Original Paper

Identifying Potential Lyme Disease Cases Using Self-Reported Worldwide Tweets: Deep Learning Modeling Approach Enhanced With Sentimental Words Through Emojis

Elda Kokoe Elolo Laison[1], MSc, MD; Mohamed Hamza Ibrahim[2], PhD; Srikanth Boligarla[3], ALM; Jiaxin Li[3], ALM; Raja Mahadevan[3], ALM; Austen Ng[3], ALM; Venkataraman Muthuramalingam[3], ALM; Wee Yi Lee[3], ALM; Yijun Yin[3], ALM; Bouchra R Nasri[1], PhD

[1]Département de médecine sociale et préventive, École de Santé Publique de l'Université de Montréal, Université de Montréal, Montréal, QC, Canada
[2]Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt
[3]Harvard Extension School, Harvard University, Cambridge, MA, United States

Corresponding Author:
Bouchra R Nasri, PhD
Département de médecine sociale et préventive
École de Santé Publique de l'Université de Montréal
Université de Montréal
7101 Park Ave
Montréal, QC, H3N 1X9
Canada
Phone: 1 514 343 7973
Email: bouchra.nasri@umontreal.ca

Abstract

Background: Lyme disease is among the most reported tick-borne diseases worldwide, making it a major ongoing public health concern. An effective Lyme disease case reporting system depends on timely diagnosis and reporting by health care professionals, and accurate laboratory testing and interpretation for clinical diagnosis validation. A lack of these can lead to delayed diagnosis and treatment, which can exacerbate the severity of Lyme disease symptoms. Therefore, there is a need to improve the monitoring of Lyme disease by using other data sources, such as web-based data.

Objective: We analyzed global Twitter data to understand its potential and limitations as a tool for Lyme disease surveillance. We propose a transformer-based classification system to identify potential Lyme disease cases using self-reported tweets.

Methods: Our initial sample included 20,000 tweets collected worldwide from a database of over 1.3 million Lyme disease tweets. After preprocessing and geolocating tweets, tweets in a subset of the initial sample were manually labeled as potential Lyme disease cases or non-Lyme disease cases using carefully selected keywords. Emojis were converted to sentiment words, which were then replaced in the tweets. This labeled tweet set was used for the training, validation, and performance testing of DistilBERT (distilled version of BERT [Bidirectional Encoder Representations from Transformers]), ALBERT (A Lite BERT), and BERTweet (BERT for English Tweets) classifiers.

Results: The empirical results showed that BERTweet was the best classifier among all evaluated models (average F1-score of 89.3%, classification accuracy of 90.0%, and precision of 97.1%). However, for recall, term frequency-inverse document frequency and k-nearest neighbors performed better (93.2% and 82.6%, respectively). On using emojis to enrich the tweet embeddings, BERTweet had an increased recall (8% increase), DistilBERT had an increased F1-score of 93.8% (4% increase) and classification accuracy of 94.1% (4% increase), and ALBERT had an increased F1-score of 93.1% (5% increase) and classification accuracy of 93.9% (5% increase). The general awareness of Lyme disease was high in the United States, the United Kingdom, Australia, and Canada, with self-reported potential cases of Lyme disease from these countries accounting for around 50% (9939/20,000) of the collected English-language tweets, whereas Lyme disease–related tweets were rare in countries from Africa and Asia. The most reported Lyme disease–related symptoms in the data were rash, fatigue, fever, and arthritis, while symptoms, such as lymphadenopathy, palpitations, swollen lymph nodes, neck stiffness, and arrythmia, were uncommon, in accordance with Lyme disease symptom frequency.

Conclusions: The study highlights the robustness of BERTweet and DistilBERT as classifiers for potential cases of Lyme disease from self-reported data. The results demonstrated that emojis are effective for enrichment, thereby improving the accuracy of tweet embeddings and the performance of classifiers. Specifically, emojis reflecting sadness, empathy, and encouragement can reduce false negatives.

(J Med Internet Res 2023;25:e47014) doi: 10.2196/47014

KEYWORDS

Lyme disease; Twitter; BERT; Bidirectional Encoder Representations from Transformers; emojis; machine learning; natural language processing

Introduction

Global warming and milder winters are causing the range of tick vectors to expand, which in turn is contributing to an increase in the incidence of tick-borne diseases [1-4]. Lyme disease is one of the most commonly reported tick-borne diseases worldwide [5]. In North America, Lyme disease is endemic in the northeastern, upper mid-West, and mid-Atlantic portions of the United States, and is prevalent in the southern regions of Canada [6-8]. In Europe, Lyme disease is mainly found in the central regions of the continent and in Scandinavian countries, and it is also found in Russia [6,9,10]. The occurrence of Lyme disease is very recent in Asia and has been reported in India, Turkey, China, Korea, Nepal, Taiwan, and Japan [11-16]. Owing to the current wide geographical spread of this disease, the early detection of potential Lyme disease cases will remain a public health concern in the forthcoming decades [17].

Lyme disease is caused by spirochetal bacteria that are part of the *Borrelia burgdorferi sensu lato* (*s.l.*) complex [17]. The *Borrelia burgdorferi sensu lato* (*s.l.*) complex contains numerous genospecies, but only a few can infect humans and cause Lyme disease. The genospecies that can infect humans have distinct geographic distributions: *Borrelia burgdorferi sensu stricto* is primarily found in North America, whereas *Borrelia afzelii* and *Borrelia garinii* are both prevalent in Asia and Europe [18,19]. Furthermore, the clinical manifestations of Lyme disease vary depending on the genospecies involved in the infection, and thus, the symptoms also vary by geographical region. In North America, *B. burgdorferi sensu stricto* (*s.s.*) typically causes Lyme arthritis and carditis, whereas both *B. afzelii* and *B. garinii* cause neuroborreliosis in Europe and Asia [9,20,21].

The infectious agent is transmitted to humans by several species of ticks from the *Ixodes* genus, whose distribution varies geographically. *Ixodes scapularis* and *Ixodes pacificus* are the most prevalent in North America, *Ixodes ricinus* is the most prevalent in Europe, and *Ixodes persulcatus* is the Lyme disease vector in Asia [16,18,22-24]. Tick vectors progress through sequential life stages: egg, larva, nymph, and adult. Ticks feed on hosts of different sizes throughout their growth stages. Specifically, nymphs primarily feed on rodents (especially *Peromyscus leucopus*, also known as white-footed mice), whereas adult ticks prefer larger mammals such as white-tailed deer (*Odocoileus virginianus*) [25].

Lyme disease has been called “the great imitator” in the literature because its clinical spectrum mimics various other unrelated diseases, making the correct diagnosis of Lyme disease based solely on clinical manifestations a difficult task, which can lead to misdiagnosis and mistreatment [26]. Lyme disease typically presents in 3 stages: early localized stage, early disseminated stage, and late disseminated stage [27,28]. The most common and usually first symptom of the early localized stage is a nonpruritic and painless rash with an erythematous center called erythema migrans (also known as the “bull’s-eye”) [27]. This symptom is present in nearly 90% of all Lyme disease cases and is accompanied by flu-like symptoms, including fever, headache, fatigue, adenopathy (lymph node), myalgia, and arthralgia [29]. The second stage is characterized by multiple skin and organ lesions, occurring months after exposure to the infected tick bite [26]. The heart, joints, and skin are the most affected organs [30-32]. The symptoms in the second stage include carditis (heart block, myocarditis, syncope, palpitations, dyspnea, and chest pain) and arthritis, which are most common in North America [31,33]. The late disseminated stage is manifested predominantly by neurological symptoms (radiculopathy, neck stiffness, meningitis, facial nerve palsy, cranial neuropathy, etc) [34,35]. Acrodermatitis chronica atrophicans and borrelial lymphocytoma are rare cutaneous manifestations of the third stage, and they are mostly noted in Europe and Asia [36].

The standard laboratory diagnosis of Lyme disease involves a 2-tier test in which an initial ELISA (enzyme-linked immunosorbent assay) screening test result is confirmed later by a western blot or an immunoblot [37]. Lyme disease is treatable with a short course of antibiotics, but if left untreated, it may lead to severe neurological, cardiac, and articular complications [38]. There is currently no vaccine against Lyme disease, and therefore, the only preventive measures are self-protection against tick bites and yard management [39].

