Ces modèles permettent la représentation vectorielle de longues séquences de texte (Qasim et al., 2022). Dans le but de ce mémoire, nous nous intéressons à ces types modèles de classification basés sur les transformateurs BERT. Ce modèle est capable de capturer des informations sémantiques contextuelles dans de grand volume de texte non structuré comme les données textuelles provenant des réseaux sociaux (Qasim et al., 2022). Aussi les modèles BERT ont l'avantage d'être spécifiques au langage et donc un sous modèle peut être développer pour une langue particulière (Devlin et al., 2019; J. Lin et al., 2021a).

Selon les développeurs de ces modèles, il s'agit des modèles de compréhension du Language (Devlin et al., 2019). Les modèles de classification basés sur les transformateurs (BERT) sont des représentations d'encodeurs bidirectionnels préservant le contexte des mots à partir de transformateurs qui sont un réseau neuronal formé pour apprendre une langue (J. Lin et al., 2021a). Le modèle BERT est caractérisé par sa capacité à prédire les mots cachés en considérant tous les mots insérés dans le modèle, par conséquent ces modèles prennent en compte la dépendance existante entre les groupes de mots (Benítez-Andrades et al., 2022). BERT a une structure de modèle dans laquelle plusieurs couches de codage sont empilées et sont chargés de trouver des relations complexes entre les représentations d'entrée et de les encoder dans la sortie (Qasim et al., 2022). La bidirectionnalité des classificateurs BERT permet aux modèles de considérer le texte

comme un ensemble en prenant en compte les mots en avant et après chaque mot (Qasim et al., 2022). Autrement dit ce modèle fonctionne avec un mécanisme d'intégration dynamique qui, non seulement prend en compte l'ordre des mots dans un texte, mais aussi la signification différente des mots selon le contexte en les reproduisant sous forme de représentations vectorielles. Les modèles de classification BERT sont formés des trois tokenisateurs spéciaux (responsable de fragmenter les mots du texte et leur attribuer un code): [CLS], [SEP] et [PAD] (Devlin et al., 2019). Le [CLS] est un jeton ajouté manuellement au début de chaque phrase et l'état caché de ce jeton correspond à la totalité de la phrase. Le second jeton [SEP] est rajouté à la fin de la première phrase et représente la prédiction de la phrase suivante. Il constitue ainsi le lien reliant la fin de la première phrase et le commencement de la phrase suivante, puis successivement. Le modèle BERT fonctionne habituellement avec une longueur de phrase maximale qui dépend du type de données. Pour les phrases plus courtes, le modèle ajoute des jetons vierges pour atteindre cette longueur maximum en insérant le jeton [PAD]. Le modèle BERT de base utilise 12 couches de blocs transformateurs avec la dimension de la couche cachée de 768 avec environ 110 millions de paramètres entraînables avec 12 têtes d'auto-attention, donc le temps d'entraînement est relativement long soit environ 1 jour (Devlin et al., 2019; J. Lin et al., 2021a).

Ce modèle a obtenu des résultats très satisfaisants dans le domaine du traitement du langage naturel et est considérée comme un des modèles de classification les plus performants (J. Lin et al., 2021b). Contrairement aux modèles de classification classiques, BERT prédit les jetons masqués en considérant tous les jetons (en avant et en arrière du jeton considéré) alors que les modèles classiques ne considèrent que les jetons précédents pour la prédiction (Devlin et al., 2019). Donc il existe plusieurs versions du modèle de base. Dans ce mémoire nous focalisons sur 3 versions à

savoir : BERTweet, DistilBERT, ALBERT (Lan et al., 2020; Nguyen et al., 2020; Sanh et al., 2019; Silva Barbon & Akabane, 2022).

(1) BERTweet est un modèle de BERT à grande échelle pour les Tweets en anglais publié en 2020 (Nguyen et al., 2020a; Silva Barbon & Akabane, 2022). Il a la même architecture que BERT de base, mais il est plus spécifique aux Tweets; (2) ALBERT, est un classificateur qui est légèrement variant du BERT pour réduire le nombre total de paramètres afin d'optimiser les configurations à grande échelle et l'efficacité de la mémoire. Ce variant du modèle BERT a moins de paramètres. Il intègre deux techniques de réduction des paramètres. Les techniques de réduction des paramètres agissent également comme une forme de régularisation qui stabilise l'entraînement et aide à la généralisation (Lan et al., 2020); (3) DistilBERT un classificateur proposé par Sanh et collaborateurs est un modèle de transformateur rapide. Il a 40% moins de paramètres que BERT et fonctionne 60% plus rapidement tout en préservant plus de 95% des performances de BERT telles que mesurées sur le benchmark de compréhension de la langue GLUE (Sanh et al., 2019; Silva Barbon & Akabane, 2022).

L'intégration des émojis dans le traitement des textes peut aussi améliorer la classification. Par ailleurs l'analyse des sentiments dans le traitement du langage naturel est un processus complexe qui permet d'extraire l'attitude d'un individu envers un évènement. L'analyse de ses sentiments est faite en fonction de la polarité du texte c'est-à-dire qu'un classificateur de sentiment détermine s'il s'agit d'un sentiment positif, négatif ou neutre en tenant compte du contexte dans lequel le sentiment est inséré (Prottasha et al., 2022). Les émojis sont fréquemment liés à un texte ou un groupe de mots et sont rarement utilisés seuls. L'intégration des émojis dans l'analyse des données provenant des réseaux sociaux peut aider à exprimer l'émotion ressentie en écrivant le message et donc renforcer ou clarifier le sens de ce tweet (Cavalheiro et al., 2022 ; Paggio & Tse,

2022). Nous avons donc pris en compte seulement les émojis qui sont intégrés à un tweet afin de former un sens complet, parce que selon des études antérieures l'intégration des émojis à un texte contribue à améliorer la clarté et la crédibilité de ce texte et donc améliore l'efficacité et la performance des modèles de classification (Bai et al., 2019 ; Dai & Wang, 2019 ; Prottasha et al., 2022 ; Zhou et al., 2017). **Demoji** est une libraire qui permet de trouver avec précision des émojis et leur attribuer leur sentiment approprié dans un texte à l'aide d'un référentiel d'émojis (Liu et al., 2021; Lu et al., 2018). Cette librairie servira donc dans notre étude à attribuer de manière approprié les sentiments véhiculés en se basant sur les émojis utilisés dans les tweets.

Nous avons utilisé un sous-domaine de l'analyse des sentiments pour détecter l'émotion de l'auteur des tweets (peur, joie, tristesse, etc.) en insérant des émojis qui les accompagnent (Albu & Spinu, 2022).

4.1.1 Gestion des données

Généralement lorsque les chercheurs analysent les données provenant des réseaux sociaux, il est adapté de diviser l'ensemble des données en 3 catégories : un jeu de données pour le test, un jeu de données de validation et un jeu de données pour l'entrainement. Le jeu de données d'entrainement est classifié manuellement dans notre cas, et un modèle de classification formé sur ce jeu de données est appliqué pour annoter automatiquement le jeu de données de validation. Le jeu de données de tests sert à évaluer la vraie performance des modèles de classification utilisées.

4.1.2 Évaluation de la performance des modèles de classification

La performance des modèles de classification est déterminée par 4 métriques à savoir : l'exactitude (*Accuracy*), la précision (*Precision*), le rappel (*Recall*) et le score moyen F1 (*F1 score*) (Benítez-Andrades et al., 2022; D. Choi et al., 2020). L'exactitude est déterminée par le rapport

entre les données prédites correctement et le nombre total des données test; le rappel indique la complétude (l'intégralité) du modèle; la précision indique l'exactitude des classificateurs et le score F1 moyen est une moyenne harmonique des scores de précision et de rappel (Gavrielov-Yusim et al., 2019).

Mathématiquement ces paramètres se calculent de la façon suivante (Keldenich, 2021):

(1) $Exactitude = \frac{Nombre\ de\ prédictions\ correctes}{nombre\ de\ données\ dans\ le\ jeu\ de\ données\ pour\ le\ test} * 100$

(2) $Précision = \frac{vrais\ positifs}{vrais\ positifs + faux\ positifs}$

(3) $Rappel = \frac{Vrais\ Positifs}{Vrais\ positifs + Faux\ Négatifs}$

(4) $Score\ moyen\ F1 = 2 * \frac{Rappel * Précision}{Rappel + Précision}$

4.2 Méthodologie de notre étude

La démarche méthodologique que nous proposons dans cette étude pour l'exploitation des données provenant des réseaux sociaux (Twitter dans le cadre de cette étude) peut se décomposer en deux étapes. La première étape consiste en premier lieu à collecter des données textuelles c'est-à-dire de recueillir des tweets sur Twitter qui semble en relation avec la maladie de Lyme ; de nettoyer les données recueillies puisqu'elles contiennent également de l'information non désirée qui rend difficile la compréhension de ces tweets (comme les liens, les hashtags, etc.) ; puis de subdiviser l'ensemble des données recueillies en 3 groupes différents (validation, entraînement et test) ; de classifier manuellement un échantillon comme étant un cas potentiel ou non de la maladie de Lyme ; et enfin de convertir les émojis qui sont des représentations de sentiments utilisés par les utilisateurs de Twitter pour communiquer une émotion en des mots correspondants (Denecke

et al., 2013; Grajales et al., 2014). La deuxième étape consiste à classifier le reste des données collectées comme cas potentiel ou non de la maladie de Lyme en utilisant des modèles de classification basés sur des transformateurs et à identifier leur distribution spatiale grâce à la géolocalisation des tweets, un de avantages de Twitter (Benítez-Andrades et al., 2022). Dans les lignes suivantes, nous expliquons chacune de ces étapes.

4.2.1 La collecte des données

La collecte des tweets relationnés à la maladie de Lyme peut se faire de deux façons. La première façon est d'avoir accès des tweets à travers API Twitter qui est une l'interface d'application de Twitter et qui permet d'avoir accès à 1% des tweets choisis aléatoirement sur le sujet désiré, gratuitement. La deuxième façon consiste à avoir accès à la totalité des données de Twitter en payant. Dans notre étude nous avons collecté nos données tweets via API Twitter à l'aide des mots clés comme *#Lyme* et *#Lyme disease* compte tenu du coût qu'implique la deuxième approche. Nous avons choisi des termes généraux pour la première extraction des données dans le but de collecter le plus de tweets relationnés à la maladie de Lyme parce qu'un choix de mots clés trop spécifiques limiterait la quantité des tweets collectés (Abilhoa & Castro, 2014; Aiello et al., 2020; Jayasiriwardene & Ganegoda, 2020). Nous avons collecté un ensemble de 1.3 million tweets allant de la période de 2010 à 2022.

4.2.2 Nettoyage et Filtrage

Nous avons ensuite nettoyé les tweets en supprimant le symbole (#), les liens URL, les balises *HTML*, les mots vides, et les mentions de nom d'utilisateur. Le nettoyage des tweets permet de réduire ce qu'on appelle « le bruit » et la redondance des tweets et donc améliore la performance des modèles de classification par la suite (Denecke et al., 2013; Edo-Osagie, Smith, Lake, Edeghere, & Iglesia, 2019; Janssens et al., 2014). La sélection des mots-clés (des termes de

recherche) pour le filtrage en amont des tweets auto-déclarés collectés dans le but de les classifier comme cas potentiels de la maladie de Lyme nécessite une attention toute particulière puisque l'exactitude des données recueillies pour construire les algorithmes d'analyse en dépend (Edo-Osagie, Smith, Lake, Edeghere, & Iglesia, 2019). Eysenbach affirme que le critère fondamental de l'étiquetage des données de santé serait de l'importance médicale de l'information collectée, donc les mots-clés portant sur la symptomatologie devraient être pris en compte (Eysenbach, 2002).

Dans cette étude, étant donné que l'objectif est de détecter les cas potentiels de la maladie de Lyme en se basant sur les tweets, nous avons opté pour l'inclusion des symptômes de la maladie de Lyme bien que non spécifiques dans l'ensemble, parce qu'un ensemble plus large de mots-clés peut nous aider à délimiter les faux négatifs dans la classification des tweets (Jayasiriwardene & Ganegoda, 2020; Jiang et al., 2018b). Nous avons ainsi inclus une liste exhaustive des symptômes de la maladie de Lyme basée sur la description des symptômes selon la CDC qui sont regroupés dans le tableau 2 (CDC, 2021)

Système affecté	**Manifestations cliniques**
Peau	*Erythema migrans, borrelial lymphocytoma, rash*
Système nerveux	*numbness, sleepiness, migraine, tingling, neck stiffness, dizziness, nerve pain, depression, memory loss*
Système ostomioarticulaire (Lyme arthritis)	*Arthritis, joint pain*

Système cardiaque (Lyme carditis)	*Carditis, irregular heart beats, palpitations,*
Symptômes non spécifiques	*Fever, nausea, headaches, vomiting, fatigue, swollen lymph nodes*

Tableau 2. – Symptômes médicaux utilisés comme mots-clés dans le filtrage par mots-clés des tweets collectés.

À la fin de cette étape nous avons limité notre extraction à des tweets en anglais et notre base de données textuelle comptait 20 000 tweets. La majorité des tweets provenant des États-Unis d'Amérique, nous nous sommes assurés de l'hétérogénéité de notre base de données en sélectionnant 10% des 20 000 tweets sélectionnés dans d'autres pays.

4.2.3 Étiquetage manuel des données et conversion des émojis :

Nous avons procédé à un étiquetage manuel des tweets comme Lyme ou non-Lyme à l'aide d'une liste de mots-clés précis sur les données de la base d'entraînement. L'étiquetage que nous adoptons se base sur la classification binaire. La classification binaire comme son nom l'indique, sépare les données textuelles de Twitter en deux classes où, une classe représente les tweets reliés à la maladie de Lyme, tandis que l'autre classe représente les tweets non reliés à la maladie de Lyme (Mackey et al., 2020; Shorten et al., 2021). Une étiquette « 1 » est attribuée aux tweets qui sont directement liés à la maladie de Lyme, tandis qu'un « 0 » est attribué à ceux qui ne le sont pas. Les conflits d'étiquetage sont tranchés à l'unanimité par l'ensemble de l'équipe de recherche. Enfin nous avons converti les emojis en mots de sentiments en utilisant la librairie **demoji**[1].

[1] Available at: https://pypi.org/project/demoji/

4.2.4 Classification des tweets avec des modèles de classification

Après la collecte, le filtrage et le nettoyage de nos données, nous avons obtenu une base constituée de 20 000 tweets en anglais provenant du monde entier de 2010 à 2022. Cette base est subdivisée en 3 jeux de données : un jeu de données pour test, un jeu de données pour l'entraînement et un jeu pour la validation de notre modèle. Le jeu de données de l'entraînement est constitué de 12 000 tweets sélectionnés aléatoirement à raison de 1 000 tweets pour chacune des années de notre étude (2010-2022). Les tweets du jeu de données d'entrainement sont ensuite distribués en 2 classes de 6 000 tweets chacun, étiquetés 0 et 1 selon qu'ils sont des cas potentiels de la maladie de Lyme ou non. Les 8 000 tweets restants sont distribués à part égale dans les jeux de données de validation et de tests. Ensuite, les tweets nettoyés ont été introduits dans des tokeniseurs de mots, qui les ont convertis en une séquence de tokens lemmatisés parsemés des trois tokens spéciaux à savoir le [CLS], [SEP] et le [PAD]. Cela est fondé sur l'utilisation des algorithmes de classifications basés sur des transformateurs de type BERT. Ceci est fait dans le but que le reste des tweets, soient classifiés en mode binaire maladie de Lyme ou non, c'est-à-dire, prédire l'étiquette de classe correspondante aux restes des tweets.

Nous avons comparé la performance des modèles basés sur les transformateurs à d'autres modèles de classification linguistiques pour assurer la validité de nos résultats. Lorsqu'un tweet obtenait la probabilité la plus élevée d'appartenir à la classe de la maladie de Lyme, nous avons utilisé la bibliothèque GeoPy[2] sur Python qui permet la géolocalisation des tweets pour estimer son emplacement géographique.

[2] Available at: https://geopy.readthedocs.io/

4.3 Considérations éthiques :

Aucun certificat éthique n'a été délivré pour la réalisation de cette étude puisque les données provenant des réseaux sociaux sont considérées dans leur ensemble comme des données publiques. De plus, depuis l'automne 2014, la politique de confidentialité de Twitter mentionne maintenant que les universitaires peuvent avoir accès aux publications des utilisateurs dans le cadre de recherche (Fiesler & Proferes, 2018).

Chapitre 5- [Manuscript Exploiting Self-reported worldwide Tweets for Identifying Potential Lyme Disease Cases Using Deep Learning Models enhanced by sentimental words through emojis]

Elda K. E. Laison [1], Mohamed H. Ibrahim[1,2], Srikanth Boligarla[3], Jiaxin Li[3], Raja Mahadevan [3], Austen Ng [3], Lee W. Yi[3], Jang Park [3], Yijun Yin [3,] Bouchra R Nasri [1].

1. Department of Social and Preventive Medicine, École de Santé Publique, University of Montreal, Montréal, Canada
2. Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt.
3. Harvard Extension School, Harvard University, Cambridge, United States.

