Assessing Twitter's Potential to Predict Lyme Disease Incidence Rates in the US

Austen Ng, Jiaxin (Fiona) Li, Raja Mahadevan, Srikanth Boligarla

A Thesis in the Field of Data Science

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2022

Abstract

Lyme disease is the most common vector-borne disease in the United States and affects approximately half a million people each year. However, there are currently no surveillance systems to provide early warning for Lyme disease outbreaks. Twitter has been evaluated by many researchers for disease spread, but its potential in the field of Lyme disease has not been fully explored in general. In this study, Twitter's potential for predicting Lyme disease incidence rates in the US was assessed. Approximately 1.3 million tweets were collected through Twitter API for the years 2010 to 2019, of which 419,000 had identifiable geolocations and were later classified as Lyme disease-related or not. Different NLP algorithms were built and tested on manually labeled tweets. The BERTweet model yielded the best F1 and accuracy score of about 90%. Classification models, time series forecasting, and spatio-temporal analysis were further conducted to examine the correlation between the classified tweets and historical Lyme disease data. These models were also applied to make predictions on future trends of Lyme disease. The study results were not conclusive for Twitter's potential to improve Lyme disease predictions due to data limitations and the quality of signal in some counties. However, it could be used as a proxy for historical Lyme disease data as a real-time surveillance tool.

Acknowledgments

We would like to thank Professor Bruce Huang, Director of Master's Degree program at Harvard University, for his flexibility and availability and helping us with identifying the project and guiding us in every step of our Capstone journey. We would like to thank our advisor Professor Bouchra Nasri for the opportunity to work on this project with her team at University of Montreal. She was always there for us for review, guidance, and encouragement. We also thank our mentors Mamadou Yamar Thioub, Dr. Elda Laison from University of Montreal for guidance and weekly reviews. We thank our mentor Leonardo Neves for guiding us on our model building process. We are thankful to Yangming (Jason) Lin (ALM '21) for inspiring us to pick this research and helping us along with other mentors. We thank Dr. Stephen F. Elston, Harvard Instructor, for guiding us in the spatial-temporal analysis and giving us his valuable input.

Table of Contents

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

Lyme disease (CDC, 2021a) is an infection caused by the bacterium Borrelia burgdorferi and rarely, Borrelia mayonii that are transmitted to humans through the bite of infected blacklegged ticks. The infection may result in skin rash and flu-like symptoms with impact to the nervous system, joints, and heart in advanced cases. Other than antibiotics' treatment, there are no vaccines available currently for this illness. While only about 30,000 Lyme disease cases are reported each year, there may be as high as 500,000 cases in the US annually (CDC, 2021a).

This project evaluates if Twitter data originating in the United States can be used to predict Lyme disease incidence rates within the United States. The hypothesis is that a correlation exists between Lyme disease cases and the number of tweets that mention this disease. Another hypothesis is that Tweets related to disease symptoms may take place on Twitter even before a possible outbreak is known to the health authorities. Building models from Twitter that include geo-location and timestamp as independent variables may minimize the time delay in detecting Lyme disease outbreaks. Twitter will be used to determine if Lyme disease incidence rate can be predicted via context search, classification, clustering, and spatio-temporal analysis.

The study evaluates data collected from all the counties within the United States. Furthermore, a limitation of this project is that the collected Tweet data may not be confirmed cases of Lyme disease. Therefore, it is assumed that any viable models built from these data points may not be able to distinguish Lyme disease from other diseases with similar symptoms or

geo-temporal origins. Improving the Lyme disease models to differentiate from other diseases of similar symptoms is beyond the scope of this research project.

This project has potential to aid future social media-based models that predict rising cases of Lyme disease in near real-time and can complement conventional models used by epidemiologists. If these models can predict that incidence rates are likely to increase before cases become widespread, health officials can preemptively inform the populace to avoid certain areas and keep the disease prevalence low.

Chapter II.

Review of Literature

In prior research (Lin et al.), machine learning models were built to predict Lyme disease incident rates in the US counties using climate, demographic, and Google Trends' data. Neural Network and Random Forest models yielded good results with low MAE (Mean Absolute Error) of 0.16. To augment this research, social media data such as Twitter feeds is being considered as a variable to model Lyme disease incidence rates. Therefore, a literature review was conducted on the use of Twitter activity on Lyme or other diseases to evaluate if there is precedence on the use of social data to track diseases.

Studies were conducted to test the performance of data from online activity - Google Trends, Twitter, and other data from social media platforms like YouTube - as an epidemiological surveillance tool; but also discussions from different medical forums offer an interesting tool for monitoring public attention to specific infectious diseases such as Lyme disease (Seifter et al., 2011; Basch et al., 2017; Pesälä et al., 2017; Yiannakoulias et al., 2017; Kapitány-Föveny et al., 2018; Tulloch et al., 2019; Kim et al., 2020; Sadilek et al., 2020; Scheerer et al., 2020; Kutera et al., 2022). It has been previously described that Google searches with terms associated with Lyme disease show similar patterns in temporal and spatial variations consistent with the trend observed in epidemiological data (Pesälä et al., 2017; Seifter et al., 2010). Very few studies have tested the performance of Twitter data to improve the predictive aspect of the surveillance system in Lyme disease incidence. Data from social networks such as Twitter also play a role in extending traditional epidemiological models (Tulloch et al., 2019; Sadileck et al., 2020).

The research done by Tulloch et al. (Tulloch et al., 2019) to model Lyme disease incidences in the UK using Twitter showed that tweets can be used to locate possible outbreaks given that the level of relevant tweets matches known historical levels of the disease. In another study, Sadilek et al. built a tool called Lymelight to help build a better survey for Lyme disease patients. Additional research was done on Lyme disease using various datasets that include tick population, climate data, and Lyme disease cases reported (ekontowicz, 2021; Sadilek et al. 2020; Rees et al., 2019; Leighton et al., 2012). Various studies were conducted to investigate the effectiveness of using social media data to detect disease outbreaks. In Yousefinaghani et al.'s work on assessing Twitter's potential to detect outbreaks of avian influenza (Yousefinaghani et al., 2019), the correlation between counts of relevant tweets and official reports was analyzed, and SH-EDS algorithm was exploited as a technique to identify anomalies of observations on a time-series basis. After analyzing over 209,000 tweets, their results showed that 75% of actual flu outbreak news correlated to flu-related tweets. Additionally, 33% of outbreak news appeared on Twitter earlier than official confirmation of cases, which bolsters the possibility of detecting flu outbreaks early before becoming widespread. This study could potentially provide a first stepping stone for building digital disease outbreak warning systems to assist epidemiologists and animal health professionals in making relevant decisions. In a similar study by Paul et al. (Paul et al., 2014), their results indicated that Twitter data was more informative than both Google Trends data, and the models provided better surveillance results than retrospective historic data. A study to determine the accuracy of tweets related to medical condition, and the results show that tweets can be related to medical condition with 96% of precision (Garzon-Alfonso and Rodriguez-Martinez, 2018).

Despite efforts being made into the studies of exploiting social media data, there are still limitations due to an extensive amount of noise and anomalies. There is also a discrepancy in the choice of keywords since according to Kutera in his study only the keyword 'Lyme disease' is a predictor of the prevalence of LD while for Kim and his collaborators, 'tick bite' is the only one that corresponds with the seasonality of Lyme disease behavior (Kim et al., 2020; Kutera et al., 2022). In the study by Di Martino et al. (Di Martino et al., 2017), it was discussed that training models in identifying epidemic outbreaks still remained challenging regarding the balance of true positives and false positives. The results of studies that tested the performance of data from social networks, however, are mixed. Some authors propose a combination of social media data with traditional epidemiological methods that would be more appropriate to develop a surveillance tool for the prediction of infectious diseases (Teng et al., 2017; Kapitány-Fövény et al., 2019).

Others consider that awareness of tick-borne diseases is on the rise, which translates into a greater volume of Google trends and other social media data, more human cases of Lyme disease reported so online data would measure public awareness and will not have much value as a prediction tool (Schwartz et al., 2021).

The literature shows that there is a wide range of keywords that researchers associated with Lyme disease, and the steps to choose these keywords vary depending on the region and the statistical model that will be used. However, this diversity of keywords chosen could explain the divergence in the results of studies that have tried to test or prove the performance of models developed from data from online activity.

Chapter III.

Methodology

The objective is to collect and classify tweets relevant to Lyme disease from 2010 to 2019, and then determine if there is any spatio-temporal correlation between the tweet counts and disease incidence rates from CDC (Figure 1). If a correlation exists, then tweet counts may be used as a predictor of Lyme disease incidence rate.

The assumption is that any tweets about Lyme disease can be used to correlate to Lyme disease cases because if an outbreak is occurring within a given community, then there would be an uptick in tweets about Lyme disease due to increased awareness. Nonetheless, the drawback of broadly classifying tweets into Lyme disease-related is that the Lyme disease tweets are very general (i.e., may not be about the Twitter user specifically having Lyme disease or it could be a campaign for creating awareness about Lyme disease). In other words, any of these general Lyme disease tweets can be generated even if there is no outbreak and can obscure any correlations to actual Lyme cases. An alternative approach is to classify specific tweets that are only about the Twitter users contracting Lyme disease and then correlate the count of these tweets to actual Lyme cases, but this would lead to a very sparse dataset and is beyond the scope of this project. It can be a future endeavor, however.

All data collection, NLP modeling, KNN modeling, ARIMA, and decision tree modeling were conducted with Python either in Google Colab, Microsoft Azure, or local desktop. Additionally, the Spatio-Temporal analysis was done in R using the Surveillance package.

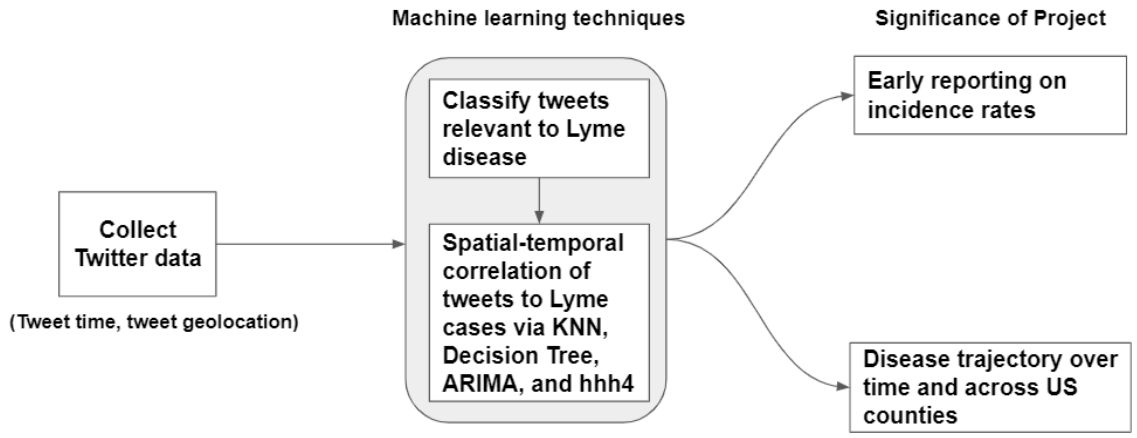

Figure 1. Project Goals and Tasks.