Surveillance is one of the public health tasks aiming to monitor trends in disease epidemiology, identify populations at risk, and report disease cases [40,41]. Surveillance systems are based on active or passive surveillance approaches. Active surveillance is a surveillance system based on periodic collection of samples or case reports from health authorities, whereas passive surveillance is a system based on reporting of clinical suspect cases to the health authorities and depends on patient willingness to seek medical attention [40,42]. In North America (both the United States and Canada), Lyme disease reporting is compulsory, and the task falls on busy health care professionals to do so promptly [43,44]. In comparison, Lyme disease reporting is not mandatory in all endemic countries in Europe; however, the European Union recently called to standardize Lyme disease reporting and make it a notifiable disease

[41]. Underreporting is a concern in Lyme disease epidemiology because the traditional surveillance system has failed to track all cases accurately [45-47]. 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 only 30,000 cases [45,48]. The United States is not the only country where underreporting of Lyme disease cases has been suggested, as this issue has been pointed out in some European countries as well [49-51].

According to a review conducted previously [26], the traditional Lyme disease surveillance system is prone to overreporting or underreporting due to multiple reasons. One reason is that the system is dependent on the reporting of cases by busy health care professionals, and therefore, only cases seen and diagnosed by professionals are reported. Another contributor to the deficiency of the Lyme disease surveillance system is the lack of accuracy of serologic tests for Lyme disease diagnosis. The clinical diagnosis of Lyme disease is based on clinical manifestations, appropriate serology, and a history of exposure to tick bites [35]. The interpretation of the results of serologic tests as positive indicators of Lyme disease is however problematic since these tests are not very sensitive in the early stage and can show false-negative results, thereby rendering treatment ineffective [52,53]. Therefore, some cases tend to get missed by health care professionals, especially in new areas, resulting in underreporting of the disease [37,48,54,55]. Additionally, the heterogeneity of Lyme disease bacterial strains contributes to the late diagnosis of Lyme disease cases [56,57]. Moreover, Lyme disease monitoring also depends on data collected from tick surveillance. Tick data can be collected through passive surveillance, which can provide insights about risk areas for tick-borne diseases, such as Lyme disease, and active tick surveillance can identify regions where tick populations are established [58,59]. In countries, such as China, where Lyme disease is not yet endemic and is not a notifiable disease, active tick surveillance is used to monitor Lyme disease cases and quantify infection risk [60].

With the extended use of the internet and social media platforms where health-related information is often shared, researchers have found an opportunity to improve disease surveillance systems by leveraging web data [61,62]. This new field of research is referred to as digital surveillance or infodemiology [63]. Among all current social media platforms, Twitter is one of the most popular social media platforms, with over 145 million daily active accounts, and it is the most widely used data source for digital health owing to certain advantages: its data can be easily accessed through the Twitter application programming interface (API), the size of the text (tweets) is limited to 280 characters (140 characters before 2017), and it is possible to geolocate the tweets [64]. A recent systematic review suggested that less than half of existing studies on digital health surveillance using Twitter data were focused mainly on prediction, and only a few studies focused on developing tools for adequate analysis of these types of data [61]. Owing to the novelty of digital surveillance in public health research, there is an unevenness in the different methodological approaches and data sets used [63]. Given the availability and the richness of text data from Twitter or from other platforms (such as Reddit), there is a need to develop reliable and accurate classification methods to process and analyze the data to study health-related issues [62]. Specifically, the development and evaluation of methodological machine learning approaches are often required for data analysis. In addition to a significant time investment, these activities usually require access to organized, validated, pretrained, and labeled data sets for various health problems to facilitate their development.

Several studies have used data from search engines and social media platforms to track Lyme disease [65-67]. For example, a previous study examined how the content of Lyme disease videos on YouTube differed depending on data sources and the people who produced the videos [65]. 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 a previous study used digital data such as Google Trends [66]. While the official reported incidence of Lyme disease correlates well with Google Trends data, it did not significantly increase the forecasting accuracy. In another study, 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, resulting in the identification of a single hotspot in eastern Ontario [67]. 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 previous studies, the number of Lyme disease searches in search engines was related to seasonal and geographic patterns of Lyme disease cases [68,69]. However, there are very few studies focusing on Lyme disease and social media. For example, a previous study showed that Twitter can be used to monitor Lyme disease through the use of Twitter data (tweets) as a proxy for monitoring disease prevalence in the United Kingdom and the Republic of Ireland [70]. A limited geographic search strategy was used to discover spatial patterns and find rare cases of Lyme disease. Another study reported that Lyme-relevant Twitter data are correlated with official reports on the disease in the United States [71].

Our study aims to fill the gap on the use of web data to study Lyme disease in a very useful way. Indeed, our study seeks to provide an accurate English worldwide tweet data set and evaluate the performance of some selected natural language processing (NLP) transformer-based models, including DistilBERT (distilled version of BERT [Bidirectional Encoder Representations from Transformers]), ALBERT (A Lite BERT), and BERTweet (BERT for English Tweets), by integrating emotional component emojis. We believe that the novelty and completeness of this data set will assist in the development and evaluation of digital Lyme disease surveillance systems and will be a useful resource for public health researchers and practitioners. It is important to mention that this 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 [71]. However, unlike this previous study, our work here provides a worldwide data set 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 new and more accurate classified data for Lyme-related tweets.

This study has the following objectives:

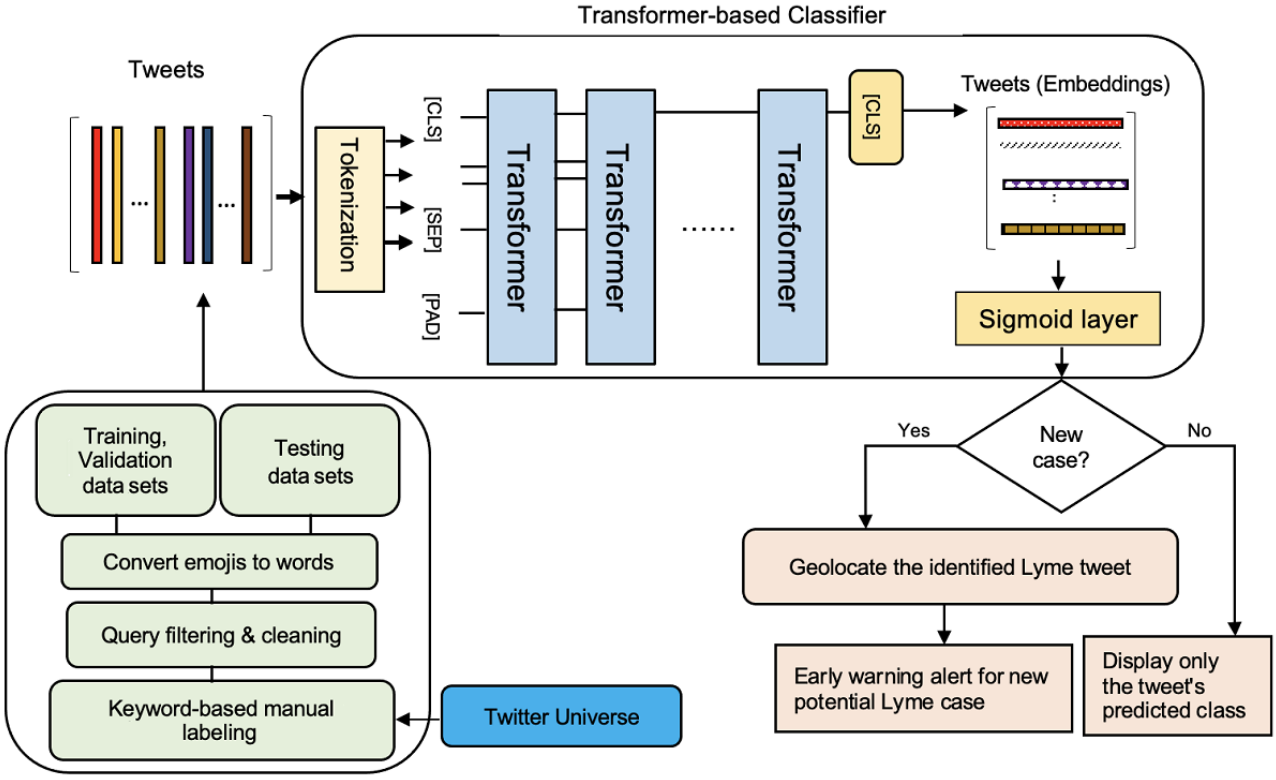
1. Provide an openly available data set to the scientific community for its use in a variety of experimental epidemiological research, at a time when there is an urgent need to integrate novel web-based data sets related to the Lyme disease epidemic (such as this classified data set) with other data sets from other sources for improving risk prediction.
2. Analyze the performance of several prominent NLP classifiers in terms of their ability to predict potential cases of Lyme disease.
3. Evaluate the effect of incorporating emojis as enrichment features to improve the performance of the transformer-based classifier.
4. Determine whether specific patterns could be identified regarding the prevalence of Lyme disease for each country based on the classified tweets.