Abstract

Background: Lyme disease is the most prevalent tick-borne disease in the Northern Hemisphere. Delayed treatment can exacerbate symptoms and result in more severe cases, making this condition a major public health concern in the coming years. Additionally, the Lyme disease surveillance system relies on healthcare professionals to report cases, which weakens the system's efficiency in having accurate data since only the cases seeking medical attention are reported. Thus, there is a need to enhance the surveillance tools of Lyme disease using other data sources such as web-data.

Objective: Worldwide Twitter data was analyzed to understand its potential and its limitations as a tool for Lyme disease surveillance. The proposed Twitter data system is primarily a transformer-based classifier that leverages self-reported tweets to identify potential cases of Lyme disease.

Methods: We first used approximately 20,000 English tweets collected worldwide from a database with more than 1.3 million tweets related to Lyme disease. Because most Lyme disease tweets are from the US, we selected only 20,000 tweets, from which about 10% represented other countries than the US, to capture more variability across countries. After preprocessing and geolocating the tweets, a set of carefully selected keywords was used to manually label a subset of tweets to classify them as potential or non-Lyme disease cases. Emojis were converted to sentiment words and then used in place of emojis in the tweets. The dataset of labelled tweets was then used to train, validate, and test the performance of three transform-based classifier variants, namely ALBERT, DistilBERT, and BERTweet, to classify the remaining and other new tweets.

Results: The empirical results showed that BERTweet is the best classifier among all classification models evaluated, with the highest average F1-score of 89.3%, classification accuracy of 90.0%, precision of 97.1%, except for the recall where TF-IDF and k-Nearest Neighbors performed better by 93.2 % against 82.6% for BERTweet. When emojis' expressions were used to enrich the tweet embeddings, the recall score for BERTweet increased by 8%, and DistilBERT had a markedly increased F1-score of 93.8% (+4%) and a classification accuracy of 94.1% (+4%), while ALBERT had a F1-score of 93.1% (5%) and a classification accuracy of 93.9% (+5%).

Conclusions: This study revealed several key findings. First, that BERTweet and DistilBERT can serve as robust NLP classifiers to identify self-reported potential cases of Lyme disease. Second, emojis are effective as enrichment features to improve the accuracy of the tweet embedding and the performance of transformer-based classifiers. In particular, the emojis reflecting sadness, empathy, and encouragement can help reduce false negatives. Third, the general awareness of Lyme disease is high in the United States, the United Kingdom, Australia, and Canada as self-reported potential cases of Lyme disease on Twitter from these countries accounted for more than 50% of the collected English tweets, while Lyme disease-related tweets are scarce in countries from Africa and Asia. Finally, the most commonly reported symptoms of Lyme disease are rash, fatigue, fever, and arthritis while symptoms such as borrelial lymphocytoma, palpitations, swollen lymph nodes, neck stiffness, and irregular heartbeat are unusual and rare.

Keywords: Lyme disease; Twitter; BERT; Emojis; machine learning

Introduction

The incidence of tick-borne diseases is increasing. Their spatial distribution is getting wider over time, owing primarily to global warming and milder winters, contributing to the tick vectors' range expansion [26, 37]. Lyme disease is the most prevalent tick-borne disease in the northern hemisphere, including North America, Europe, and some Asian countries [47, 49]. As such, the early detection of potential Lyme cases will remain a public health concern in the forthcoming decades [45].

Lyme disease is caused by a spirochetal bacterium, an infectious agent known as the *Borrelia burgdorferi sensu lato* (*s.l.*) complex [45]. The *Borrelia burgdorferi sensu lato* (*s.l.*) complex contains numerous genospecies, but only a few can infect humans and cause Lyme disease, and they have distinct geographic distributions. *B. burgdorferi sensu stricto* is primarily found in North America, whereas *B. afzelii* and *B. garinii* are prevalent in Asia and Europe [13, 16]. Furthermore, the clinical manifestations of Lyme disease vary depending on the genospecies involved in the infection and, therefore, by geographical region. In North America, *B. burgdorferi sensu stricto* (*s.s.*) causes mostly Lyme arthritis and carditis, whereas *B. afzelii* and *B. garinii* cause neuroborreliosis in Europe and Asia [49, 3, 4].

The infectious agent is transmitted to humans by several species of ticks from the *Ixodes* genus, whose distribution varies by location: In North America, *Ixodes scapularis* and *Ixodes pacificus* are the most prevalent; in Europe, *Ixodes ricinus* are the most popular, and *Ixodes persulcatus* is the Asian Lyme disease vector [8, 14, 10, 27, 36]. Tick vectors undergo various life stages such as egg, larval, nymphal, and adult. Tick hosts differ depending on their growth stages during the larval and nymphal stages. Ticks primarily feed on rodents (mice), whereas adult ticks prefer larger mammals such as deer [22].

Although seropositivity for the infectious bacteria indicates previous asymptomatic exposure rather than active exposure, the standard laboratory diagnostic of Lyme disease involves a two-tier test in which an ELISA screening test is confirmed by a second test, which can be a western blot or an immunoblot [32]. Lyme disease is a tick-borne disease treatable with a short course of antibiotics, but if left untreated, it may lead to severe neurological, cardiac, and articular complications [20]. Since there is no vaccine to protect humans, the only protective measures against tick bites are self-protection and yard management [43].

Underreporting is a concern in Lyme disease epidemiology because the traditional surveillance system has failed to track all cases accurately [31, 34, 44]. For example, a recent study estimated the number of Lyme disease cases in the United States at over 400,000, while the Centers for disease control and prevention (CDC) reported 30,000 cases [11, 44]. The United States is not the only country where underreporting of Lyme disease cases has been suggested; some European countries as well [39, 46, 50]. According to a review conducted in [6], the traditional Lyme disease surveillance system is prone to overreporting or underreporting. One reason is that the system relies on busy health care professionals to report cases. Thus, only cases seen and diagnosed by professionals are reported. Because of the difficulty related to Lyme disease diagnosis, some cases tend to get missed by healthcare professionals, especially in new areas, resulting in underreporting of the disease [11].

With the extended use of the Internet and social media platforms where health-related information is shared, researchers have found in web-data an opportunity to improve surveillance systems. This new field of research is referred to as digital surveillance or Infodemiology [28]. Among the numerous social media existent platforms, Twitter is the most widely used as a digital surveillance tool because its data can be easily accessed through the Twitter API; the brevity of the text (tweets) is limited to 280 characters (140 characters before 2017); Twitter is the most popular social media platform with over 145 million daily active accounts; and the possibility to geolocate the tweets [60].

Several studies have been proposed using data from search engines and social media platforms to track Lyme disease [52, 53, 54]. For example, [52], examined how the content of Lyme disease videos on YouTube differed depending on data sources and the people who produced the videos. It was reported that public health experts did not produce popular videos on YouTube about Lyme

disease. In addition, responsible reporting and innovative knowledge translation through videos can increase awareness of Lyme disease. To better forecast the incidence of Lyme disease in Germany, the authors in [53] utilized traditional data from Google Trends. While the reported incidence of Lyme disease correlates well with Google Trends data, it did not significantly increase the forecasting accuracy. In [54], the prevalence of Lyme disease and the frequency with which the term "Lyme" was searched in Google Trends were examined in southern Ontario, Canada, between 2015 and 2019, yielding to the identification of a single hotspot in eastern Ontario. Additionally, there was an increase in Google Trends for the term "Lyme disease", which was associated with a significant increase in Lyme disease risk. According to [55, 56], the number of Lyme disease searches in search engines is related to seasonal and geographic patterns of Lyme disease cases. Surprisingly there are very few studies on Lyme disease and social media. For example, [57] showed that Twitter can be used to monitor Lyme disease by using Twitter data as a proxy for disease prevalence in the United Kingdom and the Republic of Ireland. A limited geographic search strategy was used to discover spatial patterns and find rare cases of Lyme disease. In [7], it has been reported that Lyme-relevant Twitter data is correlated with the CDC reports.

A recent systematic review suggested that less than half of the existing studies on digital surveillance via Twitter data were focused mainly on prediction, while only a few focused on developing tools for adequate analysis of this type of data [58]. Also due to the novelty of digital surveillance in public health research, there is an unevenness in the different methodological approaches and the datasets used. In fact, given the nature of social media data, analyzing this data requires the development and evaluation of methodological approaches based on machine learning, which requires time and usually requires an organized, validated, and labelled dataset. Pre-trained and annotated datasets for various health problems are needed to facilitate their development. And

with the availability and the richness of text data from Twitter or from other platforms (such as Reddit), there is a need to develop reliable, accurate classification methods to process and analyze the data to study health-related issues [59].

Our study is coming in a very useful way in filling this gap in the literature. Indeed, our study aims to provide an accurate English worldwide tweet dataset and evaluate the performance of the selected transformer-based models with the integration of an emotional component: emojis. We believe that the novelty and completeness of this dataset will assist in the development, evaluation of digital surveillance systems regarding Lyme disease, and will be a resourceful tool for public health researchers and practitioners. It is important to mention that the current study is a continuation of a recent study where a machine learning-based model has been proposed for predicting Lyme disease cases and incidence rates in the United States using Twitter [7]. However, unlike this previous study, this work provides a worldwide dataset of English tweets and evaluates the performance of the selected advanced machine learning transformer-based models with the integration of emojis, which will lead to a new and more accurate classified data for Lyme-related tweets.

The objectives of this study are several:

1. First, to make the classified dataset openly available for academic use in a variety of experimental, epidemiological research, at a time when there is a pressing need to incorporates novel type of datasets related to the Lyme disease epidemic such as this classified dataset, with other datasets from Twitter or similar social media platforms.
2. Second, we analyzed the performance of several prominent Natural language processing (NLP) classifiers in terms of their ability to predict potential cases of Lyme disease.

3. Third, we evaluated the effect of incorporating emojis as enrichment features to improve the performance of the transformer-based classifier.
4. Finally, we explored the tweet counts to determine whether specific patterns could be identified regarding the prevalence of Lyme disease for each country.

Methods

Figure 2 [3] illustrates the methodology used to classify the tweets, which consists of two key elements: 1) collecting and preprocessing self-reported Lyme-related tweets; and 2) identifying potential Lyme-disease cases.

Data collection and Preprocessing

Using an academic research account with Twitter's application programming interface (API) and search terms like "#Lyme", and "#Lyme disease", about 20,000 English tweets were collected between 2010 and 2022. To clean tweets Hashtags, URL links, HTML markups, stop-words, username mentions, and retweets were deleted to reduce text noise and redundancy in tweets. Furthermore, accurate keywords or search terms are required to properly label extracted information from social media. Keywords used to label the collected data are important as they will have an impact on results and the quality of surveillance. Many studies have attempted to improve

[3] Cette figure est notée comme 2 étant donné que le manuscrit est inclus dans le mémoire même s'il s'agit de la figure 1 du manuscrit

the relevance of disease-related keywords by examining word frequency using a corpus of tweet texts and labeling approaches [18, 21]. As such, we compiled a list of precise keywords that often correlated with Lyme disease; these keywords were used as the basis regular expression (regex) that was applied to the cleaned tweets to manually determine whether they were relevant to Lyme disease. A label '1' is assigned to potential Lyme disease related tweets, while a '0' is assigned to those that are not. As in [7], we used two methods to specify keywords in regex: the first one entailed investigating the content of the cleaned tweets to determine the relative frequency of common colloquial Lyme-disease words such as: *have Lyme, had Lyme, having Lyme, has Lyme, get Lyme, gets Lyme, got Lyme, getting Lyme hiking, hike, forest, tick, ticks, bite, deer, deertick, tickborne*. By using this method of keyword selection, Twitter posts like "*She is terribly unwell, we suspect it's Lyme*" are labeled to be potential Lyme disease case. The second method of keyword selection involved considering most frequent Lyme disease's symptoms, transmission channels, or scientific terms as: *"erythema migrans", "carditis", "fever", "rash", "headache", "fatigue", "chills", "nausea", "vomiting", "dizziness", "sleepiness", "hallucinations", "depression", "numbness" , "tingling", "facial paralysis", "palpitations", "borrelial lymphocytoma", "anxiety", "memory loss", "joint aches", "muscle aches", "swollen lymph nodes", "neck stiffness", "nerve pain", "arthritis", "shortness of breath", "irregular heart beat", "shooting pains", "skin redness", "tick bite", "acrodermatitis chronica atrophicans".* Therefore, tweets such as "*I suffered from Lyme symptoms four years ago*" and *"My sister is developing fever after a tick bite"* were also labeled as potential cases of Lyme disease. It is important to note that the manual approach aims to ensure fair and accurate labeling of the dataset. This is because automatically labeling tweets with off-the-shelf Python regex libraries does not always provide correctly labeled tweets. It could be argued that there are differences between the two keyword selection methods we used to label the tweets manually. The rationale here was based on the fact that it has been demonstrated in [55] that the

"Lyme disease" keywords used in search engines can provide better results and, as a result, the first method of keyword selection can be viewed as a naive, general way to label any Lyme data collected from web-based sources, whereas the second method is a more specific and accurate way to confidently label tweets as Lyme.

The labeled dataset of the 20,000 cleaned tweets was split into three disjoint datasets: 1) 12,000 tweets were randomly selected from 2010 to 2022 to serve as a training dataset. Exactly 1,000 tweets were selected from each year at random. The tweets were eventually distributed between the two classes, as 6,000 were about potential Lyme-disease cases, while the remaining ones were not. 2) The remaining 8,000 tweets was then separated into two equals parts making 4, 000 tweets in each, the validating and testing datasets.

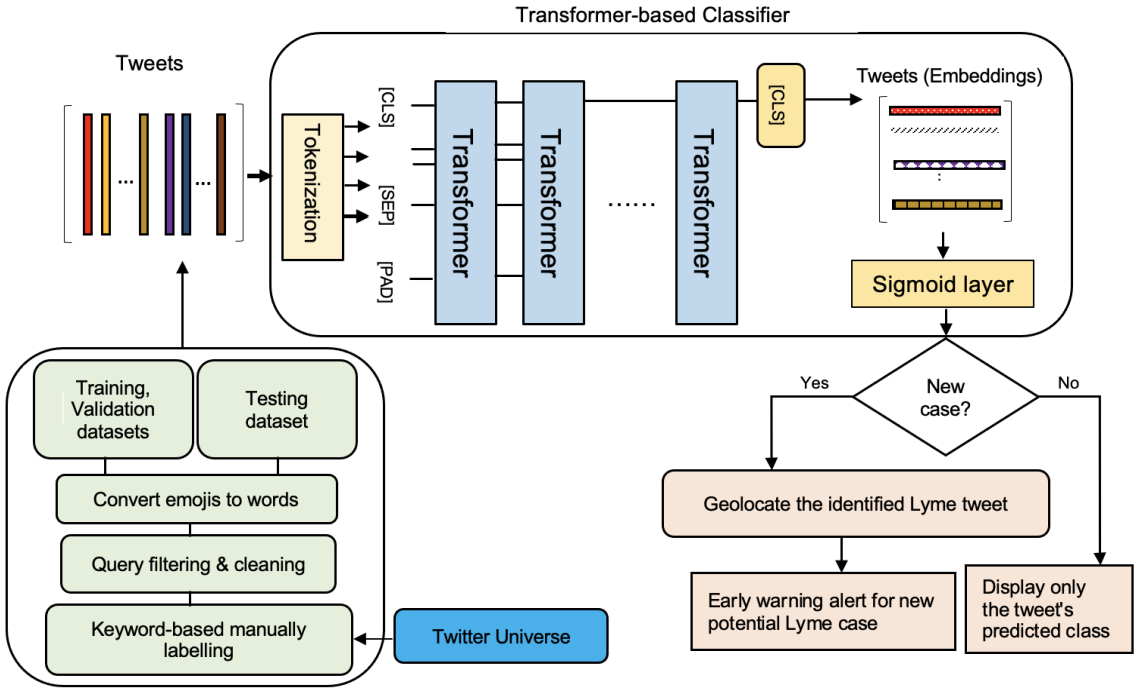

Figure 2 The two-stage approach proposed for predicting Potential Lyme disease cases.

[The first stage contains four elements: 1) We use standard search terms to collect tweets via the Twitter API; 2) We clean the tweets by removing hashtags, URL links, HTML markups, and stop-words; 3) We manually label the tweets as Lyme or non-Lyme using a list of precise keywords; and 4) We convert the emojis into sentiment words, which are then substituted for the emojis in the tweets. In the second stage, we use a transformer-based classifier to determine whether a tweet is a potential Lyme disease case or not. When a new tweet is assigned with the highest probability to the Lyme-disease class, we use the GeoPy library to estimate the tweet's location].

Detecting Lyme Disease Tweets using Transformer-based Classifiers

The process of determining whether a tweet is about Lyme disease was formulated as a binary classification problem. The training and validation datasets were used to fine-tune a set of pre-trained transformer-based classifiers so that they can identify whether a new unknown tweet is a potential Lyme disease case. Recently, transformer-based models have been highly efficient in various Natural Language processing (NLP) applications. Specifically, the Bidirectional Encoder Representations from Transformers (BERT) model [17], developed by Google AI Language in 2018, was an advancement in the transformer paradigm as it allows for the learning of token representations in both left-to-right and right-to-left directions. BERT pretraining incorporates a masked language model and next-sentence prediction, with the ability to adjust or fine-tune its parameters on other relevant datasets.