Overall workflow of the project.

This project was conducted sequentially in 5 steps:

1. Collected tweets about Lyme disease using the individual keywords '#lyme', '#lymedisease', and 'lyme'.
2. Labeled the tweets into 2 classes: tweets relevant or not relevant to Lyme disease.
3. Separated the tweets into train/test datasets to train NLP models for classifying tweets.
4. Classified an untouched set of tweets via the best performing NLP classification model.
5. Built prediction models (KNN, decision tree, ARIMA, and hhh4 geospatial time series) to try to correlate the classified tweets to Lyme disease cases.

Data Collection and Data Preprocessing

To create the train and test data sets for binary classification of tweets about Lyme disease, Twitter data was collected through the Twitter application program interface (API) with

an Academic Research account. By querying with “#lymedisease”, “#lyme”, and “lyme” keywords from the years 2010 to 2019, removing re-tweets, and restricting the language to English only, approximately 1.3 million tweets were collected (Figure 2). Of the 1.3 million tweets collected, approximately 725,000 tweets originated in the United States. Since the primary goal is to predict Lyme disease incidence rate in the United States using tweets, it is important that the tweets originated in the United States to ensure most tweets mentioning Lyme disease are referring to cases within the United States.

In addition to originating in the United States, the tweets need to have county or state geolocations that correspond to the same geolocations of Lyme disease cases within the United States. Otherwise, correlations between Lyme disease tweets and cases within the same geolocation would be impossible. Therefore, the 725,000 tweets had to be processed through the GeoPy (geopy.readthedocs.io, n.d.) library in Python to refine the county and state geolocations of these tweets. GeoPy approximates the tweet’s geolocation using the user profile’s location information through third party geocoders and other data sources. In addition to handling precise locations, GeoPy also uses fuzzy matching to infer the meaning of misspelled or abbreviated words such as BKLYN (Brooklyn, NY). Roughly 419,000 tweets have identifiable geolocations while 306,000 tweets still did not have identifiable geolocations after processing with GeoPy.

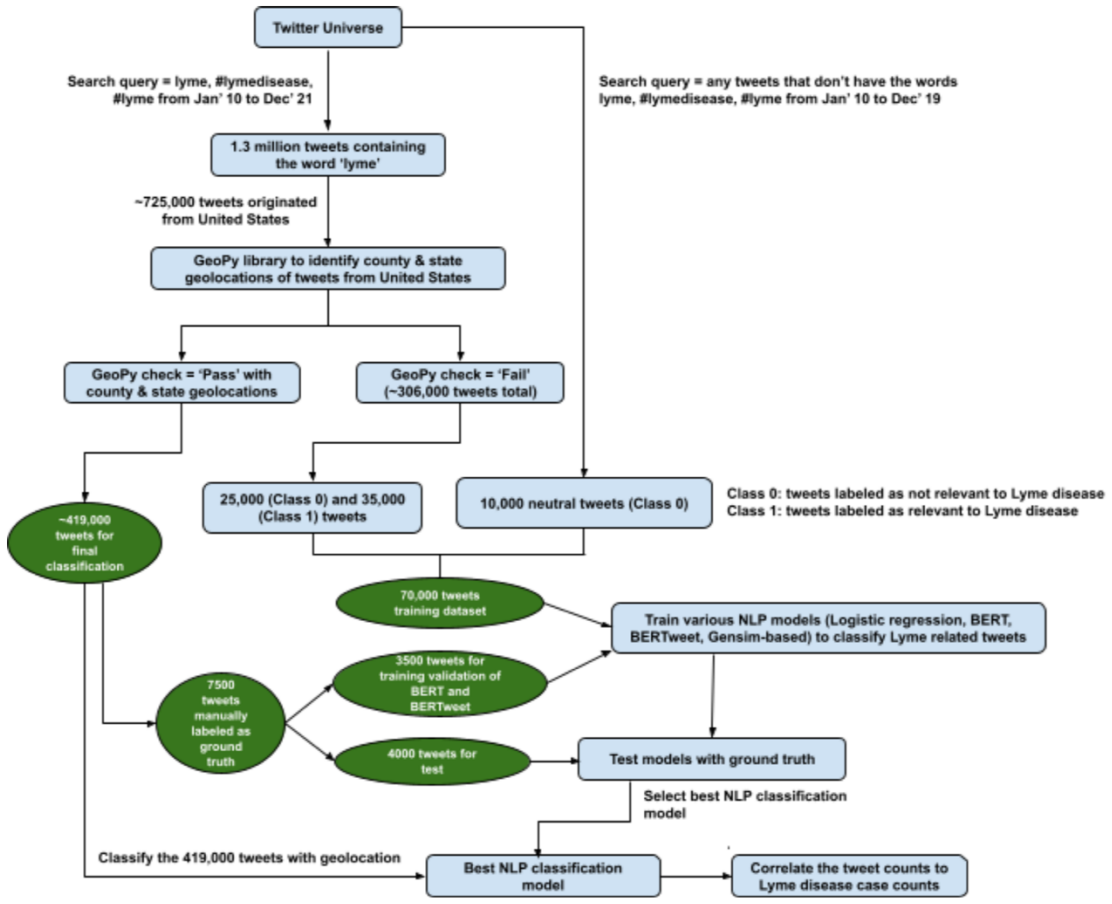

Figure 2. Tweet Data Collection and Allocation for Modeling.

Tweet data collection and processing pipeline.

Class Labeling the Data using Keywords

Each tweet needs to be labeled as either relevant or irrelevant to Lyme disease before the tweets can be split into the train and test dataset. Manually labeling hundreds of thousands of tweets is resource-intensive and impractical. Therefore, a set of keywords normally associated with Lyme disease was used to label the tweets that are either relevant or irrelevant to Lyme

disease (Table 1). Tweets relevant to Lyme disease are labeled as ‘1’ while those that are irrelevant are labeled as ‘0’.

In some cases, these keywords were chosen by inspecting the content of the collected tweets and analyzing the word frequency in all the tweets. In other cases, these keywords were selected based on known medical symptoms or modes of acquiring Lyme disease. Although all of the keywords in Table 1 were used to build the NLP classification models, subsets of these keywords may also be used to label the tweets to create a different training data set that can potentially improve tweet count correlation with Lyme disease counts for a given geolocation. Using subsets of keywords to label tweets should be explored as a future goal. Lastly, URL links in the tweets were also used to label the tweets since they also contain important words.

Table 1. Keywords for Labeling Tweets.

Keyword Category	**Keywords**
Mode of Acquiring Lyme Disease	'hiking', 'hike', 'forest', 'tick', 'ticks', 'bite', 'deer', 'deertick', 'tickborne'
Symptoms and Medical/Scientific Terms	'borreliosis', 'zoonotic', 'infection', 'erythema', 'migrans', 'carditis', 'neuroborreliosis', 'borrelia', 'bacterium', 'ixodes', 'blackleg', 'blacklegged', 'burgdorferi', 'borrelial', 'lymphocytoma', 'arthritis', 'fever', 'headache', 'headaches', 'paralysis', 'hearing', 'rash', 'fatigue', 'swollen', 'lymph', 'chill', 'chills', 'flu', 'sweat', 'inflammatory', 'inflammation', 'neck', 'knee', 'knees', 'stiffness', 'heart', 'palpitations', 'numbness', 'tingling', 'nausea', 'vomiting', 'neurologic', 'vertigo', 'dizziness', 'sleepless', 'fogginess', 'nerve', 'irritability', 'joint', 'depression', 'memory', 'malaise', 'neuro', 'long-haul', 'long haul', 'neurologist', 'dermatologist', 'late stage', 'early stage', 'antibiotic', 'specialist', 'lyme disease', 'lymedisease', 'physician', 'doctor', 'symptom', 'ache', 'pain', 'diagnose', 'diagnosis', 'patient', 'hospital', 'clinic', 'cure', 'treat', 'heal', 'disease', 'medication', 'medicine', 'therapy', 'infection', 'lyme\'s', 'tested positive', 'tested negative', 'lyme test', 'lyme\'s test', 'medical care', 'med check', 'medical checkup', 'meds', 'health', 'illness', 'bulls eye', 'bulls-eye', 'bullseye', 'bull\'s eye', 'bull\'s-eye', 'death', 'die', 'red color'

Common Vernacular/Colloquial Phrases	'have lyme', 'had lyme', 'having lyme', 'has lyme', 'get lyme', 'gets lyme', 'got lyme', 'getting lyme'

Keywords for Identifying and Labeling Tweets Relevant to Lyme Disease.

Train Dataset Preparation

Since the training data for the NLP classification models do not require geolocations, portions of the 306,000 tweets without geolocation were used as training data (Figure 1). The 419,000 tweets with geolocation were allocated for classification using the best NLP classification model in order to correlate with Lyme disease cases at the county level.

To create the training dataset, tweets were randomly selected from all 10 years so the NLP classification models can be trained to recognize differences in communication style that can occur over time. Specifically, 35,000 tweets (i.e., 3,500 randomly selected tweets from each year for all 10 years) were selected from the 306,000 tweets without geolocations and labeled as relevant to Lyme disease using the aforementioned keywords. Additionally, another 25,000 tweets (i.e., 2,500 randomly selected tweets from each year for all 10 years) were selected and labeled as irrelevant to Lyme disease (i.e., tweet topic might be about the city Lyme rather than Lyme disease) using the same keywords. Lastly, an additional 10,000 neutral tweets (i.e., tweets that do not have the word “Lyme” and are random topics unrelated to Lyme disease or the city Lyme) were also labeled as irrelevant to Lyme disease. The neutral tweets are needed to make the NLP classification models more robust towards any random tweets. Thus, the dataset for training the NLP classification models consisted of 70,000 tweets in total – 35,000 tweets about Lyme disease balanced with 35,000 tweets (25,000 + 10,000) not about Lyme disease. Any

additional data preprocessing needed for each NLP model will be discussed in their respective sections.

Test/Validation Dataset Preparation

To create the test/validation dataset (ground truth), 6,000 tweets were randomly selected from the 419,000 tweets with geolocation and manually labeled by reading each tweet. Of the 6,000 tweets, 3,800 were labeled as relevant to Lyme disease and 2,200 tweets were labeled as irrelevant to Lyme disease (i.e., tweet topic might be about the city Lyme rather than Lyme disease). Lastly, an additional 1,500 neutral tweets (i.e., tweets that do not have the word "Lyme" and have random topics unrelated to Lyme disease or the city Lyme) were also labeled as irrelevant to Lyme disease. In total, 7,500 tweets were collected for validating/testing the NLP models.

However, it was decided that 3,500 tweets of the 7,500 tweets were needed for validation of the BERT and BERTweet models during training and the remaining 4000 tweets were reserved as test data. The same 4000 tweets also served as test data for the TF-IDF logistic regression and Gensim XGBoost models.