Methods

Overview

Owing to the nature of Twitter data, the analysis requires developing and evaluating machine learning–based methodologies [61,72]. Figure 1 illustrates the methodology used to classify the tweets. This process consists of the following 2 key elements: (1) collecting and preprocessing self-reported Lyme-related tweets and (2) identifying potential Lyme-disease cases.

Figure 1. The 2-stage approach proposed for predicting potential Lyme disease cases. The first stage involves 4 elements: (1) We used standard search terms to collect tweets via the Twitter application programming interface; (2) We cleaned the tweets by removing hashtags, URL links, HTML markups, and stop-words; (3) We manually labeled the tweets as Lyme or non-Lyme using a list of precise keywords; and (4) We converted emojis into sentiment words, which were then substituted for the emojis in the tweets. In the second stage, we used a transformer-based classifier to determine whether a tweet is a potential Lyme disease case or not. When a new tweet was assigned with the highest probability to the Lyme disease class, we used the GeoPy library to estimate the tweet's location. The 3 special tokens were as follows: [CLS], which stood for classification and was typically the first token of every sequence; [SEP], which described to the pretrained 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.



Data Collection and Preprocessing

Using an academic research account with Twitter's API and search terms like "#Lyme" and "#Lyme disease," about 20,000 English tweets were collected between 2010 and 2022. Tweets were cleaned to reduce text noise and redundancy by deleting hashtags, URL links, HTML markups, stop-words, username mentions, and retweets. 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 impact the results and the quality of surveillance. Many studies have attempted to improve the relevance of disease-related keywords by examining word frequency using a corpus of tweet text and labeling approaches [72,73]. As such, we compiled a list of precise keywords that are often associated with Lyme disease. These keywords were used as the basis of regular expression that was applied to the cleaned tweets to manually determine whether they were relevant to Lyme disease.

A label "1" was assigned to potential Lyme disease–related tweets, while "0" was assigned to those that were not related. As mentioned in a previous report [71], we used 2 methods to specify keywords in regular expression. The first method 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*, and *tickborne*. By using this method of keyword selection, Twitter posts like "She is terribly unwell, we suspect it's Lyme" were labeled as potential Lyme disease cases. The second method of keyword selection involved considering the most frequent Lyme disease symptoms, transmission channels, or scientific terms, such 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 heartbeat*, *shooting pains*, *skin redness*, *tick bite*, and *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 Lyme disease cases. It is important to note that this manual approach aims to ensure fair and accurate labeling of the data set. This is because automatically labeling tweets with off-the-shelf Python regular expression libraries does not always provide correctly labeled tweets. It could also be argued that there are differences between the 2 keyword selection methods we used to manually label the tweets. However, the rationale here was based on the fact that the use of "Lyme disease" keywords in search engines has been demonstrated to improve results [68]. Therefore, the first method of keyword selection can be viewed as a naive general way to label any Lyme disease data collected from web-based sources, whereas the second method is a more specific and accurate way to confidently label tweets as related to Lyme disease. We would like to point out that, due to a lack of resources, each manually labeled tweet was only reviewed by 1 person. However, all reviewers agreed on a guideline, and several examples were provided to simplify understanding of the categories and reduce misclassification.

The labeled data set of 20,000 cleaned tweets was split into 3 disjoint data sets. Initially, a training data set (n=12,000) was compiled by randomly selecting exactly 1000 tweets from each year between 2010 and 2022. All tweets were distributed between 2 classes based on prior manual classification: potential Lyme disease cases (n=6000) and non-Lyme disease cases (n=6000). The remaining tweets (n=8000) were then separated into 2 equal parts: a validation data set (n=4000) and a testing data set (n=4000).

Detecting Lyme Disease Tweets Using Transformer-Based Classifiers

The training and validation data sets were used to fine-tune a set of pretrained transformer-based classifiers so that they could identify whether a new unknown tweet is a potential Lyme disease case. Recently, transformer-based models have been highly efficient in various NLP applications. Specifically, the BERT model [74], which was 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 data sets.

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 3 special tokens: [CLS], which stood for classification and was typically the first token of every sequence; [SEP], which described to the pretrained 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 [75]. When a token sequence exceeded the maximum length, it was truncated. Several variants of BERT classifiers have been proposed, but only the 3 most efficient ones were considered in this study: ALBERT, DistilBERT, and BERTweet [76], 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, a 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 [77] is a small and computationally efficient form of BERT. It is 60% faster than the $BERT_{base}$ model but 40% smaller owing to knowledge distillation during pretraining, all while achieving 97% of its language understanding efficiency. Compared with 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 66 million unique parameters in total. Furthermore, because DistilBERT does not require token type IDs, 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 [78] is a recent large-scale artificial intelligence 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 (transformer blocks) with a hidden size of 768 and a total of 110 million unique parameters. The model's creators produced 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 RoBERTa$_{base}$ and XLM-R$_{base}$ [78].

Embedding Enhancement

Emojis are used to express emotions succinctly and are popular communication tools 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 the feature embedding of a tweet is to use any emoji package, such as demoji, to convert emoji icons into sentiment words. The corresponding sentiment words are then substituted for emoji icons inside the tweets, resulting in tweets consisting of only word sequences that can be fed as input to the tokenizers associated with the BERT-based models.

Ethical Considerations

The use of tweets for academic research purposes is provided for in Twitter's development policy and the consent form signed by users [61]. Social media data are publicly accessible data. However, in accordance with Twitter's terms of use and to protect users' privacy, all personal information and tweets were deleted. There is no path or link from this paper (or any supplementary material related to this paper) to any individual tweets, users, or IDs.

Results

We initially compared the accuracy of the NLP classifier models (ALBERT, BERTweet, and DistilBERT) for detecting potential Lyme and non-Lyme disease tweets with the following state-of-the-art classification models: AdaBoost [79], random forest (RF) [80], logistic regression (LR) [81], Multilayer Perceptron Neural Network (MLP) [82], support vector machine (SVM) [83], k-nearest neighbors (KNN) [84], Quadratic Discriminant Analysis (QDA) [85], and Naive Bayes (NB) [82]. Using the term frequency-inverse document frequency (TF-IDF) vectorization method [86], the tweet embeddings were generated and then fed into the classifiers, except for the 3 transformer-based classifiers associated with their tokenizers. As reported previously [71], 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 [62,64]. 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 [74,87]. In order to maximize the likelihood estimation, we also evaluated the loss function by implementing binary cross-entropy [71]. As mentioned previously [71], 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 data sets. Specifically, after combining the training and validation data sets, we used 10-fold cross-validation to train the underlying classification models. Thus, 9 of the 10 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 data set to capture the following quantities: (1) the proportion of actual Lyme disease tweets correctly classified as potential Lyme disease cases (ie, true positives); (2) the proportion of actual non-Lyme tweets correctly classified as unrelated to Lyme disease (ie, true negatives); (3) the proportion of actual non-Lyme disease tweets incorrectly classified as belonging to the potential Lyme disease class (ie, false positives); and (4) the proportion of actual potential Lyme disease tweets misclassified as non-Lyme disease tweets (ie, false negatives). We computed several evaluation metrics based on confusion matrices to assess the accuracy of all tested classifiers as follows: classification accuracy [88], which measures the proportion of correct predictions (true positives and true negatives) among all examined tweets; average F1-score [89], 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.