Most BERT classifier variants were typically trained to understand tweet semantic content and context to generate word embedding representations. They are language models that require a sequence of tokens as input. Thus, the cleaned tweets were fed into word-piece tokenizers, which converted them into a sequence of lemmatized tokens peppered with three special tokens: [CLS],

which stood for classification and was typically the first token of every sequence; [SEP], which described to the pre-trained language model which token belongs to which sequence; and [PAD], which was used to fill the unused token slots to ensure that the maximum token length was met. When a token sequence exceeded the maximum length, it was truncated. Several variants of BERT classifiers have been proposed, but only the three most efficient ones were considered in this study: **ALBERT** [33] is a light variant of the BERT architecture that enhances training efficiency by factorizing embedding and sharing cross-layer parameters. We used the Albert-xlarge-v2 model, which has 12 repeated layers (called transformer blocks), 4096 hidden dimensions, 128 embedding size, and 64 attention heads with 235 million trainable parameters. The AlBertTokenizer, which is associated with ALBERT, was used to tokenize each tweet into a sequence of tokens. These tokens were then synchronously fed into ALBERT's layers, where each layer used self-attention and transmitted its intermediate encoding via a feed-forward network before passing it on to the next transformer encoder block. For each token, the ALBERT model generated an embedding vector. **DistilBERT** [42] is a small and computationally efficient form of BERT. It is 40% smaller than the $BERT_{base}$ model due to knowledge distillation during pre-training, and it is 60% faster than it, all while achieving 97% of its language understanding efficiency. Compared to BERT, the number of layers in its student architecture has been trimmed in half, and token-type embeddings have been eliminated. We used the DistilBERT-base-uncased model, which has 6 layers, 768 hidden nodes, and 66M unique parameters in total. Furthermore, DistilBERT does not require token type-ids; therefore, it is not necessary to specify which token belongs to which segment. To tokenize the input sentences of the tweets into token sequences, we used the DistilBertTokenizer equipped with the model. The DistilBERT model then outputs an embedding vector for each token. **Finally, BERTweet** [35] is a recent large-scale AI model specifically for English Tweets based on BERT. BERTweet was trained on an 80 GB uncompressed corpus containing 850 million tweets streamed

from January 2012 to August 2019 and 5 million tweets related to the COVID-19 pandemic, with each tweet containing at least 10 and no more than 64-word tokens. We specifically used the BERTweet-base model, which has 12 layers (transformers block) with a hidden size of 768 and a total of 110 million unique parameters. The model's creators produced the BertweetTokenizer, which was used to tokenize the tweets' input texts into sequences of tokens. The BERTweet model also generates an embedding vector for each token. On holy-grail NLP tasks such as entity resolution and short text classification, BERTweet outperformed state-of-the-art baselines such as $\text{RoBERTa}_{\text{base}}$ and $\text{XLM-R}_{\text{base}}$ [35].

Embedding Enhancement

Emojis, which can express emotions succinctly, are popular on social media. Several potential Lyme disease patients frequently self-report their symptoms in tweets that combine word and emoji sequences. Thus, excluding emojis during preprocessing could lead to the loss of important information. As a result, we aimed to improve the tweet's contextual encoding by including its emoji expressions. Traditionally, the more efficient way of leveraging emojis to enrich feature embedding of a tweet is to use any emoji package, such as **demoji**, to convert emoji icons into sentiment words and then substitute the emoji icons with their corresponding words inside the tweets, resulting in the tweets consisting of only sequences of words that can be fed as input to the tokenizers associated with the BERT-based models.

Results

First, we compared the accuracy of the ALBERT, BERTweet and DistilBERT models, in detecting potential Lyme and non-Lyme disease tweets with the following state-of-the-art classification models: Adaboost (Ada) [24]; Random Forest (RF) [38]; Logistic Regression (LR) [30]; Multilayer Perceptron Neural Network (MLP) [25]; Support Vector Machine (SVM) [14]; k-Nearest Neighbors (KNN) [15]; Quadratic Discriminant Analysis (QDA) [48]; and Naive Bayes (NB) [25]. Using the term frequency-inverse document frequency (TF-IDF) vectorization method [41], the tweet embeddings were generated and then fed into the classifiers, except for the three transformer-based classifiers associated with their tokenizers. As in [7], we regularized our classifier models to avoid overfitting by including extreme penalizing terms in the objective functions with L1/L2 together with solvers like liblinear, lbfgs, and saga [59, 60]. The learning rate was 0.01 and the number of estimators was 100. Since Twitter data are short text data, we chose the Adam algorithm, which has been shown to better handle potential problems associated with such data and has low sensitivity to the learning rate [17, 61]. In order to maximize the likelihood estimation, we also evaluate the loss function by implementing the binary cross-entropy [7]. As in [7] we used a learning rate of -2×10^{-5}, a weight decay of 0.001, and a batch size of 64.

To ensure consistent results across evaluations, all the classification models were built using the same training, validation, and test datasets. Specifically, after combining the training and validation datasets, we used 10-fold cross-validation to train the underlying classification models. Thus, nine of the ten folds were used in the training phase to iteratively learn the model parameters, and the remaining fold was used for validation. We used all learned classifiers to predict tweet labels during the testing phase and then recorded their confusion matrices on the testing dataset to capture the following quantities: (1) the proportion of actual Lyme tweets correctly classified as potential

Lyme cases (i.e., true positives [TPs]); (2) the proportion of actual non-Lyme tweets correctly classified as unrelated to Lyme disease (i.e., true negatives [TNs]); and (3) the proportion of actual non-Lyme tweets incorrectly classified as belonging to the potential Lyme-disease class (i.e., false positives [FPs]); and (4) the proportion of actual potential Lyme-disease tweets misclassified as non-Lyme-disease tweets (i.e., false negatives [FNs]). We computed several evaluation metrics based on confusion matrices to assess the accuracy of all tested classifiers: classification accuracy [29], which measures the proportion of correct predictions (TPs and TNs) among all examined tweets; the average F1-score [40], which quantifies the likelihood of correctly identifying Lyme-disease tweets; and precision and recall, which quantify the proportion of correctly identified tweets that are actual potential Lyme-disease cases, and vice versa, respectively. The LR classification was considered to serve as an effective baseline for comparison.

Model	**Accuracy (%)**	**F1-score (%)**	**Precision (%)**	**Recall (%)**
TF-IDF and Adaboost	76.6	76.0	96.5	62.7
TF-IDF and Random Forest	76.6	76.0	96.5	62.7
TF-IDF and Logistic Regression	76.6	76.7	93.4	65.0
TF-IDF and Multi-layer Perceptron Neural network	76.5	75.9	96.9	**62.3**
TF-IDF and SVM	76.5	76.5	93.4	64.8
TF-IDF and k-Nearest Neighbors	**71.8**	79.6	**69.4**	**93.2**
TF-IDF and Quadratic Discriminant Analysis	76.6	76.6	93.5	64.8
TF-IDF and Naive bayes	73.7	**75.6**	83.7	68.9

DistilBERT	89.2	88.2	96.8	81.0
ALBERT	88.4	87.3	96.6	79.7
BERTweet	**90.0**	**89.3**	**97.1**	82.6

Tableau 3. – Average F1-score, Accuracy, Precision, and Recall (in%) for the classification models on the test dataset. [*The highest score values are shown in red and the lowest ones in blue.]*

As shown in Table 3, the BERTweet model is the best of all the NLP models included in our study, with the highest average F1 score of 89.3%, classification accuracy of 90.0%, precision of 97.1%, and recall of 82.6%. DistilBERT was close to BERTweet and was slightly more accurate than ALBERT. LR performed adequately in classifying tweets about Lyme disease but was significantly less accurate than ALBERT, with an F1-score of 76.7% and a classification accuracy of 76.6%. The accuracy scores of the QDA, RF, and Adaboost were comparable to those of LR, with RF and Adaboost having slightly more false negatives and fewer false positives. A false negative is identified with a recall score as low as 62.7% while a false positive is identified with a precision score as high as 96.5%. Adaboost was slightly ahead of MLP, and comparable to RF, as both had a classification accuracy of 76.2% and an F1-score of 76%. SVM and the baseline LR performed similarly, with roughly the same scores.

Notably, KNN had the lowest precision score of 69.4%, producing significantly more false positives than any of the other classifiers tested, but it also had the highest recall score of 93.2%, providing a significantly fewer false negative. With a recall score of 68.9%, NB had slightly fewer true negatives than QDA, SVM, LR, and Adaboost. With a precision score of 83.7%, it had significantly more false positives than them. Overall, and apart from the transformer-based classifiers, QDA had the most consistent performance when all metrics were considered at once,

with an F1-score of 76.5%, classification accuracy of 76.5%, precision of 93.4%, and recall of 64.8%.

Second, we investigated whether the use of emoji and emotion expressions improves the contextual encoding and classification of tweets. As described in the subsection, we first used the **demoji** library to extract emoji icons and convert them into words to enrich the tweet embeddings. We then repeated the previous procedure to classify the tweets Overall, the BERTweet still outperformed the other tested variants of the BERT classification model, with the highest average F1-score of 94.9%, classification accuracy of 95.2%, precision of 98.8%, and recall of 91.2%. DistilBERT followed BERTweet and was slightly more accurate than ALBERT. The recall score for BERTweet (with emojis) was 8% higher than its recall score without emoji, and DistilBERT and ALBERT had recall scores that are at least 9% higher than their recall scores without emojis. The three classifiers were also able to reduce the produced false positives by at least 5% when emojis were utilized. As a result, DistilBERT had a significantly higher F1-score of 93.8% and an accuracy of 94.1%, while ALBERT had a higher F1-score of 93% and an accuracy of 93.9%. These results are summarized in the Table 4.

Models	**Accuracy**	**F1-score**	**Precision**	**Recall**
BERTweet	**95.2%**	**94.9%**	**98.8%**	**91.2%**
ALBERT	93.9%	93.07%	97.3%	89.2%
DistilBERT	94.1%	93.8%	97.5%	90.4%

Tableau 4. – Average F1-score, Accuracy, Precision, and Recall (in%) for the Transformer-based classification models on the test dataset after including emojis.

Finally, we explored the collected tweets to determine if we could identify certain patterns. After geolocating the tweets, we found that they originated from 46 countries all over the world. The US, UK, Canada, and Australia had the highest number of potential-related Lyme disease tweets and non-Lyme disease tweets, accounting for approximately 97% of the total. Remarkably, there were observed spikes in both Lyme and non-Lyme tweet counts for the US, as the US is a hotspot country for Lyme disease. Half (50%) of potentially Lyme-related tweets were from the US, while approximately 0.2% were reported from Canada, 0.03% from Mexico, and 0.01% from some Caribbean countries, Haiti, and Jamaica. A total of 0.03% of the potential-related Lyme disease cases are reported in 4 countries in South America including Argentina, and Venezuela. Potential Lyme disease cases reported in Europe were from Belgium, Denmark, Estonia, France, Ireland, Luxembourg, Norway, Poland, Sweden, Switzerland, and represented 0.7% of the tweets, with 0.3% coming from the United Kingdom (UK). Potential Lyme disease cases reported in Asia were from in Indonesia, Iran, the Philippines, South Korea, Taiwan, Thailand, and Vietnam and represented a total of 0.07% of the dataset. Potential Lyme disease cases from Africa represented 0.005% and came from a single country, Sudan. Finally, New Zealand and Australia reported 0.12% of the total potential Lyme disease cases on Twitter, each accounting for 0.1% and 0.02%, respectively.

Figure 3 illustrates the number of medical symptoms of Lyme disease reported in tweets. Rash, fatigue, tick bite, fever, and arthritis were the most commonly reported symptoms. In contrast,

symptoms such as neck stiffness, numbness, and lymph nodes are rarely, reported symptoms. All the classified data are available at this link [4].

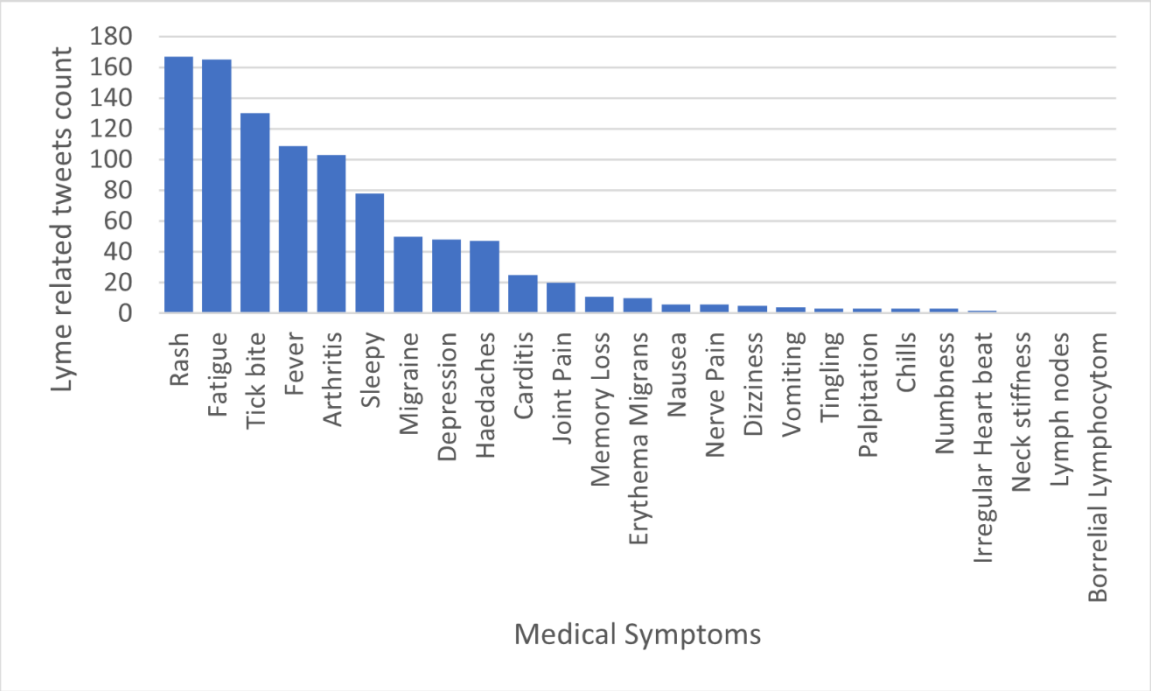

Figure 3 Top medical symptoms of Lyme disease reported in the tweets.

Discussion and Conclusion

The empirical results highlighted the improved performance of the transformer-based classifiers (*i.e.*, BERTweet, DistilBERT, and ALBERT), which we attributed to the following factors: (1) The tweet word embeddings produced by their associated tokenizers were more accurate than those generated by context-independent embedding techniques such as TF-IDF. This is because transformer-based classifiers are based on language models and can better understand

[4] https://github.com/Mohamed-Hamza-Ibrahim/Twitter_Lyme_dataset

the semantic content of short texts such as tweets in different contexts. Unlike TF-IDF, the tokenizers also consider the position and order of words in tweets, which improves their ability to understand the different meanings of individual words. (2) Unlike Adaboost, KNN and RF, transformer-based classifiers are less affected by noise and redundant words in tweets. (3) Transformer-based classifiers can learn nonlinear relationships and complex patterns in tweets because, unlike LR and SVM, their neural network architecture does not assume linearity between the dependent and independent features, which is not often the case in the extracted tweets. (4) Finally, transformer-based classifiers differ from both MLP and KNN in that they can efficiently handle feature scaling and frequently converge to the global optimum rather than getting stuck in local minima. This is due to optimizing the cross-entropy loss function, which is often convex for most weights.

The results also showed that emojis are effective as enrichment features to improve the accuracy of tweet embedding, and the performance of transformer-based classifier can be further improved by considering the sentimental semantics of emoji. Since the texts of non-Lyme and Lyme disease tweets can be similar in some cases, the emojis can play an important role in better identifying Lyme disease-related tweets. By removing these emojis, some potentially Lyme-disease related tweets could be misidentified as non-Lyme ones, resulting in more false negatives. This implies that the inclusion of sentimental or emotional words representing sadness, empathy, and encouragement emojis could significantly assist transformer-based classifiers in distinguishing potential Lyme disease related tweets from non-Lyme disease tweets.

The classification of 20,000 tweets showed a high volume of potentially Lyme-disease related tweets in the US, UK, and Canada. This may be due to two factors: (1) Lyme disease is slowly spreading (2) focusing solely on English tweets may limit the collection of tweets from non-

English speaking countries. Furthermore, borrelial lymphocytoma, palpitations, tingling, nausea, and neck stiffness are rarely reported symptoms. In fact, there are differences in the clinical manifestations seen in North America and in European countries. For example, Lyme arthritis and carditis are mainly found in North America while borrelial lymphocytoma and neurological symptoms (neck stiffness, numbness, etc.) are found in European countries. Erythema migrans, which is the most common symptoms, is not among the most common symptoms on Twitter because the general population tends to refer to this symptom as a rash. These findings correlate with the geographic distribution of Lyme disease clinical manifestations throughout the literature [2, 34, 31].

Our study is the first to provide a pre-trained, organized, and labeled Lyme related dataset with an emoji component, which will be able to quantify and compare the performance of different methodological approaches in future work. This will allow for consistency in future research and improve digital surveillance of Lyme disease. For example, a sudden increase in activity on Twitter or other social media platforms may indicate the beginning of an increase in cases and therefore justify the promotion of tick bites prevention measures in the indicated area. The methodology used in this study can be extrapolated to other social media and the performance of classification models can be evaluated.