Lyme Disease Data

Lyme disease case data from 2009 to 2019 were collected from the Centers for Disease Control and Prevention (CDC) that also contains geolocation data (i.e, county or state). National surveillance by CDC collects possible or confirmed cases that are reported by state health departments. Despite known limitations such as under-reported cases, disease misclassification, and insufficient investigation carried out by the state, the surveillance data should provide a practical estimate on the disease outbreak (CDC, 2021a).

Census Data

Population data at the county and state level were collected for the years 2009 to 2019 from the United States Census Bureau (United States Census Bureau, 2020). The population data is needed to normalize the number of tweets and Lyme disease cases given each county or state has different population sizes every year. By normalizing to the population size, Lyme disease and tweet activity incident rates can be compared across county and state level.

Adjacent County Data

A given population can travel to neighboring counties over time while sending tweets about Lyme disease from those locations. Also, people may venture into the woods for leisure in the neighboring counties containing ticks and contract the disease. Consequently, the incidence rate prediction models may need to account for these migratory patterns by using 'adjacent counties' as a predictor variable. Adjacent county data was acquired from the United States Census Bureau for inclusion into classification and time-series models for incidence rate prediction (Bureau, n.d.).

Connecticut County and Monthly Lyme Disease Data

From the Connecticut state website, monthly Lyme disease case count data for the year 2010 to 2018 was downloaded (CT.gov - Connecticut's Official State Website, n.d.). This data was later used for spatio-temporal modeling using the R package Surveillance as well as ARIMA modeling using Python. Spatio-temporal and ARIMA models were evaluated on Connecticut counties because a more frequent time interval dataset such as monthly data was needed and there were no other US states that had this frequency of data other than Connecticut.

Limitations with the Tweet Data

Accuracy of tweet geolocation is necessary in order to correlate these tweets to Lyme disease cases with matching geolocation as the tweets. Although a large amount of tweets about Lyme disease were collected, many of these tweets describing Lyme disease were not chronologically aligned with the time period from when they were tweeted. For instance, a tweet about Lyme disease may have been written recently by someone who contracted the disease 5 years ago. Chronological misalignment may prevent accurate correlation between the tweet count and the Lyme disease count or incidence rate. Additionally, some of these tweets were referring to people who contracted Lyme disease living in a different geolocation or to someone's pet contracting Lyme disease. Geolocation misalignment may also prevent accurate correlation between the tweet count and the Lyme disease count or incidence rate.

Although GeoPy provided mostly accurate tweet geolocation from the user profiles' location data, not all of these data are accurate. Providing profile location is optional for Twitter users and can be prone to outdated location information or inaccurate location entry. For instance, Twitter users occasionally do not update their profiles after relocations, which will result in mismatched geolocations between their tweets and the Lyme disease case. Geolocation misalignment may prevent accurate correlation between the tweet count and the Lyme disease incidence rate. Also, a city can sometimes be shared across multiple counties in the United States as in New York City, which can hamper the residents' ability to report their locations in a sufficiently precise manner and ultimately lead to inaccurate geolocation. Lastly, just like standardizing Lyme disease cases by the counties' population for each given year is needed to compare incidence rates across different counties, the Lyme disease tweets should also be standardized by the tweet population for each given year. However, standardizing Lyme disease

tweets by total tweet population is difficult because the tweet population is not readily available for earlier years.

Limitations with the Lyme Disease Data

A limitation of Lyme disease data is that it took a significant amount of time for CDC to provide official disease case counts. Furthermore, these disease case counts were provided annually rather than monthly, which significantly reduced the data sample size to just 10 observations (i.e., 1 observation per county for each year for 10 years). Consequently, modeling methods such as time-series could not be adequately built for all the counties in the United States. Aside from official CDC data, it is possible to use health insurance claim data as a proxy for Lyme disease case counts as performed by Swartz et. al. (Schwartz et al., 2021). However, such data was not accessible for this capstone project.

NLP Classification Modeling

Several different classification models were built and evaluated for their ability to classify the tweets as Lyme disease-related. These models include XGBoost built from Word2Vec vectorizer, logistic regression built from TF-IDF word vectorizers, BERT, and BERTweet transformer-based models. The same train and test data sets were used to build all the NLP classification models.

Gensim Word Embedding and XGBoost

Word embedding is one of the approaches to represent the words (tokens) in the high dimensional space based on the words' similarity. This is an alternate approach to represent word2vec using a bag of words.

Gensim is an NLP library that is mainly used to extract the word embedding of the given corpus (collection of documents) by training a model. In addition to training the model from the scratch for embedding, gensim also has built-in embeddings trained on large corpuses that can be fit into the given corpus. Gensim outputs the word2vec representation of each document in the corpus using the selected dimension (100, 300). These vector representations of each document can then be used for classification.

With the preprocessed and vectorized train dataset, logistic regression and XGBoost models were built to classify the tweets into Lyme disease. For the logistic regression model, default hyperparameters were used. For the XGBoost classifier model, the hyperparameters Learning rate = 0.01 and N_estimators = 100 were used.

TF-IDF (Term Frequency-Inverse Document Frequency) and Logistic Regression

The train and test datasets had to be further processed before the logistic classification model can be built. Specifically, stopwords and special characters (i.e., hashtags) had to be removed followed by tokenization and lemmatization. Stemming of the words was not evaluated for this project.

In order to train a logistic regression classifier for the tweets, each lemmatized token needs to be first converted into a numerical variable (i.e., vectorized). A variety of vectorization methods exists, but TF-IDF is one of the more reliable ones because the covariate matrix consists of the weight (i.e., TF-IDF score) of each word that is upgraded by how often the word appears

in each document (i.e., each tweet) but downgraded by how often it appears in the entire corpus. Therefore, the weight of each word in the matrix is balanced so that the importance of a given word is appropriately emphasized when training the logistic regression model for classification purposes. TF-IDF vectorization can yield reliable logistic regression models for text classification.

After TF-IDF vectorization, the logistic regression model was trained with and without 7-fold cross-validation. In addition, various hyperparameters such as L1/L2 regularization, inverse of regularization strength values ranging from 0.1 to 100, and solvers such as 'liblinear', 'lbfgs , 'sag', and 'saga' were evaluated.

BERT (Bidirectional Encoder Representations from Transformers)

BERT (Devlin et al., 2018), short for Bidirectional Encoder Representations from Transformers, is a transformer based language model pre-trained on Wikipedia and Brown Corpus. BERT models learn language embeddings through two tasks: mask language modeling and next sentence prediction. In this study, a BERT model was further trained and finetuned to classify tweets as Lyme disease related or not according to tweets' text.

For all training, validation, testing and predicting, tweets' texts were further cleaned. Hashtags, URL links, and username mentions were removed to reduce the noise within the text itself. The cleaned tweets were then vectorized using BERT tokenizer, which makes use of the WordPiece tokenization technique. For example, the word "dancing" is tokenized into "dance" and "##ing" such that similar contextual embeddings can be drawn from different samples with different word formats.

To configure the BERT model, Adam was chosen as the optimizer for its good performance handling sparse data like short tweets and low sensitivity to learning rate. In

particular, Adam was needed to handle potential issues of sparse vectorized texts due to the length limitation on the nature of tweets. In addition, Binary Cross Entropy was used as the loss function to maximize likelihood estimation, and Sigmoid was used as the activation function to better suit binary classification problems. To finetune the BERT model, the validation set which contains only hand-labeled tweets was used to assess its performance compared to the ground truth label of whether a tweet is related to Lyme disease.

BERTweet

BERTweet (Nguyen, Vu and Nguyen, 2020) is a public large-scale pre-trained model for English Tweets. It was released in 2020 and has the same architecture as BERT base. It is trained using the RoBERTa pre-training procedure on an 80GB corpus of 850M tweets. This model was evaluated for classifying tweets into Lyme or Non-Lyme disease categories. The model is proven to be effective in named-entity recognition, part-of-speech tagging, and text classification.

The pre-trained model is loaded from Hugging Face and trained on 70,000 ‘lyme’ keyword tweets and evaluated on 4000 hand-labeled tweets. The learning rate used was 2e-5 and weight decay was 0.01, and a batch size of 64. Profane words were removed in the tweets before building the tokens. Also, emojis were removed from the tweets to improve the tokenization process. The training procedure took about 1 hr with ‘accuracy’ as the best metric for early stopping value of 0.001. The model ran for 2 epochs to avoid overfitting the train dataset.

Incidence Rate Modeling

After the untouched 419,000 tweets were classified as Lyme disease-related, their counts were used to predict or correlate to actual Lyme disease cases (incidence rates). Several different

prediction models were built and evaluated for their ability to correlate or predict the classified tweets to Lyme disease cases. These models include KNN, Decision Tree, ARIMA, and hhh4 spatio-temporal time series. The same set of classified tweets were used to build all the prediction models. The KNN and Decision Tree models were evaluated for all of the US counties while the ARIMA and hhh4 spatio-temporal models were evaluated for Connecticut counties only. ARIMA and hhh4 models were evaluated on Connecticut counties because a more frequent time interval dataset such as monthly data was needed and there were no other US states that had this frequency of data other than Connecticut as per our knowledge.

Classification Models (KNN and Decision Tree)

To examine the feasibility of using Twitter data as a surveillance tool for Lyme disease, classification models were built to predict if a US county belongs to a high incidence or low incidence category using aggregated numbers of tweet counts. Data before 2010 for Twitter was very limited as the platform was not widely popular then. Also, CDC publishes Lyme disease counts once per year, restricting the usage of aggregated tweet counts to yearly instead of weekly or monthly. Thus, the resulting sample size for each geolocation was limited to 10. Therefore, ensembling methods were not investigated since bootstrapping was not feasible for such a small sample size. For the same reason, autoregressive modeling was also not considered for classification.

Instead, supervised K-Nearest Neighbors (KNN) and Decision Tree classifiers were chosen for their versatile performance. Twitter data contains noise depending on how the tweet was written (i.e., misspellings, irrelevant content), and outliers were also identified corresponding to event anomalies. Therefore, KNN and Decision Tree classifiers were selected

because they are: (1) less prone to outliers; (2) have low sensitivity on data linear separability; (3) suited for smaller datasets; and (4) able to explicitly represent decision-making processes.

Experimental Settings Twitter data, Lyme disease data, Census data, and County Adjacency data were further processed and combined to train KNN and Decision Tree classifiers. Lyme disease case counts were normalized against population for each county in the US to get the incidence rates. Based on CDC's guidelines, a state is considered a high incidence area for Lyme disease if more than 10 cases per 100,000 people are confirmed within a year. The KNN and Decision Tree classifiers will label all US counties as either high incidence county or low incidence county depending on the incidence rates for each year from 2009 to 2019.