As shown in Table 1, the BERTweet model was the best among all the NLP models included in our study. This model had the highest classification accuracy of 90.0%, average F1-score of 89.3%, 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 a classification accuracy of 76.6% and an F1-score of 76.7%. The classification accuracy scores of the QDA, RF, and AdaBoost models were comparable to those of the LR model, with both RF and AdaBoost having slightly more false negatives and fewer false positives than LR. A false negative was identified with a recall score as low as 62.7%, while a false positive was identified with a precision score as high as 96.5%. AdaBoost was slightly ahead of MLP but comparable to RF, as both AdaBoost and RF 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.

Table 1. Classification accuracy, average F1-score, precision, and recall for all classification models on the test data set.

Model	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
TF-IDF[a] 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 Multilayer Perceptron Neural Network	76.5	75.9	96.9	62.3[b]
TF-IDF and support vector machine	76.5	76.5	93.4	64.8
TF-IDF and k-nearest neighbors	71.8[b]	79.6	69.4[b]	93.2[c]
TF-IDF and Quadratic Discriminant Analysis	76.6	76.6	93.5	64.8
TF-IDF and Naive Bayes	73.7	75.6[b]	83.7	68.9
DistilBERT[d]	89.2	88.2	96.8	81.0
ALBERT[e]	88.4	87.3	96.6	79.7
BERTweet[f]	90.0[c]	89.3[c]	97.1[c]	82.6

[a]TF-IDF: term frequency-inverse document frequency.

[b]Lowest score value.

[c]Highest score value.

[d]DistilBERT: distilled version of Bidirectional Encoder Representations from Transformers.

[e]ALBERT: A Lite Bidirectional Encoder Representations from Transformers.

[f]BERTweet: Bidirectional Encoder Representations from Transformers for English Tweets.

Notably, KNN had the lowest precision score of 69.4%, producing significantly more false positives than any of the other classifiers tested. However, it also had the highest recall score of 93.2%, providing significantly fewer false negatives. When compared with the QDA, SVM, LR, and AdaBoost models, the NB classifier recall score of 68.9% showed slightly fewer true negatives, and its precision score of 83.7% demonstrated significantly more false positives. Overall, and apart from the transformer-based classifiers, QDA had the most consistent performance when all metrics were considered at once, with a classification accuracy of 76.5%, F1-score of 76.5%, precision of 93.4%, and recall of 64.8%.

Second, we investigated whether the inclusion of emojis improves the contextual encoding and classification of tweets. As described in the Methods section, 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, BERTweet still outperformed the other tested variants of the BERT classification model, with the highest classification accuracy of 95.2%, average F1-score of 94.9%, 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 emojis, and DistilBERT and ALBERT with emojis had recall scores that were at least 9% higher than their recall scores without emojis. The 3 classifiers were also able to reduce the produced false positives by at least 5% when emojis were used. As a result, DistilBERT had a significantly higher F1-score of 93.8% and accuracy of 94.1%, while ALBERT had a higher F1-score of 93% and accuracy of 93.9%. These results are summarized in Table 2.

Table 2. Classification accuracy, average F1-score, precision, and recall for the transformer-based classification models on the test data set after including emojis.

Model	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
BERTweet[a]	95.2	94.9	98.8	91.2
ALBERT[b]	93.9	93.1	97.3	89.2
DistilBERT[c]	94.1	93.8	97.5	90.4

[a]BERTweet: Bidirectional Encoder Representations from Transformers for English Tweets.

[b]ALBERT: A Lite Bidirectional Encoder Representations from Transformers.

[c]DistilBERT: distilled version of Bidirectional Encoder Representations from Transformers.

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 United States, the United Kingdom, Canada, and Australia had the highest number of potential Lyme disease–related tweets and non-Lyme disease tweets, accounting for 97.1%

(19,418/20,000) of the total. Remarkably, there were observed spikes in both Lyme disease and non-Lyme disease tweet counts for the United States, as the United States is a hotspot country for Lyme disease. Overall, the greatest proportion of potential Lyme disease–related tweets were from the United States (9827/20,000, 49.1%), whereas 0.2% (43/20,000) were reported from Canada, 0.03% (6/20,000) from Mexico, and 0.01% (2/20,000) from some Caribbean countries, such as Haiti and Jamaica. A total of 0.03% (6/20,000) of the potential Lyme disease–related tweets were reported from some South American countries, including Argentina and Venezuela. Potential Lyme disease cases reported in Europe were from Belgium, Denmark, Estonia, France, Ireland, Luxembourg, Norway, Poland, Sweden, and Switzerland, and represented 0.7% (143/20,000) of all tweets, while 0.3% (55/20,000) were from the United Kingdom. Potential Lyme disease cases reported in Asia were from Indonesia, Iran, the Philippines, South Korea, Taiwan, Thailand, and Vietnam, and represented 0.08% (15/20,000) of the data set. Potential Lyme disease cases from Africa represented 0.005% (1/20,000) of the data set and came from a single country, Sudan. Finally, New Zealand and Australia had 0.09% (18/20,000) of the total potential Lyme disease cases on Twitter, each accounting for 0.02% (4/20,000) and 0.07% (14/20,000), respectively.

Table 3 presents 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, were rarely reported. The classified data are available in the GitHub repository of this study [90].

Table 3. Top medical symptoms of Lyme disease reported in tweets.

Medical symptoms	Lyme disease–related tweet count
Rash	167
Fatigue	165
Tick bite	130
Fever	109
Arthritis	103
Sleepy	78
Migraine	48
Depression	48
Headaches	47
Carditis	25
Joint pain	20
Memory loss	11
Erythema migrans	8
Nausea	6
Nerve pain	6
Dizziness	5
Vomiting	4
Tingling	3
Palpitation	3
Chills	3
Numbness	3
Lymph nodes	3
Irregular heartbeat	2
Neck stiffness	1
Borrelial lymphocytoma	1

Discussion

Principal Findings

The empirical results of this study highlight the improved performance of transformer-based classifiers (ie, BERTweet, DistilBERT, and ALBERT), which we attribute to the following reasons. First, 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 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. Second, unlike AdaBoost, KNN, and RF, transformer-based classifiers are less affected by noise and redundant words in tweets. Third, transformer-based classifiers can learn nonlinear relationships and complex patterns in tweets because their neural network architecture does not assume linearity between dependent and independent features (unlike LR and SVM), and nonlinearity between the features is often the case in extracted tweets. Finally, transformer-based classifiers differ from both MLP and KNN because they can efficiently handle feature scaling and frequently converge to the global optimum rather than getting stuck in the local minima. This is due to the optimization of the cross-entropy loss function, which is often convex for most weights.

The results also showed that emojis are effective enrichment features to improve the accuracy of tweet embedding. The performance of transformer-based classifiers can be further improved by considering the sentimental semantics of emojis. Since the texts of non-Lyme disease and Lyme disease–related tweets can be similar in some cases, emojis can play important roles in better identifying Lyme disease–related tweets. When emojis are removed during the text cleaning process, some potential 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 the 20,000 tweets used in this study showed a high volume of potential Lyme disease–related tweets in the United States, the United Kingdom, and Canada. This may be due to 2 reasons. First, Lyme disease is spreading, and second, a focus solely on English tweets may limit the collection of tweets from non-English speaking countries. Furthermore, in the case of symptoms, borrelial lymphocytoma, palpitations, tingling, nausea, and neck stiffness are rarely reported. 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 clinical symptom, is not among the most common symptoms reported on Twitter because the general population tends to refer to this symptom as a rash. These findings correlate with the geographic distribution of the clinical manifestations of Lyme disease throughout the literature [46,47,91].