Our study has several limitations. First, the collection of Twitter data through API Twitter is suggested to have a selection bias since only 1% of the data is accessible and it is random, and therefore the data may not reflect on the reality about Lyme disease conversations on Twitter. However, in our study, we collected data over a long period of time, which may limit this bias. Also, social media data are highly susceptible to media coverage, so tweets about Lyme disease may be driven by media coverage rather than the disease incidence.

Although emoji have general meanings, their usage mainly depends on other factors such as cultural background, linguistic factors, and gender [62]. Since our study only focuses on the sentimental semantics of emoji, our model may have erroneously assigned different meanings to emoji than what the tweet author intended. Therefore, our results should be interpreted with some caution. However, since our model was trained with labeled tweets, we believe that the mislabeling of some emoji did not significantly affect the performance of our model.

Early detection of potential Lyme disease is essential to limit the spread of the disease and improve the efficiency of medical care. Given the growing importance of social media as a source of information about infected cases, platforms such as Twitter can provide simultaneous updates on the Lyme disease epidemic. This makes the use of such data as a prediction and surveillance tool for Lyme disease cases an important but underexplored challenge in the field of health informatics. In this work, we propose a Lyme-disease detection system that is primarily a transformer-based classifier that uses data from self-reported tweets to identify potentially infected cases. While Twitter was the focus of this work, the proposed system can be easily adapted to other social media platforms like Reddit. Although the incidence of Lyme disease in the African continent is very low, our model was able to collect some tweets related to Lyme disease there, which can be valuable for public health surveillance. We suggest that future work focus on collecting social media data from both English and non-English texts to improve knowledge of potential Lyme disease cases, as some of the countries with the highest incidence of Lyme disease are non-English speaking countries.

Conflicts of Interest

"None declared".

Funding:

Our work is partially funded by The *Bourse d'Intelligence Artificielle* from the ESP of *Université de Montréal*.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers.

BERTweet: Bidirectional Encoder Representations from Transformers for

English Tweets.

ALBERT: A Lite Bidirectional Encoder Representations from Transformers.

Regex: Regular expressions.

TF-IDF: Term frequency-inverse document frequency

CDC: Centers for Disease Control and Prevention

References

1. Acheampong, F.A., Nunoo-Mensah, H. and Chen, W.: Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review. *54*(8), 5789-5829 (2021).
2. Alkishe, A., Raghavan, R.K., Peterson, A.T., Likely Geographic Distributional Shifts among Medically Important Tick Species and Tick-Associated Diseases under Climate Change in North America: A Review. Insects 12, 225 (2021). https://doi.org/10.3390/insects12030225
3. Atkinson, S.F., Sarkar, S., Aviña, A., Schuermann, J.A., Williamson, P.. A determination of the spatial concordance between Lyme disease incidence and habitat probability of its primary

vector Ixodes scapularis (black-legged tick). Geospatial Health 9, 203–212 (2014). https://doi.org/10.4081/gh.2014.17

4. Barbour, A.G., Fish, D.. The biological and social phenomenon of Lyme disease. Science 260, 1610–1616 (1993). https://doi.org/10.1126/science.8503006
5. Bisanzio, D., Kraemer, M.U.G., Brewer, T., Brownstein, J.S., Reithinger, R., Geolocated Twitter social media data to describe the geographic spread of SARS-CoV-2. J. Travel Med. 27, taaa120 (2020). https://doi.org/10.1093/jtm/taaa120
6. Blanchard, L., Jones-Diette, J., Lorenc, T., Sutcliffe, K., Sowden, A., Thomas, J.,. Comparison of national surveillance systems for Lyme disease in humans in Europe and North America: a policy review. BMC Public Health 22, 1307 (2022). https://doi.org/10.1186/s12889-022-13669-w
7. Boligarla, S., Laison, E.K.E., Li, J.F., Mahadevan, R., Ng, A., Lin, Y., Thioub, M.Y., Huang, B., Ibrahim, M.H. and Nasri, B.,: Leveraging Machine Learning Approaches for Predicting Potential Lyme Disease Cases and Incidence Rates in United States Using Twitter. Preprint on research square, 1-20 (2022)
8. Bouchard, C., Beauchamp, G., Nguon, S., Trudel, L., Milord, F., Lindsay, L.R., Bélanger, D., Ogden, N.H. Associations between Ixodes scapularis ticks and small mammal hosts in a newly endemic zone in southeastern Canada: implications for Borrelia burgdorferi transmission. Ticks Tick-Borne Dis. 2, 183–190 (2011). https://doi.org/10.1016/j.ttbdis.2011.03.005
9. Burrows, H., Slatculescu, A.M., Feng, C.X., Clow, K.M., Guillot, C., Jardine, C.M., Leighton, P.A., Krause, P.J., Kulkarni, M.A.. The utility of a maximum entropy species distribution model for Ixodes scapularis in predicting the public health risk of Lyme disease in Ontario,

Canada. Ticks Tick-Borne Dis. 13, 101969 (2022). https://doi.org/10.1016/j.ttbdis.2022.101969

10. Cardenas-de la Garza, J.A., De la Cruz-Valadez, E., Ocampo-Candiani, J., Welsh, O.. Clinical spectrum of Lyme disease. Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol. 38, 201–208 (2019). https://doi.org/10.1007/s10096-018-3417-1
11. CDC. Lyme disease. Data and Surveillance. [WWW Document]. Cent. Dis. Control Prev. URL https://www.cdc.gov/Lyme/datasurveillance/index.html (accessed 6.22.22) (2021).
12. Chi, W., Chen, Y., Su, C., Shi, X., Global reoccurrence measure for keyword extraction, in: 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Presented at the 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 587–592 (2014). https://doi.org/10.1109/FSKD.2014.6980900
13. Coors, A., Hassenstein, M.J., Krause, G., Kerrinnes, T., Harries, M., Breteler, M.M.B., Castell, S.,. Regional seropositivity for Borrelia burgdorferi and associated risk factors: findings from the Rhineland Study, Germany. Parasit. Vectors 15, 241 (2022). https://doi.org/10.1186/s13071-022-05354-z
14. Cortes, C. and Vapnik, V.: Support-vector networks. Machine learning. *20*(3), 273-297 (1995).
15. Cover, T. and Hart, P.: Nearest neighbor pattern classification. IEEE transactions on information theory. *13*(1), 21-27 (1967).
16. Dehnert, M., Fingerle, V., Klier, C., Talaska, T., Schlaud, M., Krause, G., Wilking, H., Poggensee, G., Seropositivity of Lyme Borreliosis and Associated Risk Factors: A Population-Based Study in Children and Adolescents in Germany (KiGGS). PLOS ONE 7, e41321 (2012). https://doi.org/10.1371/journal.pone.0041321

17. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

18. Dias Canedo, E., Cordeiro Mendes, B., Software Requirements Classification Using Machine Learning Algorithms. Entropy Basel Switz. 22, 1057 (2020). https://doi.org/10.3390/e22091057

19. Doan, S., Yang, E.W., Tilak, S.S., Li, P.W., Zisook, D.S., Torii, M.,. Extracting health-related causality from twitter messages using natural language processing. BMC Med. Inform. Decis. Mak. 19, 79 (2019). https://doi.org/10.1186/s12911-019-0785-0

20. Donta, S.T. What We Know and Don't Know About Lyme Disease. Front. Public Health 9, 819541 (2022.). https://doi.org/10.3389/fpubh.2021.819541

21. Edo-Osagie, O., Smith, G., Lake, I., Edeghere, O., De La Iglesia, B.: Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. PloS One 14, e0210689 (2019). https://doi.org/10.1371/journal.pone.0210689

22. Eisen, R.J., Piesman, J., Zielinski-Gutierrez, E., Eisen, L. What Do We Need to Know About Disease Ecology to Prevent Lyme Disease in the Northeastern United States? J. Med. Entomol. 49, 11–22 (2012). https://doi.org/10.1603/ME11138

23. Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M. and Riedel, S.: emoji2vec: Learning emoji representations from their description, in: Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics,

Austin, TX, USA, 48–54. doi:10.18653/v1/W16-6208. URL https://aclanthology.org/W16-6208 (2016).

24. Freund Y., Schapire R. E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55 (1), 119–139 (1997).

25. Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction. Vol. 2, pp. 1-758. New York: springer (2009).

26. Jánová, E. Emerging and threatening vector-borne zoonoses in the world and in Europe: a brief update. Pathog. Glob. Health 113, 49–57. https://doi.org/10.1080/20477724.2019.1598127 (2019)

27. Ji, Z., Jian, M., Yue, P., Cao, W., Xu, X., Zhang, Y., Pan, Y., Yang, J., Chen, J., Liu, M., Fan, Y., Su, X., Wen, S., Kong, J., Li, B., Dong, Y., Zhou, G., Liu, A., Bao, F.. Prevalence of Borrelia burgdorferi in Ixodidae Tick around Asia: A Systematic Review and Meta-Analysis. Pathog. Basel Switz. 11, 143 (2022). https://doi.org/10.3390/pathogens11020143

28. Kaplan, A.M., Haenlein, M.,. Users of the world, unite! The challenges and opportunities of Social Media. Bus. Horiz. 53, 59–68 (2010). https://doi.org/10.1016/j.bushor.2009.09.003

29. Kotsiantis, S.B., Zaharakis, I. and Pintelas, P.: Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24 (2007).

30. Krishnapuram B., Carin L., Figueiredo M. A., Hartemink A. J.: Sparse multinomial logistic regression: Fast algorithms and generalization bounds, IEEE transactions on pattern analysis and machine intelligence. 27 (6), 957–968 (2005).

31. Kugeler, K.J., Schwartz, A.M., Delorey, M.J., Mead, P.S., Hinckley, A.F.,. Estimating the Frequency of Lyme Disease Diagnoses, United States, 2010-2018. Emerg. Infect. Dis. 27, 616–619 (2021). https://doi.org/10.3201/eid2702.202731

32. Kullberg, B.J., Vrijmoeth, H.D., van de Schoor, F., Hovius, J.W.,. Lyme borreliosis: diagnosis and management. BMJ 369, m1041 (2020). https://doi.org/10.1136/bmj.m1041

33. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. International Conference on Learning Representations. p. 1-17 (2020).
34. Nelson, C.A., Saha, S., Kugeler, K.J., Delorey, M.J., Shankar, M.B., Hinckley, A.F., Mead, P.S., Incidence of Clinician-Diagnosed Lyme Disease, United States, 2005-2010. Emerg. Infect. Dis. 21, 1625–1631 (2015). https://doi.org/10.3201/eid2109.150417
35. Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for English tweets. arXivpreprint arXiv:2005.10200 (2020).
36. Ogden, N.H., Bouchard, C., Kurtenbach, K., Margos, G., Lindsay, L.R., Trudel, L., Nguon, S., Milord, F.. Active and passive surveillance and phylogenetic analysis of Borrelia burgdorferi elucidate the process of Lyme disease risk emergence in Canada. Environ. Health Perspect. 118, 909–914 (2010). https://doi.org/10.1289/ehp.0901766
37. Pace, E.J., O'Reilly, M.,. Tickborne Diseases: Diagnosis and Management. Am. Fam. Physician 101, 530–540 (2020).
38. Pal M.: Random forest classifier for remote sensing classification. International journal of remote sensing 26 (1), 217–222 (2005).
39. Petrulionienė, A., Radzišauskienė, D., Ambrozaitis, A., Čaplinskas, S., Paulauskas, A., Venalis, A.,. Epidemiology of Lyme Disease in a Highly Endemic European Zone. Med. Kaunas Lith. 56, 115 (2020). https://doi.org/10.3390/medicina56030115
40. Powers, D. M.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol. 2(1), 37–63 (2011).
41. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, 242, 29–48 (2003).

42. Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019).

43. Schotthoefer, A., Stinebaugh, K., Martin, M., Munoz-Zanzi, C.,. Tickborne disease awareness and protective practices among U.S. Forest Service employees from the upper Midwest, USA. BMC Public Health 20, 1575 (2020). https://doi.org/10.1186/s12889-020-09629-x

44. Schwartz, A.M., Kugeler, K.J., Nelson, C.A., Marx, G.E., Hinckley, A.F.,. Use of Commercial Claims Data for Evaluating Trends in Lyme Disease Diagnoses, United States, 2010-2018. Emerg. Infect. Dis. 27, 499–507 (2021). https://doi.org/10.3201/eid2702.202728

45. Stanek, Gerold et al. "Lyme borreliosis." *Lancet (London, England)* vol. 379,9814 (2012): 461-73. doi:10.1016/S0140-6736(11)60103-7

46. Steinbrink, A., Brugger, K., Margos, G., Kraiczy, P., Klimpel, S.. The evolving story of Borrelia burgdorferi sensu lato transmission in Europe. Parasitol. Res. 121, 781–803 (2022). https://doi.org/10.1007/s00436-022-07445-3

47. Stone, B.L., Tourand, Y., Brissette, C.A.. Brave New Worlds: The Expanding Universe of Lyme Disease. Vector Borne Zoonotic Dis. Larchmt. N 17, 619–629 (2017). https://doi.org/10.1089/vbz.2017.2127

48. Tharwat, A.: Linear vs. quadratic discriminant analysis classifier: a tutorial. International Journal of Applied Pattern Recognition. *3*(2), 145-180 (2016).

49. Thurston, R.. Lyme disease. Work Read. Mass 62, 643–646 (2019). https://doi.org/10.3233/WOR-192897

50. Wijngaard, C.C. van den, Hofhuis, A., Simões, M., Rood, E., Pelt, W. van, Zeller, H., Bortel, W.V., 2017. Surveillance perspective on Lyme borreliosis across the European Union and

European Economic Area. Eurosurveillance 22, 30569 (2022). https://doi.org/10.2807/1560 7917.ES.2017.22.27.30569

51. Allen, C., Tsou, M.-H., Aslam, A., Nagel, A., Gawron, J.-M.: Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. PLOS ONE 11(7), 0157734 (2016). doi:10.1371/journal.pone.0157734
52. Basch, C.H., Mullican, L.A., Boone, K.D., Yin, J., Berdnik, A., Eremeeva, M.E., Fung, I.C.-H.: Lyme Disease and YouTubeTM: A Cross-Sectional Study of Video Contents. Osong Public Health and Research Perspectives 8(4), 289–292 (2017). doi:10.24171/j.phrp.2017.8.4.10
53. Kapitány-Fövény, M., Ferenci, T., Sulyok, Z., Kegele, J., Richter, H., Vályi-Nagy, I., Sulyok, M.: Can Google Trends data improve forecasting of Lyme disease incidence? Zoonoses and Public Health 66(1), 101–107 (2019). doi:10.1111/zph.12539
54. Kutera, M., Berke, O., Sobkowich, K.: Spatial epidemiological analysis of Lyme disease in southern Ontario utilizing Google Trends searches. Environmental Health Review 64(4), 105–110 (2021). doi:10.5864/d2021-025.
55. Seifter, A., Schwarzwalder, A., Geis, K., Aucott, J.: The utility of "Google Trends" for epidemiological research: Lyme disease as an example. Geospatial health 4(2), 135 (2010). doi:10.4081/gh.2010.195.
56. Kim, D., Maxwell, S., Le, Q.: Spatial and Temporal Comparison of Perceived Risks and Confirmed Cases of Lyme Disease: An Exploratory Study of Google Trends. Frontiers in Public Health 8, 395 (2020). doi:10.3389/fpubh.2020.00395.
57. Tulloch, J.S.P., Vivancos, R., Christley, R.M., Radford, A.D., Warner, J.C.: Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and

Republic of Ireland. Journal of Biomedical Informatics 100, 100060 (2019). doi:10.1016/j.yjbinx.2019.100060.

58. Takats, C., Kwan, A., Wormer, R., Goldman, D., Jones, H. E., & Romero, D. : Ethical and Methodological Considerations of Twitter Data for Public Health Research: Systematic Review. *Journal of medical Internet research*, *24*(11), e40380 (2022). doi.org/10.2196/40380
59. Babajide O. Ayinde et Jacek M. Zurada, « Deep Learning of Constrained Autoencoders for Enhanced Understanding of Data », *IEEE Transactions on Neural Networks and Learning Systems* 29, nº 9 (septembre 2018): 3969-79, https://doi.org/10.1109/TNNLS.2017.2747861.
60. Mohr, H., & Ruge, H. (2021). Fast Estimation of L1-Regularized Linear Models in the Mass-Univariate Setting. *Neuroinformatics*, *19* nº 3 (2021) : 385–392. https://doi.org/10.1007/s12021-020-09489-1
61. Alvin Subakti, Hendri Murfi, et Nora Hariadi, « The Performance of BERT as Data Representation of Text Clustering », *Journal of Big Data* 9, nº 1 (2022): 15, https://doi.org/10.1186/s40537-022-00564-9.
62. Bai, Qiyu, Qi Dan, Zhe Mu, et Maokun Yang. « A Systematic Review of Emoji: Current Research and Future Perspectives ». *Frontiers in Psychology* 10: 2221. (2019). https://doi.org/10.3389/fpsyg.2019.02221.