Given the classification predictions on the 419,000 tweets, the ones that were determined to be Lyme disease related were first aggregated as tweets counts for each county, and then normalized in the same fashion against population as tweet rates. Tweet rates served as the primary feature in the classifiers to demonstrate if Twitter is able to identify counties with high incidence rate. However, considering human behaviors such as people traveling across counties on a regular basis, endemic rates for counties were also added and used as a secondary feature to the classifiers. Endemic rate for a county is defined as, the mean Lyme disease incidence rate with stable temporal pattern.

Other than the features, different subsets of data were also generated for assessment for various purposes. Since Lyme disease in the US is consistently more prevalent in the Northeast, Upper Midwest, and Northwest regions of the US, it is of less importance to predict on the counties that are consistently high or low incidence labels. Thus, the main purpose of using Twitter as a surveillance tool should focus on the counties that have inconsistent incidence rates

over the years. Another concern for model building came from the small sample sizes given limited data. KNN and Decision Tree classifiers built on the county level with such a small sample size may not result in reliable models. Therefore, models were also built by combining samples from all counties within each state or cluster. As shown in Figure 3, clusters were assigned according to similar geolocations and Lyme disease cases for the year 2019 using CDC data as a reference shown in Figure 4.

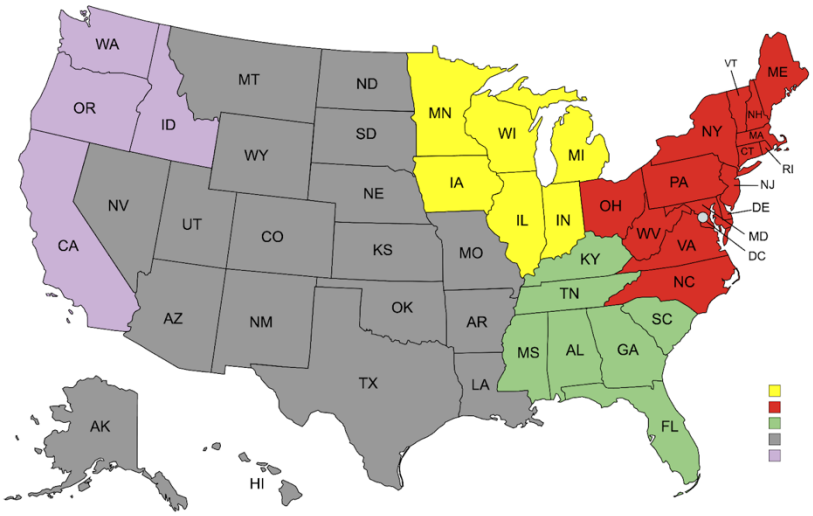

Figure 3. Cluster Assignment for Classification Models.

Cluster assignment for all US states represented on the map for classification models.

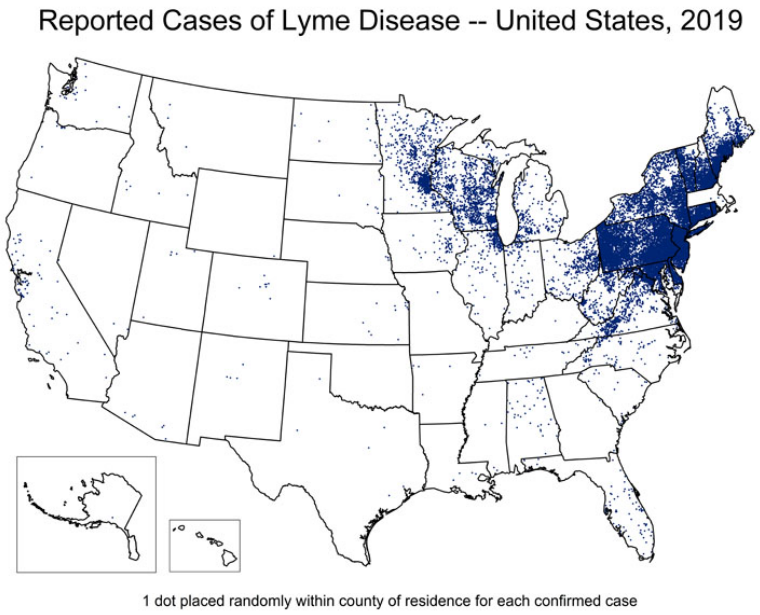

Figure 4. Reported Cases of Lyme Disease.

Reported Cases of Lyme Disease – United States, 2019. (Centers for Disease Control and Prevention, 2019)

Time-Series Forecasting

Rather than classifying a county to be Lyme disease high or low incidence county on the yearly basis, it is of higher interest to forecast the number of confirmed case onsets for each month throughout the year such that warnings could be issued in an appropriate period of time in advance. As discussed in earlier sections, monthly Lyme disease counts for all counties in the United States were not available from CDC, so the time-series forecasting was conducted on counties of Connecticut only using its monthly case counts dataset. Connecticut was chosen because it had readily available monthly data while other states did not appear to have such data.

ARIMA model was chosen for the task of time-series forecasting. ARIMA stands for Autoregressive Integrated Moving Average and is used to make future predictions using past values for time series. In addition to AR models, ARIMA allows for elimination of non-

stationarity presented in the data and also moving average smoothing. These advantages of ARIMA model help to reduce the impact of non-stationarity and smooth out trend cycles in both tweet data and Lyme disease data.

Experimental Settings To examine if using Twitter can improve the predictions on Lyme disease case counts, two ARIMA models were built and compared for all 8 counties of Connecticut. For the first model, Lyme disease case count was used as the endogenous variable. And for the second model, tweet count was added as an exogenous variable while still keeping Lyme disease case count as endogenous.

Since the models and predictions were based on each single county, Lyme disease counts and tweets counts for all months from 2010 to 2018 were directly modeled without the need to normalize by the population. This was for easier interpretation of the results. The data was split into training and test datasets in sequential order with 80% of the earlier months and 20% of the later months, respectively. To configure each model, both ACF (Auto-Correlation Function) and PACF (Partial Auto-Correlation Function) were plotted and analyzed to obtain a tight range of significant lag and moving average orders. ADF test (Augmented Dickey-Fuller test) was also performed to test whether the data was stationary and thus determine the appropriate degree of differencing. To compare the models, metrics of MAE (Mean Absolute Error) and AIC (Akaike Information Criterion) were used to assess if Twitter could improve the Lyme disease incidence rates predictions.

Spatio-Temporal Series Model

The R package Surveillance (Höhle et al., 2022) has many tools to visualize, model, and monitor spread of infectious diseases such as measles, influenza, and meningococcal disease.

There are many datasets related to these diseases as part of the package. Retrospective spatio-temporal analysis was performed using this package. Spatio-Temporal analysis is done on a special kind of object of S4 class called 'sts' – surveillance time series. The geographical information related to the US states, counties, and their adjacency information is also provided as part of the package. This information is stored as 'SpatialPolygons' object. This retrospective surveillance analysis helps identify the outbreaks by modeling disease counts into endemic (autoregressive) and epidemic (own and neighbors) entities. Spatio-Temporal analysis is performed on confirmed Lyme cases from Connecticut data (CT.gov - Connecticut's Official State Website, n.d.) and the same is performed on twitter data to observe any common patterns. CDC does not provide monthly data, for each state, required for checking seasonality and performing time series analysis. So, the available monthly data (2010-18) of Lyme Disease incidence rates for Connecticut state was analyzed. Autoregressive Poisson and Negative Binomial models were fit with 'hhh4' to a univariate and multivariate time series of counts. Hhh4 is the additive decomposition of the conditional mean into epidemic and endemic components (Paul and Meyer, 2016). The 'hhh4' function takes 2 arguments – sts object and control object. The 'sts' object is created from the actual Lyme disease counts and the control object has many parameters including neighbor effect, distribution family, autoregressive component, start, etc. Also, the neighboring county influence on the current county cases was analyzed. A predictive model was also built to predict future disease counts using just CDC cases and another model with CDC and Twitter data is created. Statistical tests were run to check the significance of the models and the distribution of the data.

Chapter IV.

Results and Discussion

Exploratory Data Analysis

CDC Data on Lyme Disease

According to the CDC, the regions with highest Lyme disease cases in the United States occur in the Northeast, Upper Midwest, and Northwest, where high incidence is defined as more than 10 Lyme disease cases per 100,000 residents. The heatmap of Lyme disease incidence rates from 2010 to 2019 for all 50 states in the US illustrates that indeed Northeast states such as Vermont, Maine, and Connecticut exhibited high incidence rates as did Wisconsin for Northwest territory and Minnesota for Upper Midwest territory (Figure 5). Furthermore, there appears to be an increase in incidence rate over time for some states, but this could be due to improved diagnosis of the disease or increased awareness among the populace such that more victims sought out medical attention than in the past. In any case, the incidence rate data was used to correlate with tweets about Lyme disease as described in subsequent sections of this report.

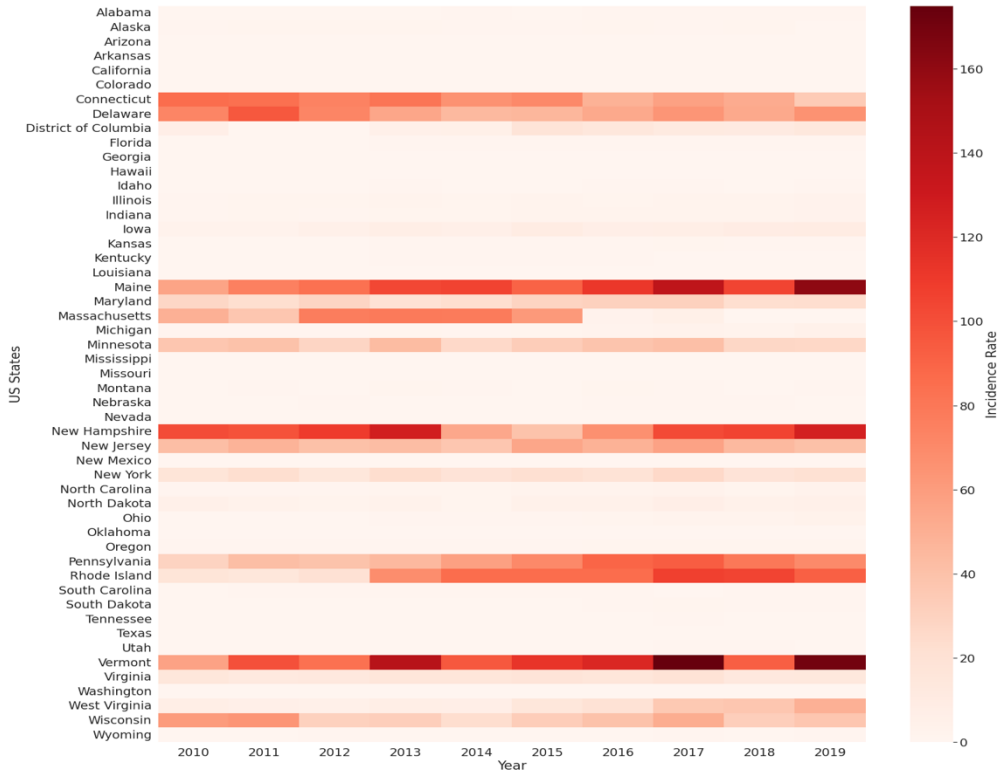

Figure 5. Heatmap of Lyme Disease Incidence Rate.