Lyme disease is endemic in both the United States and Europe. Although CDC surveillance has reported over 30,000 cases annually, other studies have estimated 476,000 cases yearly in the United States [10,47,48]. In Europe, over 200,000 cases of Lyme disease are being reported yearly [92,93]. The results of our study are in line with the literature since Lyme disease–related tweets originate mainly from the United States. However, it is difficult to compare these results, given that only English-language tweets were used in this study and that the distribution of Twitter users is related to geographical location. Our study is the first to provide a pretrained, organized, and labeled Lyme disease–related data set with an emoji component, which can be used to quantify and compare the performance of different methodological approaches in future Lyme disease–related work. This will allow for consistency in future research and improve digital surveillance of Lyme disease. For example, a sudden increase in Lyme disease–related activity on Twitter or other social media platforms may indicate the beginning of an increase in cases, which could justify the promotion of tick bite prevention measures in the indicated geographical area. For example, we found studies reporting erythema migrans in combination with *Borrelia burgdorferi* sensu lato antibodies after tick bite in some patients from regions in the Caribbean, who do not have any travel history to Lyme disease–endemic regions [94,95]. Despite the controversy surrounding the presence of Lyme disease vectors in the Caribbean [94], these findings agree with our study results that showed some Lyme disease–related tweets in the region, suggesting a need for further investigation on the presence of Lyme disease vectors in the Caribbean. Finally, to be able to compare our results to other early warning systems, we extracted Lyme disease reports on ProMED (which is the largest early warning system for emerging diseases in the world). A comparison of the Lyme disease cases identified in these reports with our data revealed some similar results. In fact, the majority of ProMED Lyme disease reports came from the United States (57%); Canada (25%); the United Kingdom (7%); and some European countries, such as Finland, Belgium, and Austria (9%). These results are in line with our findings, with some disparities in terms of ranking. Indeed, in our results, the United Kingdom ranks second for potential Lyme disease cases, while Canada ranks third. However, the trend holds for the United States and other European countries.

Limitations

Our study has several limitations. First, the collection of Twitter data through the Twitter API is suggested to have a selection bias, since only 1% of the data are accessible. Furthermore, because the data collected are randomly generated, the data may not reflect the reality of Lyme disease conversations on twitter. This study aimed to limit this bias by collecting data over a long period of time. Additionally, social media data are highly susceptible to media coverage, so tweets about Lyme disease may be driven by media coverage rather than disease incidence.

While we provided a pretrained Lyme disease–related tweet data set, our results only reflected Lyme disease–related English tweets. As we excluded tweets published in languages other than English from our data set, we were not able to access all potentially relevant Twitter discussions about Lyme disease from non-English speaking countries where Lyme disease is endemic.

Although emojis have general meanings, their usage mainly depends on other factors, such as cultural background, linguistic factors, and gender [96]. Since our study only focused on the sentimental semantics of emojis, our model may have

erroneously assigned a certain meaning to emojis different from 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 emojis did not significantly affect the performance of our model.

While we were able to geolocate the tweets, users may not register with their exact location or may register with a wrong location due to safety concerns [97]. To reduce such bias, we did not map granular tweet-specific locations, but rather expanded the spatial distribution to the country level to reduce the risk of location errors.

One common limitation of using social media data is that tweeting does not necessarily equate to the occurrence of Lyme disease [63,98,99]. Thus, our model may have included tweets about Lyme disease but not actual cases of Lyme disease. However, we believe that the model has been well-trained with various keywords related to Lyme disease, therefore improving the performance of BERT transformer models.

Additionally, according to Marques and other collaborators, *Ixodes* ticks can carry two or more pathogens and are capable of transmitting them in a single bite, thereby resulting in co-infection [4,10,100]. Given the increased public awareness of Lyme disease compared to that of other tick-borne diseases, the public use of Lyme disease as an umbrella term to describe any tick-borne disease may confound the results of this study. Specifically, Lyme disease–related Twitter discussions in endemic tick-borne disease regions may not actually be solely associated with potential Lyme disease cases when there is a risk of infection by non-Lyme disease–related pathogens [26]. Therefore, when interpreting the results of our study, these limitations should be considered.

Conclusions

The early detection of potential Lyme disease cases is essential to limit its increase 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 novel data for Lyme disease prediction and surveillance 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 potential cases of Lyme disease. While Twitter data were the focus of this work, the proposed system can be easily adapted to other social media text–based platforms like Reddit. We suggest that future research should focus on collecting social media data from both English and non-English texts to improve the knowledge of potential Lyme disease cases, as some of the countries with the highest incidence of Lyme disease are non-English speaking countries. Additionally, as our model was able to identify some Lyme disease–related tweets from regions that typically have a low reported incidence of Lyme disease (ie, African countries and the Caribbean), we believe that the results are valuable for informing emerging Lyme disease surveillance activities in these geographical areas. Despite these limitations, our study provides a steady performance model with publicly available data for researchers and policymakers to identify trends in Lyme disease discussions on social media.

Acknowledgments

We would like to thank Ariel Mundo (a postdoctoral fellow at École de Santé Publique de l'Université de Montréal) and Tanya Philipssen (a master's student in the Department of Mathematics and Statistics at the University of Victoria) for checking the fluidity and grammar of our manuscript. We would also like to thank François Hu (a postdoctoral fellow at Université de Montréal), who is our collaborator on a project using ProMED data, for his help in collecting the reports from ProMED. In addition, we would like to thank the Editor, the Associate Editor, and the referees for their helpful comments. This work was supported by the Fonds de recherche du Québec Scholar Program (J1 in Artificial Intelligence and Digital Health, BN), the Natural Sciences and Engineering Research Council of Canada through the Discovery Grant Program (BN), the Mathematics for Public Health (MfPH) Emerging Infectious Diseases Modelling Initiative (BN), One Health Modelling for Emerging Infections (OMNI) (BN), and the Bourse d'Intelligence Artificielle from the Études supérieures et postdoctorales of Université de Montréal.

Data Availability

The data sets analyzed in our study are available on GitHub [90]. However, to protect and ensure Twitter users' privacy, we only share the location and labels.

Conflicts of Interest

None declared.

References

1. Rodino KG, Theel ES, Pritt BS. Tick-Borne Diseases in the United States. Clin Chem 2020 Apr 01;66(4):537-548 [doi: 10.1093/clinchem/hvaa040] [Medline: 32232463]
2. Boulanger N, Boyer P, Talagrand-Reboul E, Hansmann Y. Ticks and tick-borne diseases. Med Mal Infect 2019 Mar;49(2):87-97 [doi: 10.1016/j.medmal.2019.01.007] [Medline: 30736991]