Chapitre 6- [Discussion et conclusions]

Les objectifs de ce mémoire étaient de développer une base de données de tweets classifiés en lien avec la maladie de Lyme et ensuite d'analyser la performances des modèles de classification basés sur les transformateurs BERT afin de prédire les cas potentiels de la maladie de Lyme en évaluant l'incorporation des émojis comme outil d'enrichissement pour améliorer la performance des modèles en utilisant les données collectées sur Twitter. Le but ultérieur de notre mémoire est de proposer un modèle de classification pré-entraîné avec une large base de données provenant de Twitter capable de détecter des cas potentiels de la maladie de Lyme avec des données non structurées comme celles des réseaux sociaux. Nous avons basé notre modèle sur les classificateurs BERT parce qu'ils performent mieux avec des données textuelles avec des informations peu pertinentes comme les tweets (Klein et al., 2021).

6.1 Interprétation des résultats observés :

Les résultats de notre étude démontrent que, dans l'ensemble, les modèles de classification basés les transformateurs BERT (ALBERT, BERTweet et DistilBERT) ont mieux performé que les modèles de classification classiques évalués dans cette étude ce qui n'est pas surprenant.

En effet les classificateurs BERT ont su démontrer leur performance à classifier avec une précision et une exactitude élevées des données non structurées comme celles provenant des réseaux sociaux, contrairement aux modèles de classification classiques (Benítez-Andrades et al., 2022; Klein et al., 2021; J. Lin et al., 2021b). Cela découle en partie d'un mécanisme d'attention et d'un traitement non essentiel du texte d'entrée qui peut capturer des dépendances à longue portée (c'est-à-dire, reliant un fait dans la première phrase du premier paragraphe à la dernière phrase du dernier paragraphe ainsi que la deuxième phrase du premier paragraphe) (Tejani et al., 2022). Aussi, la

performance de BERTweet prouve l'affirmation selon laquelle les modèles BERT spécifiques au données de Twitter surpassent les modèles génériques de BERT (Benítez-Andrades et al., 2022; Nguyen et al., 2020).

Aussi, l'intégration des émojis convertis en mots semble être un bel apport à notre étude puisque la performance des modèles BERT a augmenté dans l'ensemble bien que le modèle BERTweet a mieux performé. Ces résultats concordent avec d'autres études qui ont démontré que l'intégration des sentiments des émojis avec les données provenant des réseaux sociaux améliore considérablement la précision des modèles de classification (Bai et al., 2019; Liu et al., 2021; Lu et al., 2018; Narr et al., 2012; Shibayama et al., 2021). Par exemple, Bai et collaborateurs ont montré que l'intégration des émojis comme des fonctionnalités d'enrichissement améliorent la précision des tweets et donc la performance des modèles de classificateurs. Leur intégration prend en compte la sémantique sentimentale qui n'est pas détectée avec le texte des tweets uniquement car il peut ne pas inclure les émotions exprimées par les utilisateurs (Bai et al., 2019). Dans certains cas, les tweets liés à la maladie et ceux qui ne le sont pas peuvent partager plusieurs similitudes et l'intégration des émojis peuvent alors aider les modèles de classification à mieux les classifier (Shibayama et al., 2021).

Un point important aussi dans nos résultats est que les tweets en anglais collectés proviennent majoritairement des États-Unis et inclinées vers le monde anglophone. Ceci a été aussi observé dans d'autres études scientifiques (Takats et al., 2022). En effet, dans les tweets en anglais collectés dans notre étude, 50% des tweets potentiellement liés à la maladie de Lyme proviennent des États-Unis. Cette prépondérance des tweets aux États-Unis n'est pas surprenante, puisque la majorité des utilisateurs de Twitter viennent des États-Unis (Statista, 2022). Cependant, les États-Unis étant le pays avec les plus de cas de la maladie de Lyme, une telle observation est attendue (Kugeler et al.,

2022). Toutefois, il est étonnant de voir dans notre analyse que l'Australie et la Nouvelle-Zélande sont à la quatrième place pour les pays qui produisent le plus de tweets reliés à la maladie de Lyme derrière les États-Unis, le Royaume-Uni et le Canada puisque l'incidence de la maladie de Lyme dans cette partie du monde est incertaine et probablement faible (Dehhaghi et al., 2019). Nous pensons que cela est dû à une sensibilisation de la population australienne à la maladie de Lyme et que le fardeau de la maladie dans la région serait plus élevé que ce qui est capté par les systèmes de surveillance. Très peu de tweets classifiés comme cas potentiels de la maladie de Lyme proviennent de l'Europe si on exclut le Royaume-Uni. Les cas potentiels de maladie de Lyme signalés en Europe provenaient de Belgique, du Danemark, d'Estonie, de France, d'Irlande, du Luxembourg, de Norvège, de Pologne, de Suède et de Suisse, et représentaient 0,7 % des tweets. En effet excepté en Amérique du Nord, la majorité des pays endémiques à la maladie de Lyme en Europe et dans les autres parties du monde ne sont pas des pays anglophones d'où la possibilité d'un biais linguistique. Très peu tweets provenant de l'Afrique ont été détectés, sûrement parce que la maladie de Lyme ne présente que des cas sporadiques dans la région, aussi parce plusieurs pays africains n'ont pas une couverture internet aussi ample que dans les pays occidentaux (Nyataya et al., 2020; Yaya & Ghose, 2018).

On remarque en outre que les termes les plus associés à la maladie de Lyme sont des termes non spécifiques tels que « fièvre », « fatigue », « piqûre de tiques », « arthrite » et « dépression » et correspondent aux symptômes les plus fréquemment reportés, tandis que les termes comme « lymphocytome », « érythème migrant », caractéristiques de la maladie de Lyme, sont très peu reportés. Cela est cohérent avec les études antérieures ayant exploité les données provenant du web tels que Google trends, Twitter et autres dans le cadre de la surveillance digitale de la maladie de Lyme, ayant également reporté que les termes généraux tels que « fièvre », « piqûre de tiques » ou

simplement le nom de la maladie « Lyme » étaient les plus corrélés avec les tendances spatio-temporelles de la maladie de Lyme (D. Kim et al., 2020b; Kutera et al., 2021; Pesälä et al., 2017; Seifter et al., 2010; Tulloch et al., 2019a). En effet Seifter et collaborateurs, ont démontré que les mots clés spécifiques comme « tick bite (piqûre de tiques) » sont moins fréquents que le nom de la maladie ou d'un symptôme général comme « cold (rhume) » (Seifter et al., 2010). Une autre étude en Allemagne qui a exploité des données de Google trends sur la maladie de Lyme, a démontré que le nom général de la maladie « *borréliose* » était le mot-clé le plus fréquent (Kapitány-Fövény et al., 2019). Tulloch a suggéré de ne pas inclure les symptômes tels que « les éruptions cutanées » comme termes de requête (mots-clés) car ils entraîneraient de nombreux tweets faussement positifs (Tulloch et al., 2019b).

6.2 Apports de notre étude à la santé Publique :

Malgré le grand potentiel de l'utilisation des données Twitter pour la santé publique identifié à travers la littérature, il existe très peu d'études sur son implantation dans le cadre de la maladie de Lyme pourtant émergente. De plus, il n'existe jusqu'à date aucune base de données qui pourra être activement testée et évaluée dans un contexte opérationnel pour la pratique en santé publique.

Nos résultats contribuent à l'augmentation des preuves suggérant que les données des médias sociaux sont utiles pour la surveillance des cas de Lyme partout dans le monde. De plus, notre approche a plusieurs implications importantes en termes de méthodologie pour la surveillance de la maladie de Lyme à l'aide des données des médias sociaux.

Notre étude est la première à développer un algorithme de base de données pré-entraînées sur une liste de mots-clés basés sur les symptômes de la maladie de Lyme. Considérant la grande

variabilité des symptômes de la maladie de Lyme, et aussi parce que la plupart de ces symptômes ne sont pas spécifiques, les études qui ont utilisé les données provenant des réseaux sociaux comme outil de surveillance, ont préférablement basé leur modèle de prédiction sur des mots-clés généraux tels que « Lyme », « tick bite » pour limiter le nombre de tweets faussement positifs (Kim et al., 2020a; Kutera et al., 2021; Scheerer et al., 2020; Seifter et al., 2010; Tulloch et al., 2019). Notre étude quant à elle a pu démontrer une corrélation entre la fréquence des mots-clés basés sur les symptômes et la distribution géographique des manifestations cliniques de la maladie de Lyme. Ceci souligne le potentiel de Twitter à cartographier la répartition géographique de cette maladie et qu'il pourrait être utilisé pour identifier les zones à risque de la maladie. Cette découverte va dans le même sens que les résultats de Tulloch et collaborateurs qui ont déterminé une similitude de la distribution spatiale des tweets et les cas de Lyme (Tulloch et al., 2019). Ceci valide l'exploitation des données de twitter comme outil complémentaire de surveillance des cas de Lyme par les autorités de Santé Publique.

Les modèles de classification utilisés dans notre étude nous permettent de classer non seulement les tweets associés à la maladie de Lyme, suggérant un potentiel cas de la maladie de Lyme, mais aussi des faux positifs, c'est-à-dire des cas qui ont été classifiés comme étant des cas potentiels mais ne le sont pas en réalité. Donc l'augmentation soudaine de l'activité sur Twitter peut signifier le début d'une hausse des cas et peut donc justifier la promotion des mesures de protection contre les piqûres de tiques dans la zone indiquée.

L'ensemble de données organisé, validé et étiqueté fourni dans cette étude permettra aux chercheurs et aux praticiens de santé publique de mieux exploiter les différents aspects de la surveillance digitale de la maladie de Lyme en classifiant de nouvelles observations (tweets) et détectant les zones à risque ou en émergence de la maladie.

Notre méthodologie et nos textes classifiés peuvent aussi servir à explorer des données issues d'autres réseaux sociaux, les utilisateurs sur ces autres réseaux auront potentiellement la même façon de converser sur la maladie de Lyme sur ces plateformes. Toutefois cette hypothèse est à considérer avec réserve parce que les réseaux sociaux ont des particularités quant à leur conversation et l'accès à ces discussions, et le nombre de caractères dans les textes.

L'intégration des emojis convertis en mots dans les modèles de classificateurs basés sur BERT dans cette étude est une innovation dans la surveillance digitale de la maladie de Lyme. Tout d'abord il s'agit de la première étude à considérer la variable « sentiment » à l'aide des émojis dans l'élaboration d'une base de données sur les tweets reliés à la maladie de Lyme, mais aussi son intégration améliore la performance des modèles de classification. Une telle richesse dans la base de données proposée constitue une ressource importante pour les autorités de santé publique quant à l'élaboration des interventions basées sur les modèles de classification parce que ces modèles prendront en compte les émotions des utilisateurs à qui s'adressent ces interventions et permettront aux autorités de santé publique d'évaluer la perception de ces interventions.

Les modèles supervisés qui pourront tirer parti des ensembles de formation validés sont susceptibles d'avoir une performance beaucoup plus élevée en termes des mesures de performance évalués (rappel, précision, score F1 moyen et exactitude) ce qui permettra d'atteindre une classification plus réaliste, et cet apport va dans la même direction que les résultats de Mackey et collaborateurs (Mackey et al., 2020). L'accès à cette base de données permettra aux futurs chercheurs de gagner du temps puisque l'étiquetage des données provenant des réseaux sociaux pourra être réutilisé.

Un autre avantage de l'usage de cette base de données est l'exploration rapide des discussions en ligne sur la maladie de Lyme avant d'implémenter des interventions de santé publique concernant la maladie de Lyme.

Selon une étude sur les données de twitter, les informations sur twitter ne concernent pas seulement les utilisateurs mais aussi les familles et leurs amis donc les tweets peuvent être représentatifs d'un grand segment de la population (Mackey et al., 2020).

6.3 Limites de notre étude :

La première lacune de notre étude est le filtrage par mots-clés puisque tous les tweets ne contenant pas les mots clés inclus dans notre base de données ne seront pas considérés par notre système, laissant ainsi des cas potentiels qui ne seront pas reportés. Aussi le choix des mots-clés spécifiques tel que les symptômes de la maladie de Lyme, pourrait nous faire omettre du contenu pertinent (Y. Kim et al., 2016; Paul & Dredze, 2014; Sinnenberg et al., 2017).

La forte vulnérabilité aux événements médiatiques et l'absence d'approches traitant de cette question constituent la deuxième limite de cette étude.

Compte tenu du fait que la plupart des études qui ont porté sur la surveillance digitale des cas de Lyme sont jusqu'à présent géographiquement restreint à une partie spécifique du globe, une différence géographique des discussions sur Twitter n'a pas pu être établie. Une étude qui estimait la différence géographique des conversations sur Twitter a démontré que la fréquence des mots-clés dépend fondamentalement des régions et le vocabulaire le plus utilisé dans le milieu (van Draanen et al., 2020). Dans notre étude, la collecte de nos données est restreinte seulement à la langue anglaise mais notre méthodologie n'a pas prise en compte la variabilité linguistique des

différentes régions étudiées. Donc le biais linguistique même si tous les tweets proviennent de milieux anglophones doit être pris en compte pour la généralisation des résultats de notre étude.

Aussi, notre analyse n'a pas prise en compte la variabilité des années incluses c'est-à-dire les années comprises dans la période allant de 2010 à 2022. Cela aurait pu apporter des informations sur les tendances annuelles parce que l'usage de twitter comme outil de discussions sur les enjeux de santé pourrait émerger différemment dans les pays en fonction de l'accès à l'internet et de la hausse des cas de la maladie de Lyme dans ces derniers.

La géolocalisation des tweets n'étant pas toujours présente, elle est à considérer avec réserve puisque certains tweets ne font toujours pas référence aux personnes ayant supposément contracté la maladie de Lyme sinon à une tiers-personne avec une géolocalisation différente. Aussi nous avons utilisé le package GeoPy pour extraire la géolocalisation des tweets à partir du profil d'utilisateur tout en gardant son identité privée. Toutefois les coordonnées géographiques sur ces profils d'utilisateurs ne sont pas toujours exactes pour différentes raisons comme par exemple, certaines personnes n'actualisent pas leurs coordonnées après un déménagement ou insèrent des coordonnées fausses pour des raisons de sécurité (Bisanzio et al., 2020). Bien que les emoji aient des significations générales, leur utilisation dépend principalement d'autres facteurs tels que le contexte culturel, les facteurs linguistiques et le sexe (Acheampong et al., 2021). Comme notre étude se concentre uniquement sur la sémantique sentimentale des emoji, notre modèle peut avoir attribué par erreur aux emoji des significations différentes de celles voulues par l'auteur du tweet. Par conséquent, nos résultats doivent être interprétés avec une certaine prudence. Cependant, comme notre modèle a été entraîné avec des tweets étiquetés, nous pensons que l'étiquetage erroné de certains emoji n'a pas affecté de manière significative les performances de notre modèle.

Aussi les discussions sur les réseaux sociaux, spécialement sur Twitter, est en grande partie une réflexion des sujets abordés dans les nouvelles. Par conséquent, les discussions sur la maladie de Lyme sur les réseaux sociaux peuvent être suscitées par la médiatisation et non par l'incidence de la maladie dans une région donnée, impliquant le risque que les données ne représentent pas l'état réel de la situation de santé (Lyu et al., 2021).

Finalement, le profil des utilisateurs de Twitter diffère de celui des cas de Lyme, pouvant entraîner une surreprésentation de certaines régions et une sous-représentation dans d'autres régions, ce qui entraîne un biais démographique dont il faut tenir compte dans l'interprétation des résultats (Jahanbin et al., 2019). En effet selon Statista, l'audience mondiale de Twitter est fortement regroupés entre 25-34 ans alors que la distribution par âges des cas de la maladie de Lyme est bimodale; chez les enfants de 5-15 ans et les adultes de 45-55 ans (Mead, 2015; Statista, 2022).

6.4 Conclusion :

Les résultats de la plupart des recherches révèlent le potentiel des données provenant des réseaux sociaux pour la surveillance en temps réel des maladies. Nous ne croyons pas que ces données puissent remplacer les systèmes de surveillance traditionnels, mais elles peuvent être un supplément d'information d'une grande importance, surtout pour la conception des systèmes d'alerte précoce. En effet, malgré le potentiel de Twitter comme outil de surveillance dans le cadre de la maladie de Lyme, cette nouvelle approche de surveillance doit être prise en compte avec précaution puisqu'un tweet sur la maladie de Lyme n'est toujours pas synonyme de cas incident de la maladie de Lyme. Donc il est prudent d'affirmer que l'usage des données provenant de Twitter en tant que seul système de surveillance dans le cadre de la maladie de Lyme est peu possible à

cause de toutes les limites et les difficultés d'analyse de ces données. Mais bien que ces données soient peu susceptibles d'être à elles seules un système de surveillance performant, elles aident à combler certaines lacunes du système de surveillance traditionnel déjà en place.

De plus, notre étude apporte une contribution importante dans le domaine de l'exploitation des données de Twitter dans le contexte de la surveillance de la maladie de Lyme, puisqu'elle fournit une base de données prétraitées et classifiées à l'aide des modèles de classification capable de détecter un cas potentiel de Lyme. Notre approche toutefois est basée uniquement sur les tweets en anglais, nous comptons dans le futur inclure d'autres langues et des données d'autres réseaux sociaux pour évaluer la performance des modèles BERT.