Lyme disease incidence rate for all 50 states of the US for years 2010 to 2019.

Correlation of Tweet (Keyword Labeled) and Lyme Case Counts

EDA was performed to see if there was any correlation between the Lyme disease-related tweet counts that were labeled using keywords and annual CDC-reported lyme cases. This was done as a quick check for any signs of correlation between Lyme cases and tweets despite the tweets not optimally classified yet using the NLP classification models. The annual CDC reports on Lyme cases were downloaded from the CDC website for the years 2010 to 2019. Lyme disease-related tweet counts were filtered from the Twitter universe using a set of keywords thought to be pertinent to Lyme disease. The tweet and Lyme disease case statistics for Connecticut in 2018 is shown in Table 2.

Table 2. Comparison of Tweet and Lyme Disease Case Statistics

County	Tweet Count	Confirmed Case Count	County Population	Tweet Rate	Lyme Disease Incidence Rate
Fairfield County	695	289	945,511	73.50	30.56
Hartford County	72	217	892,767	8.06	24.30
Litchfield County	6	130	181,151	3.31	71.76
Middlesex County	7	200	163,053	4.29	122.65
New Haven County	44	333	856,895	5.13	38.86
New London County	71	387	267,089	26.58	144.89
Tolland County	12	138	150,913	7.95	91.44

Tweet and Lyme disease statistics for the state Connecticut and the year 2018.

Then correlation analysis was performed against these two datasets using the Pearson and Spearman correlation tests. For most of the states where CDC reported Lyme cases were high, a strong correlation was observed between Twitter activity and CDC data. Connecticut, Vermont, and Massachusetts were some states that had good correlation between Lyme disease cases and tweet count for 2018. For Connecticut, the ranked correlation using the Spearman method had better results than the parametric correlation using the Pearson method. However, Pearson method showed a better correlation for Massachusetts. These discrepancies in correlation results indicated that perhaps the disease distributions are different between states or that the tweets were not optimally labeled using the keyword method.

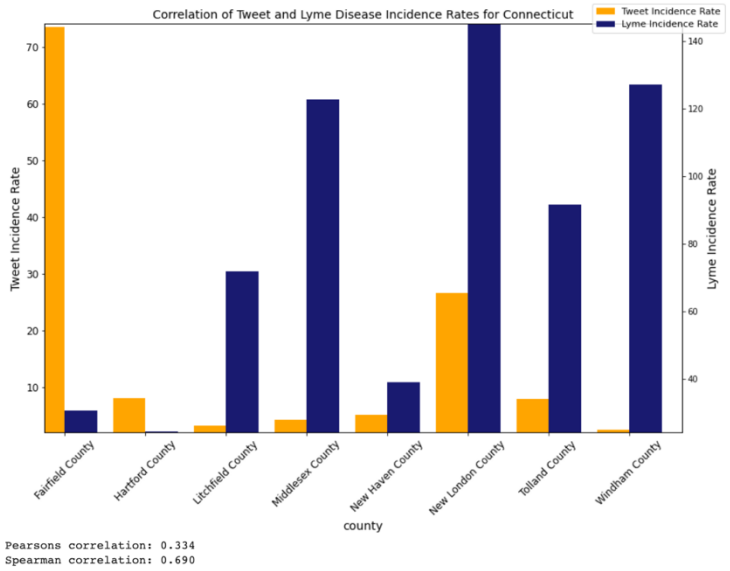

Figure 6. Correlation of Tweet and Lyme Disease Incidence Rates for Connecticut.

Correlation of tweet and Lyme disease incidence rates for Connecticut, year 2018.

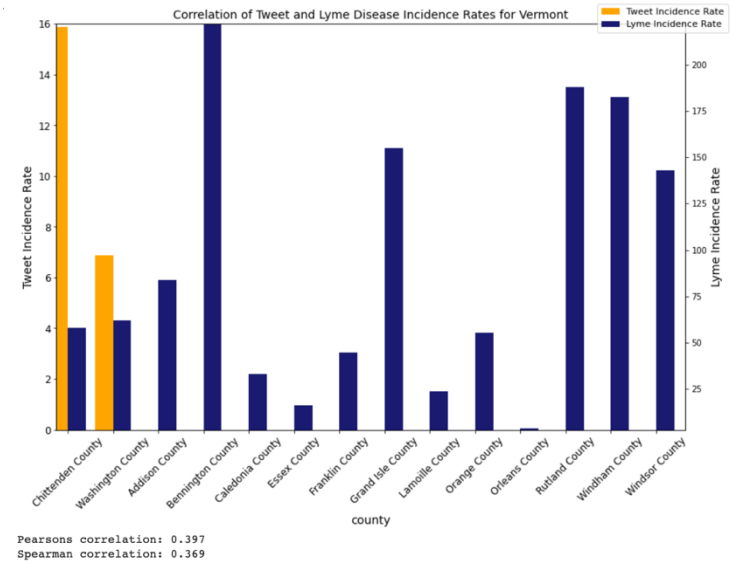

Figure 7. Correlation of Tweet and Lyme Disease Incidence Rates for Vermont.

Correlation of tweet and Lyme disease incidence rates for Vermont, year 2018.

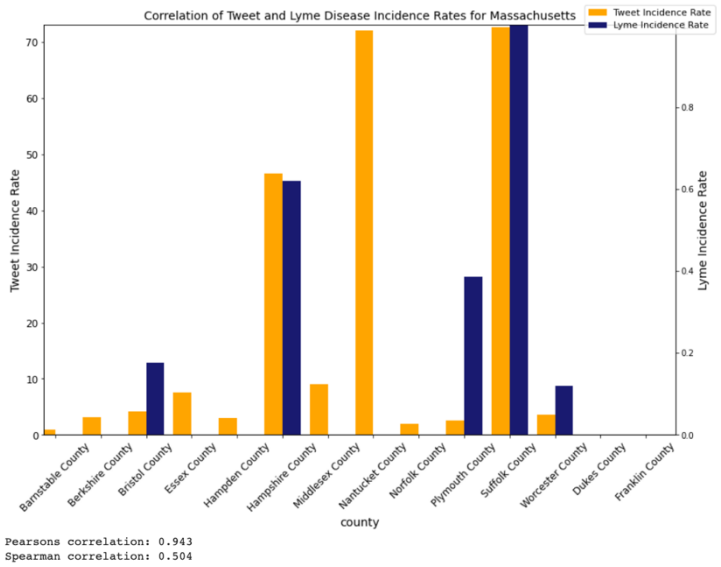

Figure 8. Correlation of Tweet and Lyme Disease Incidence Rates for Massachusetts.

Correlation of tweet and Lyme disease incidence rates for Massachusetts, year 2018.

NLP Classification Model Performance

Gensim Word2Vec

XGBoost classifier was used to classify the tweets into Lyme and non-Lyme disease categories. Word embeddings created from the Gensim model were used as features for the XGBoost model. The hyperparameters used for XGBoost were learning_rate of 0.01 and n_estimators of 100. The F1 score and accuracy values were 75% and 76%, respectively. The recall and precision values were 73% and 78% respectively. Trained Gensim model for Word2Vec embedding was not efficient, possibly due to language / context used in the tweets because of word limitations. Based on the model metrics it appeared that the word embeddings

from Gensim model along with XGBoost didn't produce favorable results. Therefore, other NLP techniques were explored to improve classification metrics.

TF-IDF Logistic Classifier

The logistic regression model was trained with and without 7-fold cross-validation. A 7-fold cross-validation did not improve the classification metrics (data not shown) on the test dataset compared to without cross-validation. In addition, various hyperparameters such as L1/L2 regularization, inverse of regularization strength values ranging from 0.1 to 100, and solvers such as 'liblinear', 'lbfgs , 'sag', and 'saga' were evaluated. Varying these hyperparameters did not significantly change the classification metrics (i.e., recall, precision, accuracy) on the train and test datasets. Ultimately, the logistic regression classifier built without cross-validation using L2 regularization, inverse regularization strength of 10, and the 'liblinear' solver was chosen to and compare against the other NLP classification models (Gensim XGBoost, BERT, and BERTweet)

After TF-IDF vectorization was performed on the train dataset, the top 1% most influential words (i.e., highest 1% TF-IDF scores) for training the logistic regression classifier model were evaluated in WordCloud (Figure 9). Words like "lyme" or "chronic" were most likely used to train the model to classify tweets that are about Lyme disease as Class 1. Similarly, words like "east" or "cat" were most likely used to train the model to classify tweets that are not about Lyme disease as Class 0.

There were some words like "road" or "new" that are ambiguous in how they train the classification model. The word "road" may have been from tweets with contents like "long road to recovery for Lyme disease" and would have trained the model to classify tweets as Class 1. Alternatively, the word "road" may have been from tweets with contents like "the road to the

city East Lyme has high traffic" and would have trained the model to classify tweets as Class 0. Also, TF-IDF vectorization penalizes words that occur too frequently in the entire corpus and would have scored the words "lyme" or "lyme disease" as less influential. However, the set of 10,000 neutral tweets (i.e., tweets completely devoid of the word "lyme" and are general topics about anything) in the train dataset helped make the words "lyme" or "lyme disease" influential for training the classification model.

Figure 9. The Most Influential Words for TF-IDF Model.

WordCloud for Top 1% Most Influential Words for Training the Model.

Based on Table 1, the TF-IDF logistic regression model performed slightly better than the basic keyword labeling method given that there were slightly more false positives (i.e., lower precision score of 94%) but fewer false negatives (i.e., higher recall score of 81%), both of which resulted in an F1 score of 87% similar to the keyword labeling method. Furthermore, the total number of accurately predicted true positives and true negatives out of the entire dataset (i.e., accuracy score of 88%) was also slightly better than the keyword labeling method.

BERT

The BERT models were trained on the NVIDIA GeForce RTX 3080 GPU. Since BERT is a transformer that was pre-trained on a massive language data, it took only a few epochs to finetune the model before the training loss converged to 0.001, which was used as an early stopping criterion. The final model presented in this study was trained with a learning rate of 1e-5, a batch size of 32, and the padding length of BERT tokenizer was set to 64.

As shown in Table 3, the test accuracy for the BERT model was 0.90 while the F1 Score was 0.89. Even though training accuracies approached 1 as the model trained, the validation accuracies remained at around 0.9 possibly due to the training dataset not optimally labeled using the filtering keywords as mentioned in the earlier section. The model interpreted wrong information from the mislabeling within the training data and thus performed worse on validations. In addition, the BERT model had a high precision score of 0.96 but a lower recall score of 0.83. These results were expected based on how the training dataset was generated. Since most common keywords were used to filter Lyme disease-related tweets, some tweets that do not contain such keywords but were in fact Lyme disease-related were misclassified. Thus, the BERT model followed the same pattern and resulted in a lower percentage of actual positives identified.