3. Cutler SJ, Vayssier-Taussat M, Estrada-Peña A, Potkonjak A, Mihalca AD, Zeller H. Tick-borne diseases and co-infection: Current considerations. Ticks Tick Borne Dis 2021 Jan;12(1):101607 [doi: 10.1016/j.ttbdis.2020.101607] [Medline: 33220628]
4. Belongia EA. Epidemiology and impact of coinfections acquired from Ixodes ticks. Vector Borne Zoonotic Dis 2002;2(4):265-273 [doi: 10.1089/153036602321653851] [Medline: 12804168]
5. Wachter J, Martens C, Barbian K, Rego ROM, Rosa P. Epigenomic Landscape of Lyme Disease Spirochetes Reveals Novel Motifs. mBio 2021 Jun 29;12(3):e0128821 [FREE Full text] [doi: 10.1128/mBio.01288-21] [Medline: 34156261]
6. Stone BL, Tourand Y, Brissette CA. Brave New Worlds: The Expanding Universe of Lyme Disease. Vector Borne Zoonotic Dis 2017 Sep;17(9):619-629 [FREE Full text] [doi: 10.1089/vbz.2017.2127] [Medline: 28727515]
7. Bisanzio D, Fernández M, Martello E, Reithinger R, Diuk-Wasser MA. Current and Future Spatiotemporal Patterns of Lyme Disease Reporting in the Northeastern United States. JAMA Netw Open 2020 Mar 02;3(3):e200319 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0319] [Medline: 32125426]
8. Lyme disease in Canada, 2009 to 2019. Public Health Agency of Canada. URL: https://www.canada.ca/en/public-health/services/reports-publications/canada-communicable-disease-report-ccdr/monthly-issue/2022-48/issue-5-may-2022/lyme-disease-canada-2009-2019.html [accessed 2022-05-19]
9. Thurston R. Lyme disease. Work 2019;62(4):643-646 [doi: 10.3233/WOR-192897] [Medline: 31104048]
10. Marques AR, Strle F, Wormser GP. Comparison of Lyme Disease in the United States and Europe. Emerg Infect Dis 2021 Aug;27(8):2017-2024 [FREE Full text] [doi: 10.3201/eid2708.204763] [Medline: 34286689]
11. Vinayaraj EV, Gupta N, Sreenath K, Thakur CK, Gulati S, Anand V, et al. Clinical and laboratory evidence of Lyme disease in North India, 2016-2019. Travel Med Infect Dis 2021;43:102134 [doi: 10.1016/j.tmaid.2021.102134] [Medline: 34216802]
12. Önal U, Aytaç Erdem H, Uyan Önal A, Reşat Sipahi O. Systematic review of Lyme disease in Turkey. Trop Doct 2019 Jul;49(3):165-170 [doi: 10.1177/0049475519843387] [Medline: 31018773]
13. Bogoch II, Watts A, Thomas-Bachli A, Huber C, Kraemer MUG, Khan K. Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel. J Travel Med 2020 Mar 13;27(2):taaa008 [FREE Full text] [doi: 10.1093/jtm/taaa008] [Medline: 31943059]
14. Acharya D, Park J. Seroepidemiologic Survey of Lyme Disease among Forestry Workers in National Park Offices in South Korea. Int J Environ Res Public Health 2021 Mar 12;18(6):2933 [FREE Full text] [doi: 10.3390/ijerph18062933] [Medline: 33809327]
15. Pun SB, Agrawal S, Jha S, Bhandari LN, Chalise BS, Mishra A, et al. First report of Lyme disease in Nepal. JMM Case Rep 2018 Mar;5(3):e005128 [FREE Full text] [doi: 10.1099/jmmcr.0.005128] [Medline: 29623212]
16. Ji Z, Jian M, Yue P, Cao W, Xu X, Zhang Y, et al. Prevalence of in Ixodidae Tick around Asia: A Systematic Review and Meta-Analysis. Pathogens 2022 Jan 24;11(2):143 [FREE Full text] [doi: 10.3390/pathogens11020143] [Medline: 35215089]
17. Stanek G, Wormser GP, Gray J, Strle F. Lyme borreliosis. Lancet 2012 Feb 04;379(9814):461-473 [doi: 10.1016/S0140-6736(11)60103-7] [Medline: 21903253]
18. Coors A, Hassenstein MJ, Krause G, Kerrinnes T, Harries M, Breteler MMB, et al. Regional seropositivity for Borrelia burgdorferi and associated risk factors: findings from the Rhineland Study, Germany. Parasit Vectors 2022 Jul 04;15(1):241 [FREE Full text] [doi: 10.1186/s13071-022-05354-z] [Medline: 35786209]
19. Dehnert M, Fingerle V, Klier C, Talaska T, Schlaud M, Krause G, et al. Seropositivity of Lyme borreliosis and associated risk factors: a population-based study in Children and Adolescents in Germany (KiGGS). PLoS One 2012;7(8):e41321 [FREE Full text] [doi: 10.1371/journal.pone.0041321] [Medline: 22905101]
20. Atkinson SF, Sarkar S, Aviña A, Schuermann JA, Williamson P. A determination of the spatial concordance between Lyme disease incidence and habitat probability of its primary vector Ixodes scapularis (black-legged tick). Geospat Health 2014 Nov;9(1):203-212 [FREE Full text] [doi: 10.4081/gh.2014.17] [Medline: 25545937]
21. Barbour AG, Benach JL. Discovery of the Lyme Disease Agent. mBio 2019 Sep 17;10(5):e02166-19 [FREE Full text] [doi: 10.1128/mBio.02166-19] [Medline: 31530679]
22. Ogden NH, Bouchard C, Kurtenbach K, Margos G, Lindsay LR, Trudel L, et al. Active and passive surveillance and phylogenetic analysis of Borrelia burgdorferi elucidate the process of Lyme disease risk emergence in Canada. Environ Health Perspect 2010 Jul;118(7):909-914 [FREE Full text] [doi: 10.1289/ehp.0901766] [Medline: 20421192]
23. Cardenas-de la Garza JA, De la Cruz-Valadez E, Ocampo-Candiani J, Welsh O. Clinical spectrum of Lyme disease. Eur J Clin Microbiol Infect Dis 2019 Feb;38(2):201-208 [doi: 10.1007/s10096-018-3417-1] [Medline: 30456435]
24. Bouchard C, Beauchamp G, Nguon S, Trudel L, Milord F, Lindsay LR, et al. 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 2011 Dec;2(4):183-190 [doi: 10.1016/j.ttbdis.2011.03.005] [Medline: 22108010]
25. Eisen RJ, 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 2012 Jan 01;49(1):11-22 [doi: 10.1603/me11138] [Medline: 22308766]
26. Baker PJ. Is It Possible to Make a Correct Diagnosis of Lyme Disease on Symptoms Alone? Review of Key Issues and Public Health Implications. Am J Med 2019 Oct;132(10):1148-1152 [FREE Full text] [doi: 10.1016/j.amjmed.2019.04.001] [Medline: 31028718]

27. Aguero-Rosenfeld ME, Wang G, Schwartz I, Wormser GP. Diagnosis of lyme borreliosis. Clin Microbiol Rev 2005 Jul;18(3):484-509 [FREE Full text] [doi: 10.1128/CMR.18.3.484-509.2005] [Medline: 16020686]
28. Allehebi ZO, Khan FM, Robbins M, Simms E, Xiang R, Shawwa A, et al. Lyme Disease, Anaplasmosis, and Babesiosis, Atlantic Canada. Emerg Infect Dis 2022 Jun;28(6):1292-1294 [FREE Full text] [doi: 10.3201/eid2806.220443] [Medline: 35608954]
29. Van Hout MC. The Controversies, Challenges and Complexities of Lyme Disease: A Narrative Review. J Pharm Pharm Sci 2018;21(1):429-436 [FREE Full text] [doi: 10.18433/jpps30254] [Medline: 30458921]
30. Sanchez JL. Clinical Manifestations and Treatment of Lyme Disease. Clin Lab Med 2015 Dec;35(4):765-778 [doi: 10.1016/j.cll.2015.08.004] [Medline: 26593256]
31. Shen RV, McCarthy CA. Cardiac Manifestations of Lyme Disease. Infect Dis Clin North Am 2022 Sep;36(3):553-561 [doi: 10.1016/j.idc.2022.03.001] [Medline: 36116834]
32. Scheerer C, Rüth M, Tizek L, Köberle M, Biedermann T, Zink A. Googling for Ticks and Borreliosis in Germany: Nationwide Google Search Analysis From 2015 to 2018. J Med Internet Res 2020 Oct 16;22(10):e18581 [FREE Full text] [doi: 10.2196/18581] [Medline: 33064086]
33. Bockenstedt LK, Wormser GP. Review: unraveling Lyme disease. Arthritis Rheumatol 2014 Sep;66(9):2313-2323 [FREE Full text] [doi: 10.1002/art.38756] [Medline: 24965960]
34. Ross Russell AL, Dryden MS, Pinto AA, Lovett JK. Lyme disease: diagnosis and management. Pract Neurol 2018 Dec;18(6):455-464 [doi: 10.1136/practneurol-2018-001998] [Medline: 30282764]
35. Borchers AT, Keen CL, Huntley AC, Gershwin ME. Lyme disease: a rigorous review of diagnostic criteria and treatment. J Autoimmun 2015 Feb;57:82-115 [doi: 10.1016/j.jaut.2014.09.004] [Medline: 25451629]
36. Müllegger R, Glatz M. Skin manifestations of lyme borreliosis: diagnosis and management. Am J Clin Dermatol 2008;9(6):355-368 [doi: 10.2165/0128071-200809060-00002] [Medline: 18973402]
37. Kullberg BJ, Vrijmoeth HD, van de Schoor F, Hovius JW. Lyme borreliosis: diagnosis and management. BMJ 2020 May 26;369:m1041 [doi: 10.1136/bmj.m1041] [Medline: 32457042]
38. Donta ST. What We Know and Don't Know About Lyme Disease. Front Public Health 2021;9:819541 [FREE Full text] [doi: 10.3389/fpubh.2021.819541] [Medline: 35127630]
39. 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 2020 Oct 20;20(1):1575 [FREE Full text] [doi: 10.1186/s12889-020-09629-x] [Medline: 33081728]
40. Villalonga-Olives E. Foundations of Epidemiology. Int J Epidemiol 2017;46(1):370-371 [doi: 10.1093/ije/dyw272]
41. 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 2022 Jul 07;22(1):1307 [FREE Full text] [doi: 10.1186/s12889-022-13669-w] [Medline: 35799156]
42. Hadorn DC, Stärk K. Evaluation and optimization of surveillance systems for rare and emerging infectious diseases. Vet Res 2008;39(6):57 [FREE Full text] [doi: 10.1051/vetres:2008033] [Medline: 18651991]
43. White J, Noonan-Toly C, Lukacik G, Thomas N, Hinckley A, Hook S, et al. Lyme Disease Surveillance in New York State: an Assessment of Case Underreporting. Zoonoses Public Health 2018 Mar;65(2):238-246 [doi: 10.1111/zph.12307] [Medline: 27612955]
44. Boulanger V, Poirier É, MacLaurin A, Quach C. Divergences between healthcare-associated infection administrative data and active surveillance data in Canada. Can Commun Dis Rep 2022 Jan 26;48(1):4-16 [FREE Full text] [doi: 10.14745/ccdr.v48i01a02] [Medline: 35273464]
45. Schwartz AM, Kugeler KJ, Nelson CA, Marx GE, Hinckley AF. Use of Commercial Claims Data for Evaluating Trends in Lyme Disease Diagnoses, United States, 2010-2018. Emerg Infect Dis 2021;27(2):499-507 [FREE Full text] [doi: 10.3201/eid2702.202728] [Medline: 33496238]
46. Nelson CA, Saha S, Kugeler KJ, Delorey MJ, Shankar MB, Hinckley AF, et al. Incidence of Clinician-Diagnosed Lyme Disease, United States, 2005-2010. Emerg Infect Dis 2015 Sep;21(9):1625-1631 [FREE Full text] [doi: 10.3201/eid2109.150417] [Medline: 26291194]
47. Kugeler KJ, Schwartz AM, Delorey MJ, Mead PS, Hinckley AF. Estimating the Frequency of Lyme Disease Diagnoses, United States, 2010-2018. Emerg Infect Dis 2021 Feb;27(2):616-619 [FREE Full text] [doi: 10.3201/eid2702.202731] [Medline: 33496229]
48. Lyme Disease - Data and Surveillance. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/lyme/datasurveillance/index.html [accessed 2022-06-22]
49. Petrulionienė A, Radzišauskienė D, Ambrozaitis A, Čaplinskas S, Paulauskas A, Venalis A. Epidemiology of Lyme Disease in a Highly Endemic European Zone. Medicina (Kaunas) 2020 Mar 05;56(3):115 [FREE Full text] [doi: 10.3390/medicina56030115] [Medline: 32151097]
50. Steinbrink A, Brugger K, Margos G, Kraiczy P, Klimpel S. The evolving story of Borrelia burgdorferi sensu lato transmission in Europe. Parasitol Res 2022 Mar;121(3):781-803 [FREE Full text] [doi: 10.1007/s00436-022-07445-3] [Medline: 35122516]