Références bibliographiques

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection : A review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789-5829. https://doi.org/10.1007/s10462-021-09958-2

Adebisi, T., Aregbesola, A., Asamu, F., Arisukwu, O., & Oyeyipo, E. (2021). Using SNSs for early detection of disease outbreak in developing countries : Evidence from COVID-19 pandemic in Nigeria. *Heliyon*, *7*(6), e07184. https://doi.org/10.1016/j.heliyon.2021.e07184

Aenishaenslin, C., Bouchard, C., Koffi, J. K., & Ogden, N. H. (2017). Exposure and preventive behaviours toward ticks and Lyme disease in Canada : Results from a first national survey. *Ticks and Tick-Borne Diseases*, *8*(1), 112-118. https://doi.org/10.1016/j.ttbdis.2016.10.006

Aguero-Rosenfeld, M. E., Wang, G., Schwartz, I., & Wormser, G. P. (2005). Diagnosis of Lyme borreliosis. *Clinical Microbiology Reviews*, *18*(3), 484-509. https://doi.org/10.1128/CMR.18.3.484-509.2005

Aguilar-Gallegos, N., Romero-García, L. E., Martínez-González, E. G., García-Sánchez, E. I., & Aguilar-Ávila, J. (2020). Dataset on dynamics of Coronavirus on Twitter. *Data in Brief*, *30*, 105684. https://doi.org/10.1016/j.dib.2020.105684

Alkishe, A., Raghavan, R. K., & Peterson, A. T. (2021). Likely Geographic Distributional Shifts among Medically Important Tick Species and Tick-Associated Diseases under Climate Change in North America : A Review. *Insects*, *12*(3), 225. https://doi.org/10.3390/insects12030225

Allen, C., Tsou, M.-H., Aslam, A., Nagel, A., & Gawron, J.-M. (2016). Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. *PLOS ONE*, *11*(7), e0157734. https://doi.org/10.1371/journal.pone.0157734

Al-Rawi, A., Siddiqi, M., Morgan, R., Vandan, N., Smith, J., & Wenham, C. (2020). COVID-19 and the Gendered Use of Emojis on Twitter : Infodemiology Study. *Journal of Medical Internet Research*, *22*(11), e21646. https://doi.org/10.2196/21646

Arias, M., Arratia, A., & Xuriguera, R. (2014). Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology*, *5*(1), 8:1-8:24. https://doi.org/10.1145/2542182.2542190

Atkinson, S. F., Sarkar, S., Aviña, A., Schuermann, J. A., & Williamson, P. (2014). A determination of the spatial concordance between Lyme disease incidence and habitat probability of its primary vector Ixodes scapularis (black-legged tick). *Geospatial Health*, *9*(1), 203-212. https://doi.org/10.4081/gh.2014.17

Aucott, J. N. (2015). Posttreatment Lyme disease syndrome. *Infectious Disease Clinics of North America*, *29*(2), 309-323. https://doi.org/10.1016/j.idc.2015.02.012

Bacon, R. M., Kugeler, K. J., Mead, P. S., & Centers for Disease Control and Prevention (CDC). (2008). Surveillance for Lyme disease—United States, 1992-2006. *Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002)*, *57*(10), 1-9.

Bai, Q., Dan, Q., Mu, Z., & Yang, M. (2019). A Systematic Review of Emoji : Current Research and Future Perspectives. *Frontiers in Psychology*, *10*, 2221. https://doi.org/10.3389/fpsyg.2019.02221

Basch, C. H., Mullican, L. A., Boone, K. D., Yin, J., Berdnik, A., Eremeeva, M. E., & Fung, I. C.-H. (2017). Lyme Disease and YouTube TM : A Cross-Sectional Study of Video Contents. *Osong*

Public Health and Research Perspectives, *8*(4), 289-292. https://doi.org/10.24171/j.phrp.2017.8.4.10

Batheja, S., Nields, J. A., Landa, A., & Fallon, B. A. (2013). Post-treatment Lyme syndrome and central sensitization. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *25*(3), 176-186. https://doi.org/10.1176/appi.neuropsych.12090223

Beck, A. R., Marx, G. E., & Hinckley, A. F. (2021). Diagnosis, Treatment, and Prevention Practices for Lyme Disease by Clinicians, United States, 2013-2015. *Public Health Reports (Washington, D.C.: 1974)*, *136*(5), 609-617. https://doi.org/10.1177/0033354920973235

Beckmann, S., Freund, R., Pehl, H., Rodgers, A., & Venegas, T. (2019). Rodent species as possible reservoirs of Borrelia burgdorferi in a prairie ecosystem. *Ticks and Tick-Borne Diseases*, *10*(5), 1162-1167. https://doi.org/10.1016/j.ttbdis.2019.06.011

Belongia, E. A. (2002). Epidemiology and impact of coinfections acquired from Ixodes ticks. *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)*, *2*(4), 265-273. https://doi.org/10.1089/153036602321653851

Bisanzio, D., Fernández, M. P., Martello, E., Reithinger, R., & Diuk-Wasser, M. A. (2020). Current and Future Spatiotemporal Patterns of Lyme Disease Reporting in the Northeastern United States. *JAMA Network Open*, *3*(3), e200319-e200319. https://doi.org/10.1001/jamanetworkopen.2020.0319

Bisanzio, D., Kraemer, M. U. G., Brewer, T., Brownstein, J. S., & Reithinger, R. (2020). Geolocated Twitter social media data to describe the geographic spread of SARS-CoV-2. *Journal of Travel Medicine*, *27*(5), taaa120. https://doi.org/10.1093/jtm/taaa120

Blanchard, L., Jones-Diette, J., Lorenc, T., Sutcliffe, K., Sowden, A., & Thomas, J. (2022a). Comparison of national surveillance systems for Lyme disease in humans in Europe and North America : A policy review. *BMC Public Health*, *22*(1), 1307. https://doi.org/10.1186/s12889-022-13669-w

Blanchard, L., Jones-Diette, J., Lorenc, T., Sutcliffe, K., Sowden, A., & Thomas, J. (2022b). Comparison of national surveillance systems for Lyme disease in humans in Europe and North America : A policy review. *BMC Public Health*, *22*(1), 1307. https://doi.org/10.1186/s12889-022-13669-w

Bobe, J. R., Jutras, B. L., Horn, E. J., Embers, M. E., Bailey, A., Moritz, R. L., Zhang, Y., Soloski, M. J., Ostfeld, R. S., Marconi, R. T., Aucott, J., Ma'ayan, A., Keesing, F., Lewis, K., Ben Mamoun, C., Rebman, A. W., McClune, M. E., Breitschwerdt, E. B., Reddy, P. J., … Fallon, B. A. (2021). Recent Progress in Lyme Disease and Remaining Challenges. *Frontiers in Medicine*, *8*, 666554. https://doi.org/10.3389/fmed.2021.666554

Bockenstedt, L. K., & Wormser, G. P. (2014). Review : Unraveling Lyme disease. *Arthritis & Rheumatology (Hoboken, N.J.)*, *66*(9), 2313-2323. https://doi.org/10.1002/art.38756

Bonds, M. H., Dobson, A. P., & Keenan, D. C. (2012). Disease Ecology, Biodiversity, and the Latitudinal Gradient in Income. *PLOS Biology*, *10*(12), e1001456. https://doi.org/10.1371/journal.pbio.1001456

Borchers, A. T., Keen, C. L., Huntley, A. C., & Gershwin, M. E. (2015). Lyme disease : A rigorous review of diagnostic criteria and treatment. *Journal of Autoimmunity*, *57*, 82-115. https://doi.org/10.1016/j.jaut.2014.09.004

Bouchard, C., Beauchamp, G., Leighton, P. A., Lindsay, R., Bélanger, D., & Ogden, N. H. (2013). Does high biodiversity reduce the risk of Lyme disease invasion? *Parasites & Vectors*, *6*, 195. https://doi.org/10.1186/1756-3305-6-195

Bouchard, C., Beauchamp, G., Nguon, S., Trudel, L., Milord, F., Lindsay, L. R., Bélanger, D., & Ogden, N. H. (2011). Associations between Ixodes scapularis ticks and small mammal hosts in a newly endemic zone in southeastern Canada : Implications for Borrelia burgdorferi transmission. *Ticks and Tick-Borne Diseases*, *2*(4), 183-190. https://doi.org/10.1016/j.ttbdis.2011.03.005

Bour, C., Ahne, A., Schmitz, S., Perchoux, C., Dessenne, C., & Fagherazzi, G. (2021). The Use of Social Media for Health Research Purposes : Scoping Review. *Journal of Medical Internet Research*, *23*(5), e25736. https://doi.org/10.2196/25736

Bregnard, C., Rais, O., & Voordouw, M. J. (2020). Climate and tree seed production predict the abundance of the European Lyme disease vector over a 15-year period. *Parasites & Vectors*, *13*(1), 408. https://doi.org/10.1186/s13071-020-04291-z

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter : An analysis of the 2012-2013 influenza epidemic. *PloS One*, *8*(12), e83672. https://doi.org/10.1371/journal.pone.0083672

Brown, J. D. (2018). A description of « Australian Lyme disease » epidemiology and impact : An analysis of submissions to an Australian senate inquiry. *Internal Medicine Journal*, *48*(4), 422-426. https://doi.org/10.1111/imj.13746

Brownstein, J. S., Skelly, D. K., Holford, T. R., & Fish, D. (2005). Forest fragmentation predicts local scale heterogeneity of Lyme disease risk. *Oecologia*, *146*(3), 469-475. https://doi.org/10.1007/s00442-005-0251-9

Burgdorfer, W., Barbour, A. G., Hayes, S. F., Benach, J. L., Grunwaldt, E., & Davis, J. P. (1982). Lyme disease-a tick-borne spirochetosis? *Science (New York, N.Y.)*, *216*(4552), 1317-1319. https://doi.org/10.1126/science.7043737

Burgdorfer, W., Hayes, S. F., & Corwin, D. (1989). Pathophysiology of the Lyme disease spirochete, Borrelia burgdorferi, in ixodid ticks. *Reviews of Infectious Diseases*, *11 Suppl 6*, S1442-1450. https://doi.org/10.1093/clinids/11.supplement_6.s1442

Byrd, K., Mansurov, A., & Baysal, O. (2016). Mining Twitter data for influenza detection and surveillance. *Proceedings of the International Workshop on Software Engineering in Healthcare Systems*, 43-49. https://doi.org/10.1145/2897683.2897693

Campbell, G. L., Fritz, C. L., Fish, D., Nowakowski, J., Nadelman, R. B., & Wormser, G. P. (1998). Estimation of the incidence of Lyme disease. *American Journal of Epidemiology*, *148*(10), 1018-1026. https://doi.org/10.1093/oxfordjournals.aje.a009568

Canada, P. H. A. of. (2022, janvier 28). *Lyme disease surveillance in Canada : Preliminary annual report 2019* [Statistics]. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/Lyme-disease-surveillance-report-2019.html

Cardenas-de la Garza, J. A., De la Cruz-Valadez, E., Ocampo-Candiani, J., & Welsh, O. (2019). Clinical spectrum of Lyme disease. *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology*, *38*(2), 201-208. https://doi.org/10.1007/s10096-018-3417-1

Carneiro, H. A., & Mylonakis, E. (2009a). Google trends : A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *49*(10), 1557-1564. https://doi.org/10.1086/630200

Carneiro, H. A., & Mylonakis, E. (2009b). Google Trends : A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, *49*(10), 1557-1564. https://doi.org/10.1086/630200

Carriveau, A., Poole, H., & Thomas, A. (2019). Lyme Disease. *The Nursing Clinics of North America*, *54*(2), 261-275. https://doi.org/10.1016/j.cnur.2019.02.003

Cartter, M. L., Lynfield, R., Feldman, K. A., Hook, S. A., & Hinckley, A. F. (2018). Lyme disease surveillance in the United States : Looking for ways to cut the Gordian knot. *Zoonoses and Public Health*, *65*(2), 227-229. https://doi.org/10.1111/zph.12448

Castillo-Salgado, C. (2010). Trends and directions of global public health surveillance. *Epidemiologic Reviews*, *32*, 93-109. https://doi.org/10.1093/epirev/mxq008

CDC. (2021, avril 29). *Lyme disease. Data and Surveillance.* [INTERNET]. Center of Diseases Control and Prevention. https://www.cdc.gov/Lyme/datasurveillance/index.html

Chen, J., & Wang, Y. (2021). Social Media Use for Health Purposes : Systematic Review. *Journal of Medical Internet Research*, *23*(5), e17917. https://doi.org/10.2196/17917

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter : Content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One*, *5*(11), e14118. https://doi.org/10.1371/journal.pone.0014118

Chiang, A. L. (2021). Navigating and Leveraging Social Media. *Gastrointestinal Endoscopy Clinics of North America*, *31*(4), 695-707. https://doi.org/10.1016/j.giec.2021.05.006

Choi, D., Sumner, S. A., Holland, K. M., Draper, J., Murphy, S., Bowen, D. A., Zwald, M., Wang, J., Law, R., Taylor, J., Konjeti, C., & De Choudhury, M. (2020). Development of a Machine

Learning Model Using Multiple, Heterogeneous Data Sources to Estimate Weekly US Suicide Fatalities. *JAMA Network Open*, *3*(12), e2030932. https://doi.org/10.1001/jamanetworkopen.2020.30932

Chomel, B. (2015). Lyme disease. *Revue Scientifique Et Technique (International Office of Epizootics)*, *34*(2), 569-576. https://doi.org/10.20506/rst.34.2.2380

Chorianopoulos, K., & Talvis, K. (2016). Flutrack.org : Open-source and linked data for epidemiology. *Health Informatics Journal*, *22*(4), 962-974. https://doi.org/10.1177/1460458215599822

Clark, R. P., & Hu, L. T. (2008). Prevention of Lyme disease and other tick-borne infections. *Infectious Disease Clinics of North America*, *22*(3), 381-396, vii. https://doi.org/10.1016/j.idc.2008.03.007

Clayton, J. L., Jones, S. G., Dunn, J. R., Schaffner, W., & Jones, T. F. (2015). Enhancing Lyme Disease Surveillance by Using Administrative Claims Data, Tennessee, USA. *Emerging Infectious Diseases*, *21*(9), 1632-1634. https://doi.org/10.3201/eid2109.150344

Conway, M., Hu, M., & Chapman, W. W. (2019). Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and ConsumerGenerated Data. *Yearbook of Medical Informatics*, *28*(1), 208-217. https://doi.org/10.1055/s-0039-1677918

Cook, M. J. (2015). Lyme borreliosis : A review of data on transmission time after tick attachment. *International Journal of General Medicine*, *8*, 1-8. https://doi.org/10.2147/IJGM.S73791

Couper, L. I., MacDonald, A. J., & Mordecai, E. A. (2021). Impact of prior and projected climate change on US Lyme disease incidence. *Global Change Biology*, *27*(4), 738-754. https://doi.org/10.1111/gcb.15435

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. *Proceedings of the First Workshop on Social Media Analytics*, 115-122. https://doi.org/10.1145/1964858.1964874

Curriero, F. C., Wychgram, C., Rebman, A. W., Corrigan, A. E., Kvit, A., Shields, T., & Aucott, J. N. (2021). The Lyme and Tickborne Disease Dashboard : A map-based resource to promote public health awareness and research collaboration. *PloS One*, *16*(12), e0260122. https://doi.org/10.1371/journal.pone.0260122

De Silva, A. M., & Fikrig, E. (1995). Growth and migration of Borrelia burgdorferi in Ixodes ticks during blood feeding. *The American Journal of Tropical Medicine and Hygiene*, *53*(4), 397-404. https://doi.org/10.4269/ajtmh.1995.53.397

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *Undefined*. https://doi.org/10.18653/v1/N19-1423

Dias Canedo, E., & Cordeiro Mendes, B. (2020). Software Requirements Classification Using Machine Learning Algorithms. *Entropy (Basel, Switzerland)*, *22*(9), 1057. https://doi.org/10.3390/e22091057

Doan, S., Yang, E. W., Tilak, S. S., Li, P. W., Zisook, D. S., & Torii, M. (2019). Extracting health-related causality from twitter messages using natural language processing. *BMC Medical Informatics and Decision Making*, *19*(Suppl 3), 79. https://doi.org/10.1186/s12911-019-0785-0

Dumas, A., Bouchard, C., Lindsay, L. R., Ogden, N. H., & Leighton, P. A. (2022). Fine-scale determinants of the spatiotemporal distribution of Ixodes scapularis in Quebec (Canada). *Ticks and Tick-Borne Diseases*, *13*(1), 101833. https://doi.org/10.1016/j.ttbdis.2021.101833

Edo-Osagie, O., Smith, G., Lake, I., Edeghere, O., & De La Iglesia, B. (2019). Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. *PloS One*, *14*(7), e0210689. https://doi.org/10.1371/journal.pone.0210689

Eisen, R. J., & Eisen, L. (2018). The Blacklegged Tick, Ixodes scapularis : An Increasing Public Health Concern. *Trends in Parasitology*, *34*(4), 295-309. https://doi.org/10.1016/j.pt.2017.12.006

Eisen, R. J., Piesman, J., Zielinski-Gutierrez, E., & Eisen, L. (2012). What Do We Need to Know About Disease Ecology to Prevent Lyme Disease in the Northeastern United States? *Journal of Medical Entomology*, *49*(1), 11-22. https://doi.org/10.1603/ME11138

Eysenbach, G. (2002). Infodemiology : The epidemiology of (mis)information. *The American Journal of Medicine*, *113*(9), 763-765. https://doi.org/10.1016/s0002-9343(02)01473-0