By inspecting misclassified tweets samples from the test data, a few reasons could be found to explain the model's performance. (1) Tweets were generally short in length (140-280 characters). So, for some tweets insufficient information was given to the model to make correct predictions. (2) Tweets can contain URL's or pictures on top of text, and such information was not made available to the BERT model. (3) In some cases, the tweets could also be ambiguous in

their contextual meanings. Despite these limitations, the metrics indicated that the BERT model performed relatively well overall identifying Lyme related tweets.

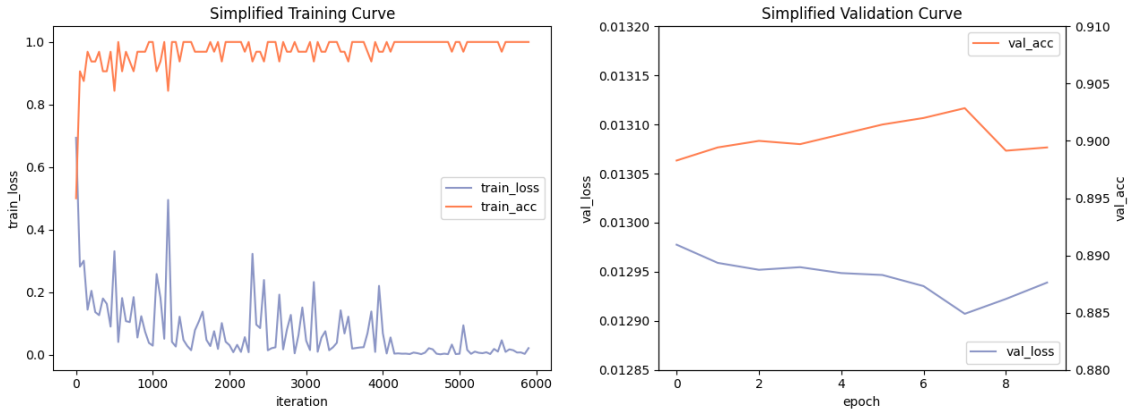

Figure 10. Simplified Training and Validation Curves for BERT Model.

Simplified training and validation loss and accuracy curves for the BERT model.

BERTweet

The virtual machine used for the training is Standard_NC12 NVIDIA TESLA K80 (12 cores, 112 GB RAM, 680 GB disk) on Azure. The F1 score and accuracy values are 90%. The recall and precision values are 85% and 95% respectively. Though the model was pre-trained on Tweets, it performed very similarly to the standard BERT base model.

Selecting the Best Classification Model

The classification metrics for all the models on the test dataset of 4,000 manually labeled tweets with balanced binary classes are shown in Table 3 and compared against the metrics of just using the basic keyword labeling method as described in earlier section. The keyword

labeling method actually had a decent ability to classify tweets into the Lyme disease-related category given the relatively high classification metric scores.

The Gensim XGBoost classification model performed worse than just using the most basic keyword labeling method given that there were more false positives (i.e., lower precision score) and false negatives (i.e., lower recall score), both of which resulted in lower F1 score than the keyword labeling method. Furthermore, the total number of accurately predicted true positives and true negatives out of the entire dataset was also lower (i.e., lower accuracy score) than the keyword labeling method. Gradient boosting is prone to overfit the training data and may have led to poorer performance. Also, the Gensim hyperparameters for creating word similarity may not have been optimized during the vectorization stage.

The TF-IDF logistic regression model performed slightly better than the basic keyword labeling method given that there were slightly more false positives (i.e., lower precision score) but fewer false negatives (i.e., higher recall score), both of which resulted in a similar F1 score to the keyword labeling method. Furthermore, the total number of accurately predicted true positives and true negatives out of the entire dataset was also similar (i.e., similar accuracy score) to the keyword labeling method.

The 2 best performing models were BERT and BERTweet given the highest accuracy, F1, precision, and recall scores. The training times were also acceptable for these 2 models. BERTweet was chosen as the model for classifying the untouched 419,000 tweets to correlate or predict Lyme disease cases.

Table 3. Performance Comparison of NLP Classification Models

Model Name	Accuracy	F1 Score	Precision	Recall	Training Time
Keyword Label	0.84	0.86	0.97	0.77	Not Applicable
Gensim XGBoost	0.76	0.75	0.78	0.73	10 min
TF-IDF Logistic Regression (no CV)	0.88	0.87	0.94	0.81	7 sec
BERT	0.90	0.89	0.96	0.83	34 min
BERTweet	0.90	0.90	0.95	0.85	53 min

Comparison of metrics for all NLP models on 4,000 test dataset.

Exploratory Data Analysis of Classified Tweets

Tweet Counts and Lyme Disease Counts Distributions

To understand the effective sample sizes over the years 2010 and 2019, a comparison of tweet counts and Lyme disease counts was plotted as in Figure 11. As the plot showed, tweet counts and Lyme disease counts roughly followed a similar trend with a few exceptions. For years 2010 to 2012 when Twitter was still in the early phase of adoption by the general public, tweet counts were expected to be lower and did not correlate to the Lyme disease counts. After initial inspection of the data, the observed spike in tweet counts for 2015 and 2016 were possibly due to the increased popularity of Twitter and general awareness of the disease. The Pearson

correlation between tweet counts and Lyme disease counts was 0.82 while the Spearman correlation was 0.92, both with significant p-values of less than 0.05.

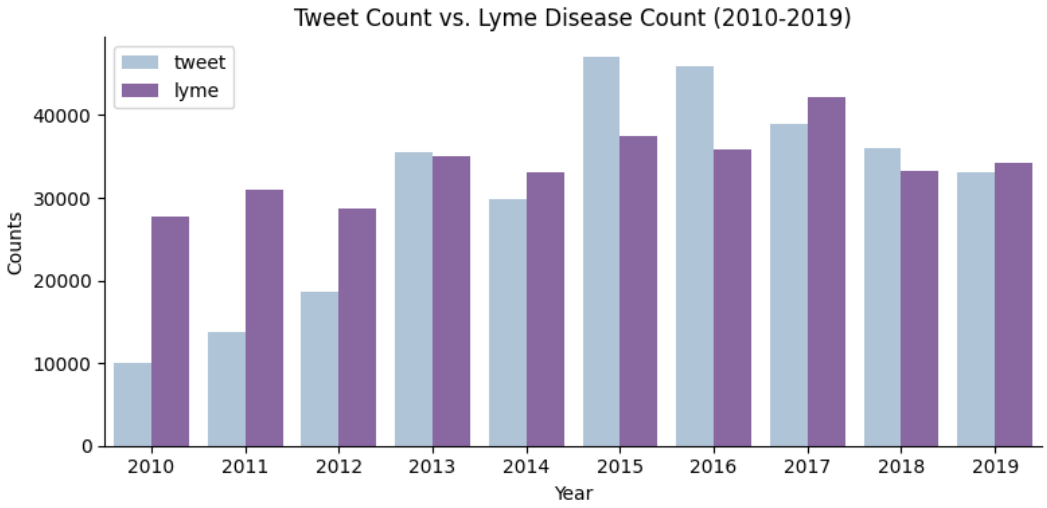

Figure 11. Tweet Count vs. Lyme Disease Count (2010-2019).

Comparison of total tweet counts and Lyme disease counts for all years 2010 to 2019.

As shown in Figure 12, Lyme disease cases showed a seasonal trend that peaked during the summer months of June and July. Tweet data was also explored to examine if a similar pattern could be established. Figure 13 showed that discussions about Lyme disease increased in May, June, and July. Additionally, May had the highest peak in tweet counts, which was about one month earlier than the Lyme disease peaks. To explain this, three main factors were considered: (1) it takes about up to one month for symptoms of Lyme disease to show; (2) Lyme disease could be hard and takes time to diagnose; (3) organizations or Twitter users send out warnings about Lyme disease in advance of time. Thus, Twitter can potentially provide information on how Lyme disease spreads over the months of a year, if historical monthly case counts were available to build the correlation.

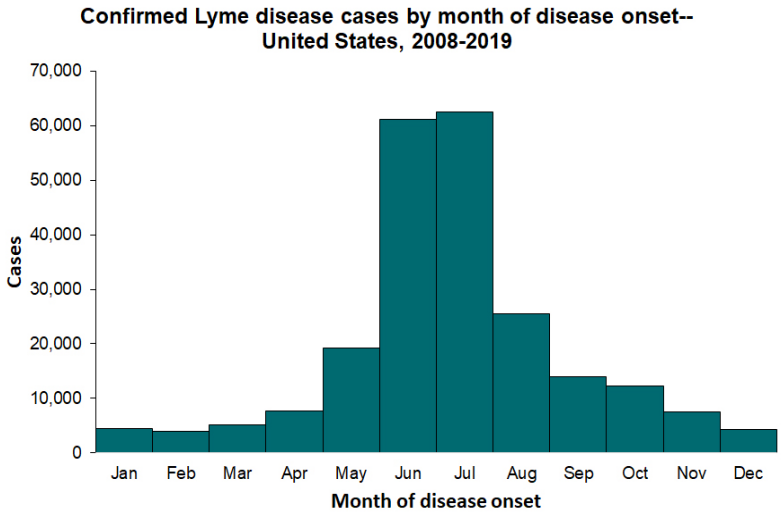

Figure 12. Confirmed Lyme Disease Cases by Month of Disease Onset

Monthly Confirmed Lyme disease for years 2018 to 2019, by CDC. (CDC, 2019)

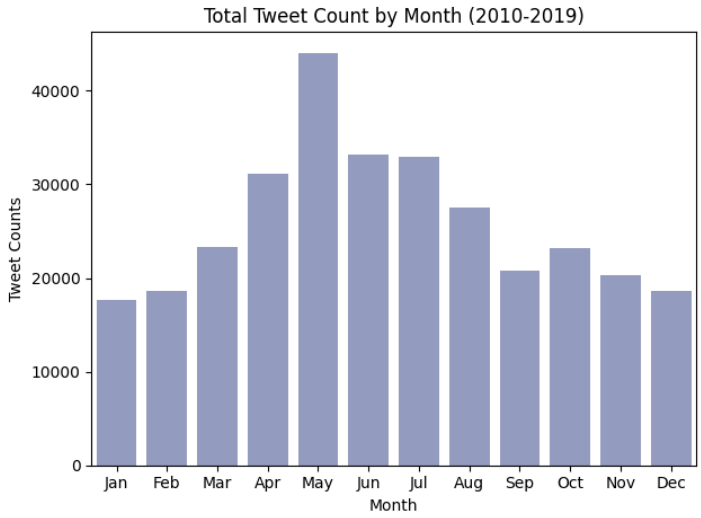

Figure 13. Total Tweet Count by Month (2010-2019).

Total tweet counts aggregated on each month for the years 2010 to 2019.