51. van den Wijngaard C, Hofhuis A, Simões M, Rood E, van Pelt W, Zeller H, et al. Surveillance perspective on Lyme borreliosis across the European Union and European Economic Area. Euro Surveill 2017 Jul 06;22(27):30569 [FREE Full text] [doi: 10.2807/1560-7917.ES.2017.22.27.30569] [Medline: 28703098]
52. Mead PS. Epidemiology of Lyme disease. Infect Dis Clin North Am 2015 Jun;29(2):187-210 [doi: 10.1016/j.idc.2015.02.010] [Medline: 25999219]
53. Mead P. Epidemiology of Lyme Disease. Infect Dis Clin North Am 2022 Sep;36(3):495-521 [doi: 10.1016/j.idc.2022.03.004] [Medline: 36116831]
54. John TM, Taege AJ. Appropriate laboratory testing in Lyme disease. Cleve Clin J Med 2019 Nov;86(11):751-759 [FREE Full text] [doi: 10.3949/ccjm.86a.19029] [Medline: 31710588]
55. Pace E, O'Reilly M. Tickborne Diseases: Diagnosis and Management. Am Fam Physician 2020 May 01;101(9):530-540 [FREE Full text] [Medline: 32352736]
56. Brownstein JS, Skelly DK, Holford TR, Fish D. Forest fragmentation predicts local scale heterogeneity of Lyme disease risk. Oecologia 2005 Dec;146(3):469-475 [doi: 10.1007/s00442-005-0251-9] [Medline: 16187106]
57. Rudenko N, Golovchenko M, Grubhoffer L, Oliver JH. Updates on Borrelia burgdorferi sensu lato complex with respect to public health. Ticks Tick Borne Dis 2011 Sep;2(3):123-128 [FREE Full text] [doi: 10.1016/j.ttbdis.2011.04.002] [Medline: 21890064]
58. Ogden NH, Koffi JK, Pelcat Y, Lindsay LR. Environmental risk from Lyme disease in central and eastern Canada: a summary of recent surveillance information. Can Commun Dis Rep 2014 Mar 06;40(5):74-82 [FREE Full text] [doi: 10.14745/ccdr.v40i05a01] [Medline: 29769885]
59. Chilton NB, Curry PS, Lindsay LR, Rochon K, Lysyk TJ, Dergousoff SJ. Passive and Active Surveillance for Ixodes scapularis (Acari: Ixodidae) in Saskatchewan, Canada. J Med Entomol 2020 Jan 09;57(1):156-163 [doi: 10.1093/jme/tjz155] [Medline: 31618432]
60. Zhao G, Wang Y, Fan Z, Ji Y, Liu M, Zhang W, et al. Mapping ticks and tick-borne pathogens in China. Nat Commun 2021 Feb 17;12(1):1075 [FREE Full text] [doi: 10.1038/s41467-021-21375-1] [Medline: 33597544]
61. Takats C, Kwan A, Wormer R, Goldman D, Jones HE, Romero D. Ethical and Methodological Considerations of Twitter Data for Public Health Research: Systematic Review. J Med Internet Res 2022 Nov 29;24(11):e40380 [FREE Full text] [doi: 10.2196/40380] [Medline: 36445739]
62. Ayinde BO, Zurada JM. Deep Learning of Constrained Autoencoders for Enhanced Understanding of Data. IEEE Trans Neural Netw Learn Syst 2018 Sep;29(9):3969-3979 [doi: 10.1109/TNNLS.2017.2747861] [Medline: 28961128]
63. Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons 2010 Jan;53(1):59-68 [doi: 10.1016/j.bushor.2009.09.003]
64. Mohr H, Ruge H. Fast Estimation of L1-Regularized Linear Models in the Mass-Univariate Setting. Neuroinformatics 2021 Jul;19(3):385-392 [FREE Full text] [doi: 10.1007/s12021-020-09489-1] [Medline: 32935193]
65. Basch CH, Mullican LA, Boone KD, Yin J, Berdnik A, Eremeeva ME, et al. Lyme Disease and YouTube : A Cross-Sectional Study of Video Contents. Osong Public Health Res Perspect 2017 Aug;8(4):289-292 [FREE Full text] [doi: 10.24171/j.phrp.2017.8.4.10] [Medline: 28904853]
66. Kapitány-Fövény M, Ferenci T, Sulyok Z, Kegele J, Richter H, Vályi-Nagy I, et al. Can Google Trends data improve forecasting of Lyme disease incidence? Zoonoses Public Health 2019 Feb;66(1):101-107 [doi: 10.1111/zph.12539] [Medline: 30447056]
67. Kutera M, Berke O, Sobkowich K. Spatial epidemiological analysis of Lyme disease in southern Ontario utilizing Google Trends searches. Environ. Health Rev 2021 Dec;64(4):105-110 [doi: 10.5864/d2021-025]
68. Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. Geospat Health 2010 May;4(2):135-137 [doi: 10.4081/gh.2010.195] [Medline: 20503183]
69. 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. Front Public Health 2020;8:395 [FREE Full text] [doi: 10.3389/fpubh.2020.00395] [Medline: 32923420]
70. Tulloch JSP, Vivancos R, Christley RM, Radford AD, Warner JC. Mapping tweets to a known disease epidemiology; a case study of Lyme disease in the United Kingdom and Republic of Ireland. J Biomed Inform 2019;100S:100060 [FREE Full text] [doi: 10.1016/j.yjbinx.2019.100060] [Medline: 34384577]
71. Boligarla S, Laison E, Li J, Mahadevan R, Ng A, Lin Y, et al. Leveraging Machine Learning Approaches for Predicting Potential Lyme Disease Cases and Incidence Rates in United States Using Twitter. Research Square. 2022. URL: https://www.researchsquare.com/article/rs-2136402/v1 [accessed 2023-09-15]
72. 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 2019;14(7):e0210689 [FREE Full text] [doi: 10.1371/journal.pone.0210689] [Medline: 31318885]
73. Dias Canedo E, Cordeiro Mendes B. Software Requirements Classification Using Machine Learning Algorithms. Entropy (Basel) 2020 Sep 21;22(9):1057 [FREE Full text] [doi: 10.3390/e22091057] [Medline: 33286826]
74. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2019; Minneapolis, MN [doi: 10.18653/v1/N19-1423]
75. Benítez-Andrades J, Alija-Pérez J, Vidal M, Pastor-Vargas R, García-Ordás M. Traditional Machine Learning Models and Bidirectional Encoder Representations From Transformer (BERT)-Based Automatic Classification of Tweets About Eating Disorders: Algorithm Development and Validation Study. JMIR Med Inform 2022 Feb 24;10(2):e34492 [FREE Full text] [doi: 10.2196/34492] [Medline: 35200156]
76. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv. URL: https://arxiv.org/pdf/1909.11942.pdf [accessed 2023-09-15]
77. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv. URL: https://arxiv.org/pdf/1910.01108.pdf [accessed 2023-09-15]
78. Nguyen DQ, Vu T, Nguyen AT. BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: Conference on Empirical Methods in Natural Language Processing: System Demonstrations; October 2020; Online [doi: 10.18653/v1/2020.emnlp-demos.2]
79. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 1997 Aug;55(1):119-139 [doi: 10.1006/jcss.1997.1504]
80. Pal M. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 2007 Feb 22;26(1):217-222 [doi: 10.1080/01431160412331269698]
81. Krishnapuram B, Carin L, Figueiredo MAT, Hartemink AJ. Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans Pattern Anal Mach Intell 2005 Jun;27(6):957-968 [doi: 10.1109/TPAMI.2005.127] [Medline: 15943426]
82. Hastie T, Tibshirani R, Friedman J. Model Assessment and Selection. In: The Elements of Statistical Learning. Springer Series in Statistics. New York, NY: Springer; 2009:219-259
83. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297 [doi: 10.1007/BF00994018]
84. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans. Inform. Theory 1967 Jan;13(1):21-27 [doi: 10.1109/tit.1967.1053964]
85. Tharwat A. Linear vs. quadratic discriminant analysis classifier: a tutorial. IJAPR 2016;3(2):145 [doi: 10.1504/IJAPR.2016.079050]
86. Ramos J. Using TF-IDF to Determine Word Relevance in Document Queries. CiteSeerX. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c [accessed 2023-09-15]
87. Subakti A, Murfi H, Hariadi N. The performance of BERT as data representation of text clustering. J Big Data 2022;9(1):15 [FREE Full text] [doi: 10.1186/s40537-022-00564-9] [Medline: 35194542]
88. Kotsiantis S. Supervised Machine Learning: A Review of Classification Techniques. In: Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. Amsterdam, Netherlands: IOS Press; 2007.
89. Powers D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv. URL: https://arxiv.org/ftp/arxiv/papers/2010/2010.16061.pdf [accessed 2023-09-15]
90. nasrilab/Twitter_study. Github. URL: https://github.com/nasrilab/Twitter_study/tree/main/data [accessed 2023-09-15]
91. Alkishe A, Raghavan RK, Peterson AT. Likely Geographic Distributional Shifts among Medically Important Tick Species and Tick-Associated Diseases under Climate Change in North America: A Review. Insects 2021 Mar 05;12(3):225 [FREE Full text] [doi: 10.3390/insects12030225] [Medline: 33807736]
92. Stanek G, Strle F. Lyme disease: European perspective. Infect Dis Clin North Am 2008 Jun;22(2):327-39, vii [doi: 10.1016/j.idc.2008.01.001] [Medline: 18452805]
93. Stanek G, Strle F. Lyme borreliosis-from tick bite to diagnosis and treatment. FEMS Microbiol Rev 2018 May 01;42(3):233-258 [doi: 10.1093/femsre/fux047] [Medline: 29893904]
94. Gondard M, Cabezas-Cruz A, Charles RA, Vayssier-Taussat M, Albina E, Moutailler S. Ticks and Tick-Borne Pathogens of the Caribbean: Current Understanding and Future Directions for More Comprehensive Surveillance. Front Cell Infect Microbiol 2017;7:490 [FREE Full text] [doi: 10.3389/fcimb.2017.00490] [Medline: 29238699]
95. Sharma A, Jaimungal S, Basdeo-Maharaj K, Chalapathi Rao AV, Teelucksingh S. Erythema migrans-like illness among Caribbean islanders. Emerg Infect Dis 2010 Oct;16(10):1615-1617 [FREE Full text] [doi: 10.3201/eid1610.100587] [Medline: 20875293]
96. Bai Q, Dan Q, Mu Z, Yang M. A Systematic Review of Emoji: Current Research and Future Perspectives. Front Psychol 2019;10:2221 [FREE Full text] [doi: 10.3389/fpsyg.2019.02221] [Medline: 31681068]
97. Bisanzio D, Kraemer MUG, Brewer T, Brownstein JS, Reithinger R. Geolocated Twitter social media data to describe the geographic spread of SARS-CoV-2. J Travel Med 2020 Aug 20;27(5):taaa120 [FREE Full text] [doi: 10.1093/jtm/taaa120] [Medline: 32701135]