Eysenbach, G. (2009). Infodemiology and Infoveillance : Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *Journal of Medical Internet Research*, *11*(1), e1157. https://doi.org/10.2196/jmir.1157

Gabarron, E., Larbi, D., Dorronzoro, E., Hasvold, P. E., Wynn, R., & Årsand, E. (2020). Factors Engaging Users of Diabetes Social Media Channels on Facebook, Twitter, and Instagram : Observational Study. *Journal of Medical Internet Research*, *22*(9), e21204. https://doi.org/10.2196/21204

Gasmi, S., Ogden, N. H., Leighton, P. A., Lindsay, L. R., & Thivierge, K. (2016). Analysis of the human population bitten by Ixodes scapularis ticks in Quebec, Canada : Increasing risk of Lyme

disease. *Ticks and Tick-Borne Diseases*, *7*(6), 1075-1081. https://doi.org/10.1016/j.ttbdis.2016.09.006

Gasmi, S., Ogden, N., Lindsay, L., Burns, S., Fleming, S., Badcock, J., Hanan, S., Gaulin, C., Leblanc, M., Russell, C., Nelder, M., Hobbs, L., Graham-Derham, S., Lachance, L., Scott, A., Galanis, E., & Koffi, J. (2017). Surveillance for Lyme disease in Canada : 2009–2015. *Canada Communicable Disease Report*, *43*(10), 194-199. https://doi.org/10.14745/ccdr.v43i10a01

Gavrielov-Yusim, N., Kürzinger, M.-L., Nishikawa, C., Pan, C., Pouget, J., Epstein, L. B., Golant, Y., Tcherny-Lessenot, S., Lin, S., Hamelin, B., & Juhaeri, J. (2019). Comparison of text processing methods in social media-based signal detection. *Pharmacoepidemiology and Drug Safety*, *28*(10), 1309-1317. https://doi.org/10.1002/pds.4857

Geebelen, L., Van Cauteren, D., Devleesschauwer, B., Moreels, S., Tersago, K., Van Oyen, H., Speybroeck, N., & Lernout, T. (2019). Combining primary care surveillance and a meta-analysis to estimate the incidence of the clinical manifestations of Lyme borreliosis in Belgium, 2015-2017. *Ticks and Tick-Borne Diseases*, *10*(3), 598-605. https://doi.org/10.1016/j.ttbdis.2018.12.007

Gern, L., Rouvinez, E., Toutoungi, L. N., & Godfroid, E. (1997). Transmission cycles of Borrelia burgdorferi sensu lato involving Ixodes ricinus and/or I. hexagonus ticks and the European hedgehog, Erinaceus europaeus, in suburban and urban areas in Switzerland. *Folia Parasitologica*, *44*(4), 309-314.

Gilbert, L. (2021). The Impacts of Climate Change on Ticks and Tick-Borne Disease Risk. *Annual Review of Entomology*, *66*, 373-388. https://doi.org/10.1146/annurev-ento-052720-094533

Gordillo-Pérez, G., Torres, J., Solórzano-Santos, F., Garduño-Bautista, V., Tapia-Conyer, R., & Muñoz, O. (2003). Estudio seroepidemiológico de borreliosis de Lyme en la Ciudad de México y el noroeste de la República Mexicana. *Salud Pública de México*, *45*(5), 351-355.

Gray, J. S., Kahl, O., Janetzki-Mittman, C., Stein, J., & Guy, E. (1994). Acquisition of Borrelia burgdorferi by Ixodes ricinus ticks fed on the European hedgehog, Erinaceus europaeus L. *Experimental & Applied Acarology*, *18*(8), 485-491. https://doi.org/10.1007/BF00051470

Guo, Y., Ge, Y., Yang, Y.-C., Al-Garadi, M. A., & Sarker, A. (2022). Comparison of Pretraining Models and Strategies for Health-Related Social Media Text Classification. *Healthcare (Basel, Switzerland)*, *10*(8), 1478. https://doi.org/10.3390/healthcare10081478

Gupta, A., & Katarya, R. (2020). Social media based surveillance systems for healthcare using machine learning : A systematic review. *Journal of Biomedical Informatics*, *108*, 103500. https://doi.org/10.1016/j.jbi.2020.103500

Hanincova, K., Mukherjee, P., Ogden, N. H., Margos, G., Wormser, G. P., Reed, K. D., Meece, J. K., Vandermause, M. F., & Schwartz, I. (2013). Multilocus sequence typing of Borrelia burgdorferi suggests existence of lineages with differential pathogenic properties in humans. *PloS One*, *8*(9), e73066. https://doi.org/10.1371/journal.pone.0073066

Hanson, M. S., & Edelman, R. (2003). Progress and controversy surrounding vaccines against Lyme disease. *Expert Review of Vaccines*, *2*(5), 683-703. https://doi.org/10.1586/14760584.2.5.683

Hasan, M. K., Ghazal, T. M., Alkhalifah, A., Abu Bakar, K. A., Omidvar, A., Nafi, N. S., & Agbinya, J. I. (2021). Fischer Linear Discrimination and Quadratic Discrimination Analysis-Based

Data Mining Technique for Internet of Things Framework for Healthcare. *Frontiers in Public Health*, *9*, 737149. https://doi.org/10.3389/fpubh.2021.737149

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Model Assessment and Selection. In *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (p. 219-259). Springer New York. https://doi.org/10.1007/978-0-387-84858-7_7

Hussain, S., Hussain, A., Aziz, U., Song, B., Zeb, J., George, D., Li, J., & Sparagano, O. (2021). The Role of Ticks in the Emergence of Borrelia burgdorferi as a Zoonotic Pathogen and Its Vector Control : A Global Systemic Review. *Microorganisms*, *9*(12), 2412. https://doi.org/10.3390/microorganisms9122412

Jahanbin, K., Rahmanian, F., Rahmanian, V., & Jahromi, A. S. (2019). Application of Twitter and web news mining in infectious disease surveillance systems and prospects for public health. *GMS Hygiene and Infection Control*, *14*, Doc19. https://doi.org/10.3205/dgkh000334

Jalilifard, A., Carid'a, V. F., Mansano, A. F., & Cristo, R. (2021). Semantic Sensitive TF-IDF to Determine Word Relevance in Documents. *Undefined*. https://doi.org/10.1007/978-981-33-6987-0_27

Jánová, E. (2019). Emerging and threatening vector-borne zoonoses in the world and in Europe : A brief update. *Pathogens and Global Health*, *113*(2), 49-57. https://doi.org/10.1080/20477724.2019.1598127

Ji, Z., Jian, M., Yue, P., Cao, W., Xu, X., Zhang, Y., Pan, Y., Yang, J., Chen, J., Liu, M., Fan, Y., Su, X., Wen, S., Kong, J., Li, B., Dong, Y., Zhou, G., Liu, A., & Bao, F. (2022). Prevalence of Borrelia burgdorferi in Ixodidae Tick around Asia : A Systematic Review and Meta-Analysis. *Pathogens (Basel, Switzerland)*, *11*(2), 143. https://doi.org/10.3390/pathogens11020143

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 338-345.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990-993. https://doi.org/10.1038/nature06536

Kahl, O., Janetzki-Mittmann, C., Gray, J. S., Jonas, R., Stein, J., & de Boer, R. (1998). Risk of infection with Borrelia burgdorferi sensu lato for a host in relation to the duration of nymphal Ixodes ricinus feeding and the method of tick removal. *Zentralblatt Fur Bakteriologie: International Journal of Medical Microbiology*, *287*(1-2), 41-52. https://doi.org/10.1016/s0934-8840(98)80142-4

Kapitány-Fövény, M., Ferenci, T., Sulyok, Z., Kegele, J., Richter, H., Vályi-Nagy, I., & Sulyok, M. (2019). Can Google Trends data improve forecasting of Lyme disease incidence? *Zoonoses and Public Health*, *66*(1), 101-107. https://doi.org/10.1111/zph.12539

Kilpatrick, A. M., Dobson, A. D. M., Levi, T., Salkeld, D. J., Swei, A., Ginsberg, H. S., Kjemtrup, A., Padgett, K. A., Jensen, P. M., Fish, D., Ogden, N. H., & Diuk-Wasser, M. A. (2017). Lyme disease ecology in a changing world : Consensus, uncertainty and critical gaps for improving control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1722), 20160117. https://doi.org/10.1098/rstb.2016.0117

Kilpatrick, H. J., LaBonte, A. M., & Stafford, K. C. (2014). The relationship between deer density, tick abundance, and human cases of Lyme disease in a residential community. *Journal of Medical Entomology*, *51*(4), 777-784. https://doi.org/10.1603/me13232

Kim, D., Maxwell, S., & Le, Q. (2020a). Spatial and Temporal Comparison of Perceived Risks and Confirmed Cases of Lyme Disease : An Exploratory Study of Google Trends. *Frontiers in Public Health*, *8*, 395. https://doi.org/10.3389/fpubh.2020.00395

Kim, D., Maxwell, S., & Le, Q. (2020b). Spatial and Temporal Comparison of Perceived Risks and Confirmed Cases of Lyme Disease : An Exploratory Study of Google Trends. *Frontiers in Public Health*, *8*, 395. https://doi.org/10.3389/fpubh.2020.00395

Klein, A. Z., Magge, A., O'Connor, K., Flores Amaro, J. I., Weissenbacher, D., & Gonzalez Hernandez, G. (2021). Toward Using Twitter for Tracking COVID-19 : A Natural Language Processing Pipeline and Exploratory Data Set. *Journal of Medical Internet Research*, *23*(1), e25314. https://doi.org/10.2196/25314

Kugeler, K. J., Cervantes, K., Brown, C. M., Horiuchi, K., Schiffman, E., Lind, L., Barkley, J., Broyhill, J., Murphy, J., Crum, D., Robinson, S., Kwit, N. A., Mullins, J., Sun, J., & Hinckley, A. F. (2022). Potential quantitative effect of a laboratory-based approach to Lyme disease surveillance in high-incidence states. *Zoonoses and Public Health*, *69*(5), 451-457. https://doi.org/10.1111/zph.12933

Kugeler, K. J., Farley, G. M., Forrester, J. D., & Mead, P. S. (2015a). Geographic Distribution and Expansion of Human Lyme Disease, United States. *Emerging Infectious Diseases*, *21*(8), 1455-1457. https://doi.org/10.3201/eid2108.141878

Kugeler, K. J., Farley, G. M., Forrester, J. D., & Mead, P. S. (2015b). Geographic Distribution and Expansion of Human Lyme Disease, United States. *Emerging Infectious Diseases*, *21*(8), 1455-1457. https://doi.org/10.3201/eid2108.141878

Kurokawa, C., Lynn, G. E., Pedra, J. H. F., Pal, U., Narasimhan, S., & Fikrig, E. (2020). Interactions between Borrelia burgdorferi and ticks. *Nature Reviews. Microbiology*, *18*(10), 587-600. https://doi.org/10.1038/s41579-020-0400-5

Kurtenbach, K., Hanincová, K., Tsao, J. I., Margos, G., Fish, D., & Ogden, N. H. (2006). Fundamental processes in the evolutionary ecology of Lyme borreliosis. *Nature Reviews. Microbiology*, *4*(9), 660-669. https://doi.org/10.1038/nrmicro1475

Kutera, M., Berke, O., & Sobkowich, K. (2021). Spatial epidemiological analysis of Lyme disease in southern Ontario utilizing Google Trends searches. *Environmental Health Review*, *64*(4), 105-110. https://doi.org/10.5864/d2021-025

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT : A Lite BERT for Self-supervised Learning of Language Representations. *Undefined*. https://www.semanticscholar.org/reader/7a064df1aeada7e69e5173f7d4c8606f4470365b

Lantos, P. M., Rumbaugh, J., Bockenstedt, L. K., Falck-Ytter, Y. T., Aguero-Rosenfeld, M. E., Auwaerter, P. G., Baldwin, K., Bannuru, R. R., Belani, K. K., Bowie, W. R., Branda, J. A., Clifford, D. B., DiMario, F. J., Jr, Halperin, J. J., Krause, P. J., Lavergne, V., Liang, M. H., Meissner, H. C., Nigrovic, L. E., … Zemel, L. S. (2021). Clinical Practice Guidelines by the Infectious Diseases Society of America (IDSA), American Academy of Neurology (AAN), and American College of Rheumatology (ACR) : 2020 Guidelines for the Prevention, Diagnosis and Treatment of Lyme Disease. *Clinical Infectious Diseases*, *72*(1), e1-e48. https://doi.org/10.1093/cid/ciaa1215

Laranjo, L., Arguel, A., Neves, A. L., Gallagher, A. M., Kaplan, R., Mortimer, N., Mendes, G. A., & Lau, A. Y. S. (2015). The influence of social networking sites on health behavior change : A

systematic review and meta-analysis. *Journal of the American Medical Informatics Association: JAMIA*, *22*(1), 243-256. https://doi.org/10.1136/amiajnl-2014-002841

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT : A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

Lee, W. C., Lee, M. J., Choi, K. H., Chung, H. S., & Choe, N. H. (2019). A comparative study of the trends in epidemiological aspects of Lyme disease infections in Korea and Japan, 2011-2016. *Journal of Vector Borne Diseases*, *56*(3), 268-271. https://doi.org/10.4103/0972-9062.289396

Lindsay, L. R., Mathison, S. W., Barker, I. K., McEwen, S. A., Gillespie, T. J., & Surgeoner, G. A. (1999). Microclimate and habitat in relation to Ixodes scapularis (Acari : Ixodidae) populations on Long Point, Ontario, Canada. *Journal of Medical Entomology*, *36*(3), 255-262. https://doi.org/10.1093/jmedent/36.3.255

Lou, Y., & Wu, J. (2017). Modeling Lyme disease transmission. *Infectious Disease Modelling*, *2*(2), 229-243. https://doi.org/10.1016/j.idm.2017.05.002

Lyu, J. C., Han, E. L., & Luli, G. K. (2021). COVID-19 Vaccine-Related Discussion on Twitter : Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research*, *23*(6), e24435. https://doi.org/10.2196/24435

Mac, S., Evans, G. A., Patel, S. N., Pullenayegum, E. M., & Sander, B. (2021). Estimating the population health burden of Lyme disease in Ontario, Canada : A microsimulation modelling approach. *CMAJ Open*, *9*(4), E1005-E1012. https://doi.org/10.9778/cmajo.20210024

Mac, S., Silva, S. R. da, & Sander, B. (2019). The economic burden of Lyme disease and the cost-effectiveness of Lyme disease interventions : A scoping review. *PLOS ONE*, *14*(1), e0210280. https://doi.org/10.1371/journal.pone.0210280

Mackey, T. K., Li, J., Purushothaman, V., Nali, M., Shah, N., Bardier, C., Cai, M., & Liang, B. (2020). Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales : Infoveillance Study on Twitter and Instagram. *JMIR Public Health and Surveillance*, *6*(3), e20794. https://doi.org/10.2196/20794

Mahroum, N., Bragazzi, N. L., Sharif, K., Gianfredi, V., Nucci, D., Rosselli, R., Brigo, F., Adawi, M., Amital, H., & Watad, A. (2018). Leveraging Google Trends, Twitter, and Wikipedia to Investigate the Impact of a Celebrity's Death From Rheumatoid Arthritis. *Journal of Clinical Rheumatology: Practical Reports on Rheumatic & Musculoskeletal Diseases*, *24*(4), 188-192. https://doi.org/10.1097/RHU.0000000000000692

Maksimyan, S., Syed, M. S., & Soti, V. (2021). Post-Treatment Lyme Disease Syndrome : Need for Diagnosis and Treatment. *Cureus*, *13*(10), e18703. https://doi.org/10.7759/cureus.18703

Marques, A. R., Strle, F., & Wormser, G. P. (2021). Comparison of Lyme Disease in the United States and Europe. *Emerging Infectious Diseases*, *27*(8), 2017-2024. https://doi.org/10.3201/eid2708.204763

Mayor, E., & Bietti, L. M. (2021). Twitter, time and emotions. *Royal Society Open Science*, *8*(5), 201900. https://doi.org/10.1098/rsos.201900

Nam, Y. H., Willis, S. J., Mendelsohn, A. B., Forrow, S., Gessner, B. D., Stark, J. H., Brown, J. S., & Pugh, S. (2022). Healthcare claims-based Lyme disease case-finding algorithms in the United

States : A systematic literature review. *PloS One*, *17*(10), e0276299. https://doi.org/10.1371/journal.pone.0276299

Nasser, N., Karim, L., El Ouadrhiri, A., Ali, A., & Khan, N. (2021). N-Gram based language processing using Twitter dataset to identify COVID-19 patients. *Sustainable Cities and Society*, *72*, 103048. https://doi.org/10.1016/j.scs.2021.103048

Nelson, C. A., Saha, S., Kugeler, K. J., Delorey, M. J., Shankar, M. B., Hinckley, A. F., & Mead, P. S. (2015). Incidence of Clinician-Diagnosed Lyme Disease, United States, 2005-2010. *Emerging Infectious Diseases*, *21*(9), 1625-1631. https://doi.org/10.3201/eid2109.150417

Nguyen, C. T., Cifu, A. S., & Pitrak, D. (2022). Prevention and Treatment of Lyme Disease. *JAMA*, *327*(8), 772-773. https://doi.org/10.1001/jama.2021.25302

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020a). BERTweet : A pre-trained language model for English Tweets. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9-14. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020b). BERTweet : A pre-trained language model for English Tweets. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9-14. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Nyataya, J., Maraka, M., Lemtudo, A., Masakhwe, C., Mutai, B., Njaanake, K., Estambale, B. B., Nyakoe, N., Siangla, J., & Waitumbi, J. N. (2020). Serological Evidence of Yersiniosis, Tick-Borne Encephalitis, West Nile, Hepatitis E, Crimean-Congo Hemorrhagic Fever, Lyme Borreliosis, and

Brucellosis in Febrile Patients Presenting at Diverse Hospitals in Kenya. *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)*, *20*(5), 348-357. https://doi.org/10.1089/vbz.2019.2484

Ogden, N. H., Feil, E. J., Leighton, P. A., Lindsay, L. R., Margos, G., Mechai, S., Michel, P., & Moriarty, T. J. (2015). Evolutionary aspects of emerging Lyme disease in Canada. *Applied and Environmental Microbiology*, *81*(21), 7350-7359. https://doi.org/10.1128/AEM.01671-15

Ogden, N. H., St-Onge, L., Barker, I. K., Brazeau, S., Bigras-Poulin, M., Charron, D. F., Francis, C. M., Heagy, A., Lindsay, L. R., Maarouf, A., Michel, P., Milord, F., O'Callaghan, C. J., Trudel, L., & Thompson, R. A. (2008). Risk maps for range expansion of the Lyme disease vector, Ixodes scapularis, in Canada now and with climate change. *International Journal of Health Geographics*, *7*, 24. https://doi.org/10.1186/1476-072X-7-24

Ojo, S., Sari, A., & Ojo, T. P. (2022). Path Loss Modeling : A Machine Learning Based Approach Using Support Vector Regression and Radial Basis Function Models. *Open Journal of Applied Sciences*, *12*(6), Art. 6. https://doi.org/10.4236/ojapps.2022.126068

Ostfeld, R. S., & Brunner, J. L. (2015). Climate change and Ixodes tick-borne diseases of humans. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1665), 20140051. https://doi.org/10.1098/rstb.2014.0051

Paul, M., & Dredze, M. (2011). You Are What You Tweet : Analyzing Twitter for Public Health. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), Art. 1.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830.