Tweet Rate and Lyme Incidence Rate on Map

To understand the pattern of how tweets and Lyme disease cases are spread over different regions of the United States, plots for both tweet rates and Lyme disease incidence rates were generated for each year. The plots for selected years (2013, 2016, and 2019) are shown as in Figure 14. For Lyme disease data, a similar distribution could be identified across the years. Northeast, Upper Midwest, and Northwest remained to be the regions with highest incidence rates. Similarly for Twitter data, a consistent pattern showed that West and Northeast regions had a higher tweet rate across the years despite a small increase in total tweets in later years. With both Lyme disease data and Twitter data showing stable distributions, it would be possible to find correlations between Lyme disease incidence rates and tweet rates.

However, the plots also revealed that the distributions for Lyme disease incidence rates and tweets rates are not identical. It is common that a county with a high incidence rate was not associated with a high tweet rate, or vice versa. This indicated that all counties within the United States did not present the same correlation between the two rates. Therefore, the correlations needed to be determined for each county or region depending on geographical locations in order to obtain precise predictions on Lyme disease incidence rates.

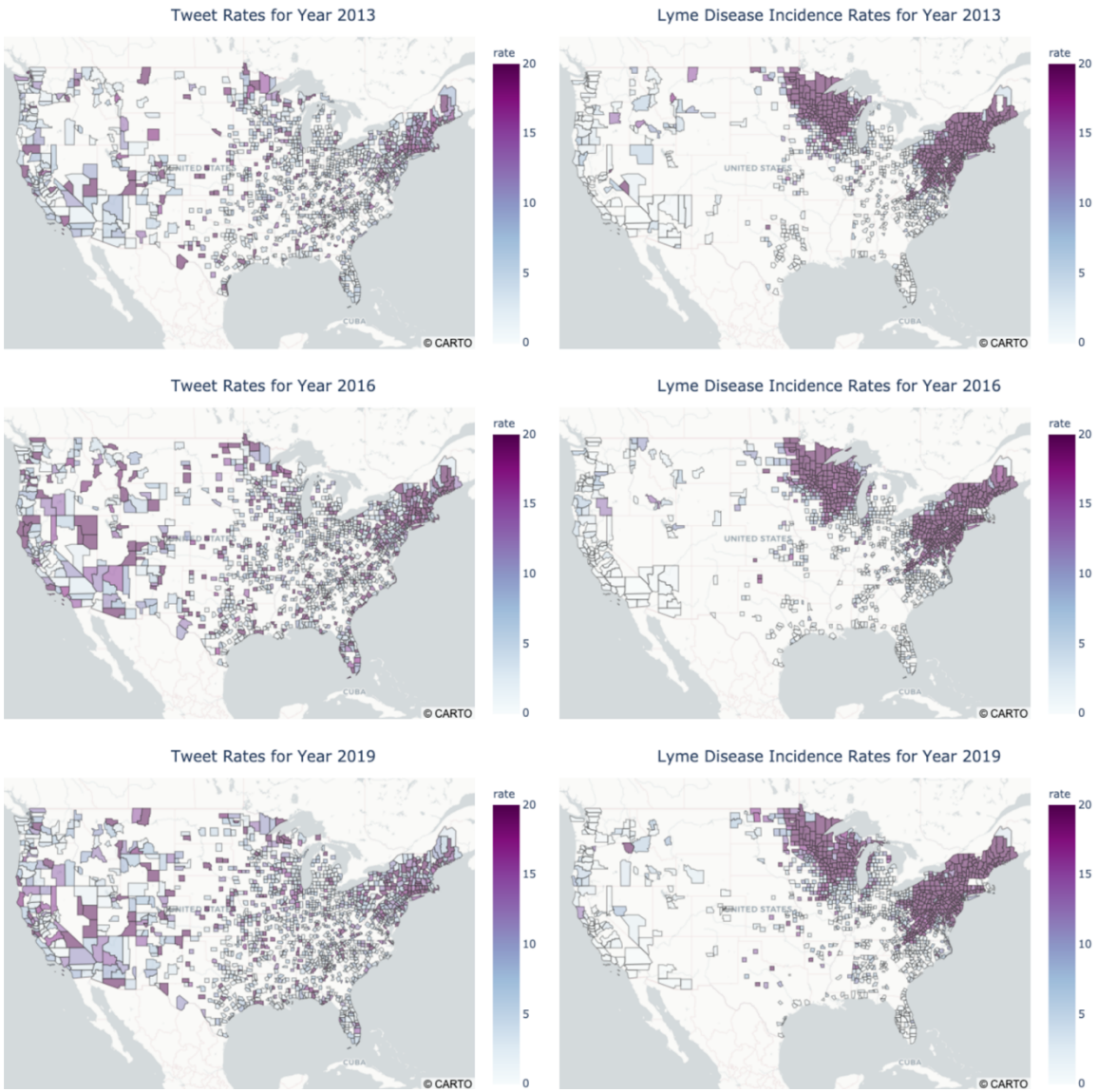

Figure 14. Tweet Rates and Incidence Rates Distributions on Geo Maps

Comparison of selected geo maps for Lyme disease tweet rates and incidence rates in the US (years 2013, 2016, and 2019).

Incidence Rate Models

Classification Models

As shown in Table 4 and Table 5, KNN and Decision Tree classifiers have comparable results for predicting whether a county is one with high Lyme disease incidence rate or not. For KNN classifiers, 3-nearest-neighbor was chosen for county level models. Since with state level models there were more data samples, 7-nearest-neighbor was chosen to obtain smoother classification boundaries. For Decision Tree classifiers, max depths of 1 and 2 were set as hyperparameters for models with 1 and 2 features respectively for county and state levels models. Since the Decision Tree classifier is more prone to overfit, especially with the small sizes of data samples, the trees were kept shallow to mitigate the potential overfitting problem.

Prediction results on the dataset with all counties showed an accuracy above 0.90 across all models with different features and geographical levels, but the accuracy dropped to the range of 0.58 to 0.72 if only counties with inconsistent incidence rates over the years 2010 to 2019 were considered. The decrease in accuracy scores was expected mainly for the reason that, for counties with consistent Lyme disease rates, the classifiers would not have made any wrong predictions due to only one single class (high incidence or low incidence) being present in the data.

After detailed inspections of the data for counties with inconsistent Lyme disease rates, several possible explanations could be deduced. First, there is a high possibility that tweets rates were not sufficiently accurate. As discussed in early sections, geo-location information associated with tweets were identified through users' self reporting and additionally were processed through Geopy. With such limitations, tweet rates for some counties could be not truly representative, so that the model was not able to successfully correlate tweet rates with Lyme

disease incidence rates. Second, CDC data which was referred to as ground truth for this study, could be unfactual. As Lyme disease case counts were reported to CDC by each state, the abilities to capture and identify positive cases may vary from region to region. Similar to the problem with tweet rates, the model might fail to capture the true correlation if no accurate ground truth is available. Lastly, the noise within tweets that are classified as Lyme disease related could be masking the correlation between tweet rates and Lyme disease rates. For this study, Lyme disease-related tweets covered a range of topics that did not directly correspond to Lyme disease positive cases such as fundraising events and campaigns for Lyme disease awareness. Such noise in the data could be a source leading to inaccurate tweet rates, which may lead to wrong predictions by the models.

The results also revealed that models built on state or cluster level have similar performance than those built on county level. However, when tweet rate was used as the single independent variable, the model performance tended to decrease as larger geographical regions were taken into account. This could indicate (1) there was variance in Twitter activity for different geographical locations, or (2) inaccuracy of tweets data such that tweet rates were not truly representative for some counties and lead to wrong correlation obtained for a region. However, if endemic rate (mean incidence rates of adjacent counties from previous year) was used as an independent variable, grouping the counties into larger regions increased the model accuracies. With endemic rates, spatial factors were also taken into consideration as Lyme disease could spread when humans or ticks travel to other places. Aside from model performance, county level models were trained on very small sample sizes, thus the results could be considered unreliable and biased. So, for the purpose of prediction, such models should be avoided.

To comprehensively understand the feasibility of using Twitter data to identify high incidence counties for Lyme disease, the focus will be on predictions of inconsistent counties. Using tweet rate, endemic rate, or both as feature set resulted in accuracies of 0.66, 0.72, and 0.70 respectively. Even though using Twitter did not add value to improve predictions on Lyme disease incidence rates for the experiments conducted in this study, the results revealed that Twitter data provided similar amounts of information as historical Lyme disease data did. Therefore, Twitter has the potential to act as a proxy for the Lyme disease data, especially when official reports of confirmed cases are not available or delayed.

Table 4. Test Metrics for KNN Models

Feature	Model Dataset	Model Level	Accuracy	F1 Score	Precision	Recall
Tweet Rate	**All Counties**	**County**	0.94	0.79	0.91	0.68
		State	0.91	0.71	0.80	0.64
		Cluster	0.87	0.49	0.71	0.56
	Inconsistent Counties	**County**	0.66	0.54	0.74	0.37
		State	0.58	0.37	0.59	0.25
		Cluster	0.61	0.39	0.70	0.3
Endemic Rate	**All Counties**	**County**	0.94	0.82	0.88	0.78
		State	0.94	0.82	0.88	0.76
		Cluster	0.94	0.82	0.86	0.75
	Inconsistent Counties	**County**	0.69	0.63	0.72	0.56
		State	0.72	0.66	0.74	0.53
		Cluster	0.71	0.66	0.72	0.51
Tweet rate + Endemic Rate	**All Counties**	**County**	0.93	0.80	0.83	0.77
		State	0.94	0.81	0.87	0.74
		Cluster	0.94	0.82	0.86	0.75

Feature	Model Dataset	Model Level	Accuracy	F1 Score	Precision	Recall
Endemic Rate	**Inconsistent Counties**	**County**	0.65	0.58	0.64	0.54
		State	0.69	0.61	0.72	0.53
		Cluster	0.70	0.65	0.71	0.49

Comparison of metrics on predictions results of year 2019 for all KNN models.

Table 5. Test Metrics for Decision Tree Models

Feature	Model Dataset	Model Level	Accuracy	F1 Score	Precision	Recall
Tweet Rate	**All Counties**	**County**	0.93	0.78	0.90	0.71
		State	0.91	0.71	0.82	0.64
		Cluster	0.84	0.55	0.54	0.37
	Inconsistent Counties	**County**	0.64	0.49	0.72	0.42
		State	0.58	0.36	0.63	0.27
		Cluster	0.60	0.41	0.66	0.27
Endemic Rate	**All Counties**	**County**	0.94	0.83	0.88	0.78
		State	0.94	0.82	0.90	0.76
		Cluster	0.93	0.79	0.85	0.78
	Inconsistent Counties	**County**	0.70	0.64	0.73	0.56
		State	0.71	0.63	0.77	0.60
		Cluster	0.68	0.59	0.71	0.60
Tweet rate + Endemic Rate	**All Counties**	**County**	0.95	0.83	0.90	0.77
		State	0.94	0.81	0.88	0.76
		Cluster	0.93	0.80	0.85	0.78
	Inconsistent Counties	**County**	0.71	0.63	0.77	0.55
		State	0.70	0.62	0.75	0.53
		Cluster	0.68	0.59	0.72	0.60

Comparison of metrics on predictions results of year 2019 for all Decision Tree models.