98. Aiello AE, Renson A, Zivich PN. Social Media- and Internet-Based Disease Surveillance for Public Health. Annu Rev Public Health 2020 Apr 02;41:101-118 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
99. Milinovich GJ, Avril SMR, Clements ACA, Brownstein JS, Tong S, Hu W. Using internet search queries for infectious disease surveillance: screening diseases for suitability. BMC Infect Dis 2014 Dec 31;14(1):690 [FREE Full text] [doi: 10.1186/s12879-014-0690-1] [Medline: 25551277]
100. Eisen RJ, Paddock CD. Tick and Tickborne Pathogen Surveillance as a Public Health Tool in the United States. J Med Entomol 2021 Jul 16;58(4):1490-1502 [FREE Full text] [doi: 10.1093/jme/tjaa087] [Medline: 32440679]

Abbreviations

ALBERT: A Lite Bidirectional Encoder Representations from Transformers
API: application programming interface
BERT: Bidirectional Encoder Representations from Transformers
BERTweet: Bidirectional Encoder Representations from Transformers for English Tweets
CDC: Centers for Disease Control and Prevention
DistilBERT: distilled version of Bidirectional Encoder Representations from Transformers
KNN: k-nearest neighbors
LR: logistic regression
MLP: Multilayer Perceptron Neural Network
NB: Naive Bayes
NLP: natural language processing
QDA: Quadratic Discriminant Analysis
RF: random forest
SVM: support vector machine
TF-IDF: term frequency-inverse document frequency

Edited by A Mavragani; submitted 13.03.23; peer-reviewed by S Maxwell, J Chen; comments to author 29.06.23; revised version received 26.07.23; accepted 31.08.23; published 16.10.23

Please cite as:
Laison EKE, Hamza Ibrahim M, Boligarla S, Li J, Mahadevan R, Ng A, Muthuramalingam V, Lee WY, Yin Y, Nasri BR
Identifying Potential Lyme Disease Cases Using Self-Reported Worldwide Tweets: Deep Learning Modeling Approach Enhanced With Sentimental Words Through Emojis
J Med Internet Res 2023;25:e47014
URL: https://www.jmir.org/2023/1/e47014
doi: 10.2196/47014
PMID: 37843893