Pesälä, S., Virtanen, M. J., Sane, J., Mustonen, P., Kaila, M., & Helve, O. (2017). Health Information–Seeking Patterns of the General Public and Indications for Disease Surveillance : Register-Based Study Using Lyme Disease. *JMIR Public Health and Surveillance*, *3*(4), e8306. https://doi.org/10.2196/publichealth.8306

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227-2237. https://doi.org/10.18653/v1/N18-1202

Piesman, J., & Gern, L. (2004a). Lyme borreliosis in Europe and North America. *Parasitology*, *129 Suppl*, S191-220. https://doi.org/10.1017/s0031182003004694

Piesman, J., & Gern, L. (2004b). Lyme borreliosis in Europe and North America. *Parasitology*, *129 Suppl*, S191-220. https://doi.org/10.1017/s0031182003004694

Pollett, S., Althouse, B. M., Forshey, B., Rutherford, G. W., & Jarman, R. G. (2017). Internet-based biosurveillance methods for vector-borne diseases : Are they novel public health tools or just novelties? *PLoS Neglected Tropical Diseases*, *11*(11), e0005871. https://doi.org/10.1371/journal.pntd.0005871

Ratti, V., Winter, J. M., & Wallace, D. I. (2021). Dilution and amplification effects in Lyme disease : Modeling the effects of reservoir-incompetent hosts on Borrelia burgdorferi sensu stricto transmission. *Ticks and Tick-Borne Diseases*, *12*(4), 101724. https://doi.org/10.1016/j.ttbdis.2021.101724

Rebman, A. W., Yang, T., Mihm, E. A., Novak, C. B., Yoon, I., Powell, D., Geller, S. A., & Aucott, J. N. (2021). The presenting characteristics of erythema migrans vary by age, sex, duration, and body location. *Infection*, *49*(4), 685-692. https://doi.org/10.1007/s15010-021-01590-0

Ripoche, M., Gasmi, S., Adam-Poupart, A., Koffi, J. K., Lindsay, L. R., Ludwig, A., Milord, F., Ogden, N. H., Thivierge, K., & Leighton, P. A. (2018). Passive Tick Surveillance Provides an Accurate Early Signal of Emerging Lyme Disease Risk and Human Cases in Southern Canada. *Journal of Medical Entomology*, *55*(4), 1016-1026. https://doi.org/10.1093/jme/tjy030

Rosenberg, R., Lindsey, N. P., Fischer, M., Gregory, C. J., Hinckley, A. F., Mead, P. S., Paz-Bailey, G., Waterman, S. H., Drexler, N. A., Kersh, G. J., Hooks, H., Partridge, S. K., Visser, S. N., Beard, C. B., & Petersen, L. R. (2018). Vital Signs : Trends in Reported Vectorborne Disease Cases - United States and Territories, 2004-2016. *MMWR. Morbidity and Mortality Weekly Report*, *67*(17), 496-501. https://doi.org/10.15585/mmwr.mm6717e1

Sadilek, A., Hswen, Y., Bavadekar, S., Shekel, T., Brownstein, J. S., & Gabrilovich, E. (2020). Lymelight : Forecasting Lyme disease risk using web search data. *Npj Digital Medicine*, *3*(1), Art. 1. https://doi.org/10.1038/s41746-020-0222-x

Sanchez, J. L. (2015). Clinical Manifestations and Treatment of Lyme Disease. *Clinics in Laboratory Medicine*, *35*(4), 765-778. https://doi.org/10.1016/j.cll.2015.08.004

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT : Smaller, faster, cheaper and lighter. *Undefined*. https://www.semanticscholar.org/reader/a54b56af24bb4873ed0163b77df63b92bd018ddc

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., & Brownstein, J. S. (2015). Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology*, *11*(10), e1004513. https://doi.org/10.1371/journal.pcbi.1004513

Sarker, A., Lakamana, S., Hogg-Bremer, W., Xie, A., Al-Garadi, M. A., & Yang, Y.-C. (2020). Self-reported COVID-19 symptoms on Twitter : An analysis and a research resource. *Journal of the American Medical Informatics Association: JAMIA*, *27*(8), 1310-1315. https://doi.org/10.1093/jamia/ocaa116

Scheerer, C., Rüth, M., Tizek, L., Köberle, M., Biedermann, T., & Zink, A. (2020). Googling for Ticks and Borreliosis in Germany : Nationwide Google Search Analysis From 2015 to 2018. *Journal of Medical Internet Research*, *22*(10), e18581. https://doi.org/10.2196/18581

Schwartz, A. M., Kugeler, K. J., Nelson, C. A., Marx, G. E., & Hinckley, A. F. (2021). Use of Commercial Claims Data for Evaluating Trends in Lyme Disease Diagnoses, United States, 2010-2018. *Emerging Infectious Diseases*, *27*(2), 499-507. https://doi.org/10.3201/eid2702.202728

Seifter, A., Schwarzwalder, A., Geis, K., & Aucott, J. (2010). The utility of « Google Trends » for epidemiological research : Lyme disease as an example. *Geospatial Health*, *4*(2), 135-137. https://doi.org/10.4081/gh.2010.195

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PloS One*, *16*(7), e0254034. https://doi.org/10.1371/journal.pone.0254034

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Deep Learning applications for COVID-19. *Journal of Big Data*, *8*(1), 18. https://doi.org/10.1186/s40537-020-00392-9

Silva Barbon, R., & Akabane, A. T. (2022). Towards Transfer Learning Techniques-BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different

Languages : A Case Study. *Sensors (Basel, Switzerland)*, *22*(21), 8184. https://doi.org/10.3390/s22218184

Simon, J. A., Marrotte, R. R., Desrosiers, N., Fiset, J., Gaitan, J., Gonzalez, A., Koffi, J. K., Lapointe, F.-J., Leighton, P. A., Lindsay, L. R., Logan, T., Milord, F., Ogden, N. H., Rogic, A., Roy-Dufresne, E., Suter, D., Tessier, N., & Millien, V. (2014). Climate change and habitat fragmentation drive the occurrence of Borrelia burgdorferi, the agent of Lyme disease, at the northeastern limit of its distribution. *Evolutionary Applications*, *7*(7), 750-764. https://doi.org/10.1111/eva.12165

Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a Tool for Health Research : A Systematic Review. *American Journal of Public Health*, *107*(1), e1-e8. https://doi.org/10.2105/AJPH.2016.303512

Stanek, G., & Strle, F. (2008). Lyme disease : European perspective. *Infectious Disease Clinics of North America*, *22*(2), 327-339, vii. https://doi.org/10.1016/j.idc.2008.01.001

Stanek, G., & Strle, F. (2018). Lyme borreliosis-from tick bite to diagnosis and treatment. *FEMS Microbiology Reviews*, *42*(3), 233-258. https://doi.org/10.1093/femsre/fux047

Stanek, G., Wormser, G. P., Gray, J., & Strle, F. (2012). Lyme borreliosis. *Lancet (London, England)*, *379*(9814), 461-473. https://doi.org/10.1016/S0140-6736(11)60103-7

Stark, J. H., Li, X., Zhang, J. C., Burn, L., Valluri, S. R., Liang, J., Pan, K., Fletcher, M. A., Simon, R., Jodar, L., & Gessner, B. D. (2022). Systematic Review and Meta-analysis of Lyme Disease Data and Seropositivity for Borrelia burgdorferi, China, 2005–2020. *Emerging Infectious Diseases*, *28*(12), 2389-2397. https://doi.org/10.3201/eid2812.212612

Steere, A. C. (2001a). Lyme disease. *The New England Journal of Medicine*, *345*(2), 115-125. https://doi.org/10.1056/NEJM200107123450207

Steere, A. C. (2001b). Lyme disease. *The New England Journal of Medicine*, *345*(2), 115-125. https://doi.org/10.1056/NEJM200107123450207

Steere, A. C., Dhar, A., Hernandez, J., Fischer, P. A., Sikand, V. K., Schoen, R. T., Nowakowski, J., McHugh, G., & Persing, D. H. (2003). Systemic symptoms without erythema migrans as the presenting picture of early Lyme disease. *The American Journal of Medicine*, *114*(1), 58-62. https://doi.org/10.1016/s0002-9343(02)01440-7

Stevens, R., Bonett, S., Bannon, J., Chittamuru, D., Slaff, B., Browne, S. K., Huang, S., & Bauermeister, J. A. (2020). Association Between HIV-Related Tweets and HIV Incidence in the United States : Infodemiology Study. *Journal of Medical Internet Research*, *22*(6), e17196. https://doi.org/10.2196/17196

Stewart, P. E., & Rosa, P. A. (2018). Physiologic and Genetic Factors Influencing the Zoonotic Cycle of Borrelia burgdorferi. *Current Topics in Microbiology and Immunology*, *415*, 63-82. https://doi.org/10.1007/82_2017_43

Stone, B. L., Tourand, Y., & Brissette, C. A. (2017). Brave New Worlds : The Expanding Universe of Lyme Disease. *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)*, *17*(9), 619-629. https://doi.org/10.1089/vbz.2017.2127

Stoové, M. A., & Pedrana, A. E. (2014). Making the most of a brave new world : Opportunities and considerations for using Twitter as a public health monitoring tool. *Preventive Medicine*, *63*, 109-111. https://doi.org/10.1016/j.ypmed.2014.03.008

Tang, J., Henderson, A., & Gardner, P. (2021). Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced data sets. *The Analyst*, *146*(19), 5880-5891. https://doi.org/10.1039/d0an02155e

Tejani, A. S., Ng, Y. S., Xi, Y., Fielding, J. R., Browning, T. G., & Rayan, J. C. (2022). Performance of Multiple Pretrained BERT Models to Automate and Accelerate Data Annotation for Large Datasets. *Radiology. Artificial Intelligence*, *4*(4), e220007. https://doi.org/10.1148/ryai.220007

Tseng, Y.-J., Cami, A., Goldmann, D. A., DeMaria, A., & Mandl, K. D. (2015). Using Nation-Wide Health Insurance Claims Data to Augment Lyme Disease Surveillance. *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)*, *15*(10), 591-596. https://doi.org/10.1089/vbz.2015.1790

Tulloch, J. S. P., Vivancos, R., Christley, R. M., Radford, A. D., & Warner, J. C. (2019a). Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and Republic of Ireland. *Journal of Biomedical Informatics*, *100S*, 100060. https://doi.org/10.1016/j.yjbinx.2019.100060

Tulloch, J. S. P., Vivancos, R., Christley, R. M., Radford, A. D., & Warner, J. C. (2019b). Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and Republic of Ireland. *Journal of Biomedical Informatics*, *100S*, 100060. https://doi.org/10.1016/j.yjbinx.2019.100060

Tulloch, J. S. P., Vivancos, R., Christley, R. M., Radford, A. D., & Warner, J. C. (2019c). Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and Republic of Ireland. *Articles initially published in Journal of Biomedical Informatics: X 1-4, 2019*, *100*, 100060. https://doi.org/10.1016/j.yjbinx.2019.100060

Tutt-Guérette, M.-A., Yuan, M., Szaroz, D., McKinnon, B., Kestens, Y., Guillot, C., Leighton, P., & Zinszer, K. (2021). Modelling Spatiotemporal Patterns of Lyme Disease Emergence in Québec. *International Journal of Environmental Research and Public Health*, *18*(18), 9669. https://doi.org/10.3390/ijerph18189669

Vinayaraj, E. V., Gupta, N., Sreenath, K., Thakur, C. K., Gulati, S., Anand, V., Tripathi, M., Bhatia, R., Vibha, D., Dash, D., Soneja, M., Kumar, U., Padma, M. V., & Chaudhry, R. (2021). Clinical and laboratory evidence of Lyme disease in North India, 2016-2019. *Travel Medicine and Infectious Disease*, *43*, 102134. https://doi.org/10.1016/j.tmaid.2021.102134

Wen, S., Xu, Q., Liu, D., Lin, Z., Lin, Z., Chen, S., & Chen, M. (2021). A seroepidemiological investigation of Lyme disease in Qiongzhong County, Hainan Province in 2019-2020. *Annals of Palliative Medicine*, *10*(4), 4721-4727. https://doi.org/10.21037/apm-21-693

Werden, L., Barker, I. K., Bowman, J., Gonzales, E. K., Leighton, P. A., Lindsay, L. R., & Jardine, C. M. (2014). Geography, deer, and host biodiversity shape the pattern of Lyme disease emergence in the Thousand Islands Archipelago of Ontario, Canada. *PloS One*, *9*(1), e85640. https://doi.org/10.1371/journal.pone.0085640

Wormser, G. P., Dattwyler, R. J., Shapiro, E. D., Halperin, J. J., Steere, A. C., Klempner, M. S., Krause, P. J., Bakken, J. S., Strle, F., Stanek, G., Bockenstedt, L., Fish, D., Dumler, J. S., & Nadelman, R. B. (2006a). The Clinical Assessment, Treatment, and Prevention of Lyme Disease, Human Granulocytic Anaplasmosis, and Babesiosis : Clinical Practice Guidelines by the Infectious Diseases Society of America. *Clinical Infectious Diseases*, *43*(9), 1089-1134. https://doi.org/10.1086/508667

Wormser, G. P., Dattwyler, R. J., Shapiro, E. D., Halperin, J. J., Steere, A. C., Klempner, M. S., Krause, P. J., Bakken, J. S., Strle, F., Stanek, G., Bockenstedt, L., Fish, D., Dumler, J. S., & Nadelman, R. B. (2006b). The clinical assessment, treatment, and prevention of Lyme disease, human granulocytic anaplasmosis, and babesiosis : Clinical practice guidelines by the Infectious Diseases Society of America. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *43*(9), 1089-1134. https://doi.org/10.1086/508667

Yaya, S., & Ghose, B. (2018). Patterns of computer and Internet use and its association with HIV knowledge in selected countries in sub-Saharan Africa. *PloS One*, *13*(6), e0199236. https://doi.org/10.1371/journal.pone.0199236

Yiannakoulias, N., Tooby, R., & Sturrock, S. L. (2017). Celebrity over science ? An analysis of Lyme disease video content on YouTube. *Social Science & Medicine*, *191*, 57-60. https://doi.org/10.1016/j.socscimed.2017.08.042

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis : A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1253.

Zhang, X., Meltzer, M. I., Peña, C. A., Hopkins, A. B., Wroth, L., & Fix, A. D. (2006). Economic impact of Lyme disease. *Emerging Infectious Diseases*, *12*(4), 653-660. https://doi.org/10.3201/eid1204.050602

Zhao, W., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.-P., & Li, X. (2011). *Topical Keyphrase Extraction from Twitter* (p. 388).

Zhou, G., Xu, X., Zhang, Y., Yue, P., Luo, S., Fan, Y., Chen, J., Liu, M., Dong, Y., Li, B., Kong, J., Wen, S., Liu, A., & Bao, F. (2021). Antibiotic prophylaxis for prevention against Lyme disease

following tick bite : An updated systematic review and meta-analysis. *BMC Infectious Diseases*, *21*(1), 1141. https://doi.org/10.1186/s12879-021-06837-