Time-Series Forecasting

Orders of lag and moving average were selected for each ARIMA model by examining plots of ACF and PACF. Shown in Figure 15 the ACF and PACF plots for Windham County as an example, both order of lag and order of moving average were set to 2. As ADF tests indicated non-stationarity in all time series for the counties, the degree of differencing was set to 1 based on model performances.

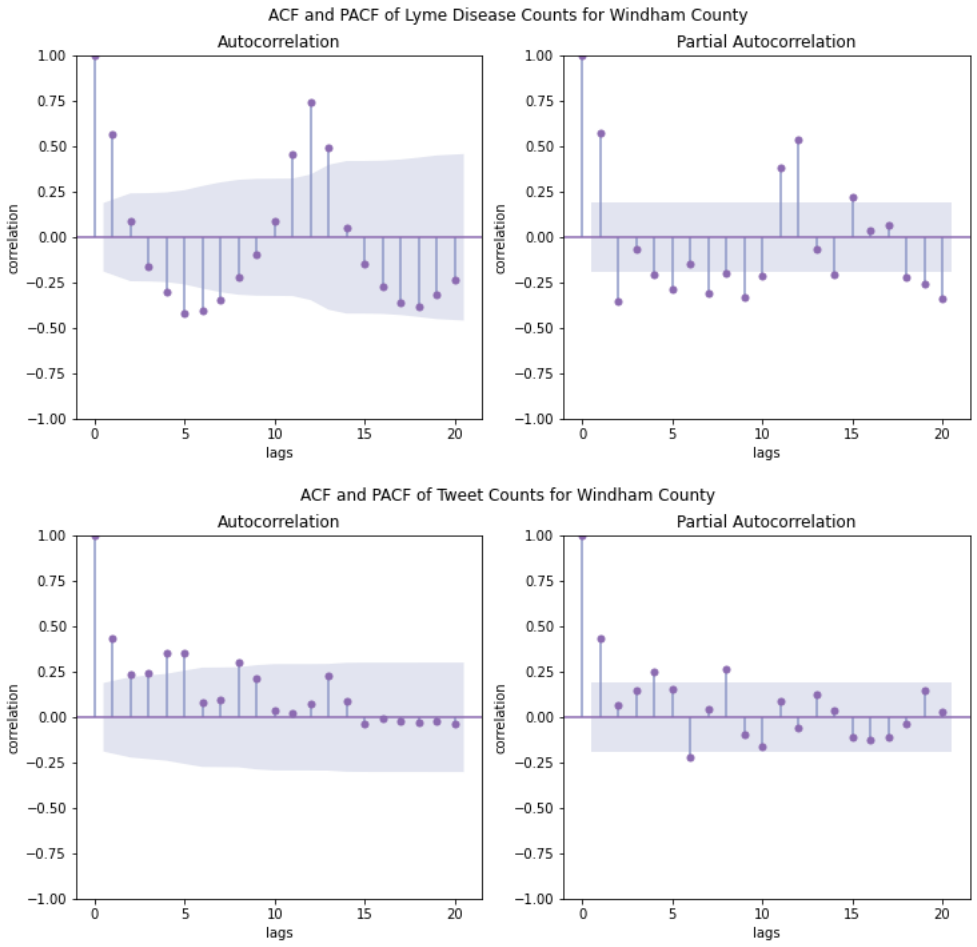

Figure 15. ACF and PACF for Time Series of Windham County.

Example of ACF and PACF of both Lyme disease counts and tweet counts

The resulting MAE's and AIC's from Table 6 suggested that ARIMA models built using tweet count as an exogenous variable performed slightly better than the models built using only Lyme disease count itself. MAE's were reduced by roughly 1 to 5 counts for all the counties, while AIC's did not decrease significantly. While the p-values of tweet count as the exogenous variable for some of the counties are within 0.05, the p-values for other counties resulted in the range of 0.1 to 0.5 which indicated that tweet count is not significant in predicting Lyme disease counts in those models. It's likely that the limitations of tweet data discussed in earlier sections resulted in inconsistent p-values across different models for all the counties.

Table 6. Test Metrics for ARIMA Models for All Counties.

	ARIMA (endog=Lyme)		**ARIMA(endog=Lyme, exog=Tweet)**	
County	**MAE (count)**	**AIC**	**MAE (count)**	**AIC**
Fairfield	16.60	754.00	11.04	747.38
Hartford	12.42	704.72	9.64	697.37
Litchfield	12.08	642.99	10.88	633.42
Middlesex	9.55	639.04	8.21	625.34
New Haven	18.10	742.94	14.22	733.87
New London	16.32	776.93	11.10	770.57
Tolland	9.86	675.16	8.97	676.23
Windham	8.69	652.13	5.96	645.64

Comparisons on MAE and AIC for all ARIMA models for all counties in Connecticut.

.

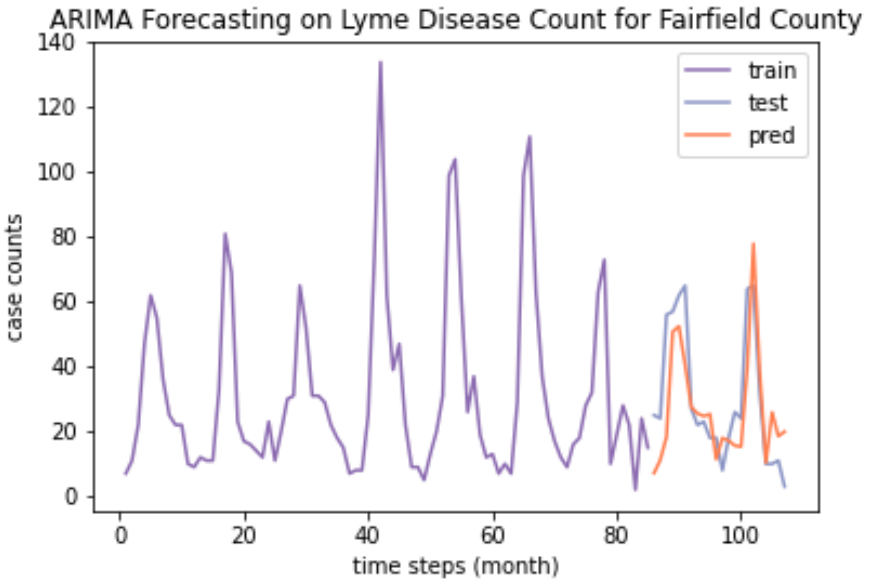

Figure 16. ARIMA Forecasting on Lyme Disease Count for Fairfield County.

Example of forecasting (pred) vs. ground truth (test) for ARIMA model.

Surveillance Models

The aggregated tweet counts and disease counts were plotted from 2010 to 2018 for the 8 counties in Connecticut.

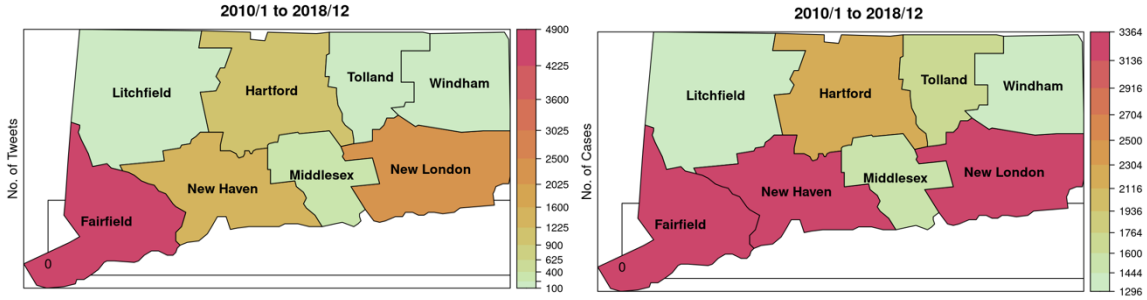

Figure 17. Total Tweet and Case Counts for Connecticut, 2010-2018.

Total tweets and case counts for all counties in Connecticut, years 2010 to 2018.

The counties appeared to have similar heatmaps for Tweets and Lyme disease cases. Fairfield county had the most tweets and most Lyme cases overall followed by New Haven and New London. Similarly, Lichfield, Tolland, and Windham had lowest Tweet counts and case counts.

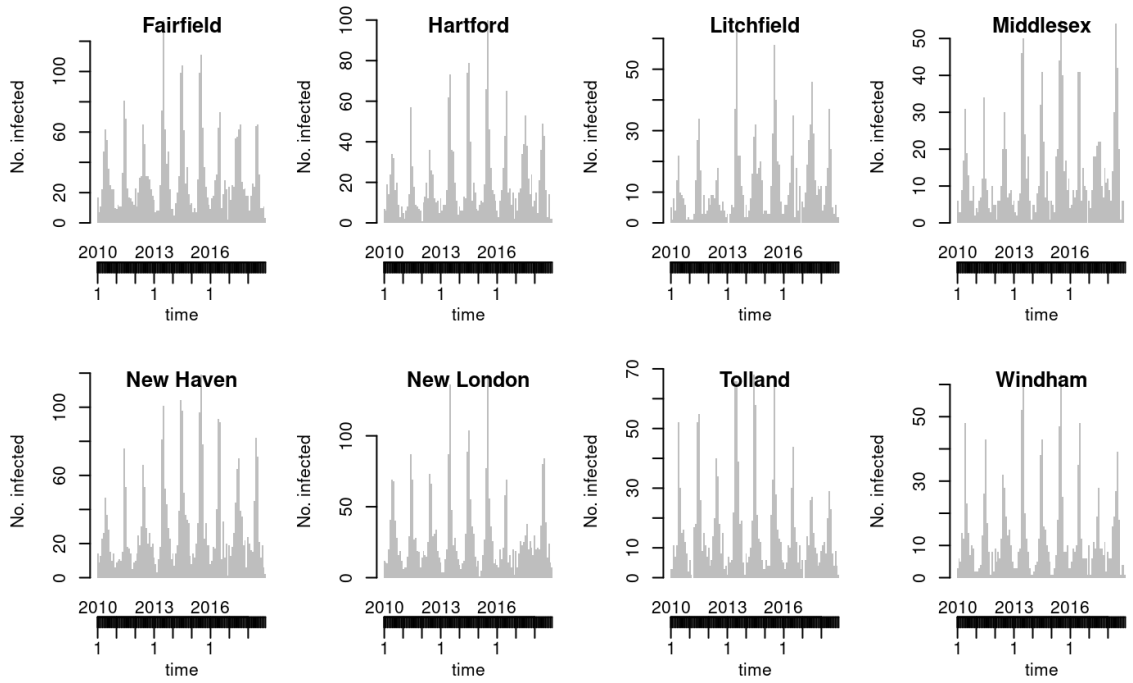

Figure 18. Lyme Disease Spread for Connecticut.

Lyme Disease spread across 8 counties of Connecticut from 2010 to 2018.

The data was assumed to have Poisson distribution. However, due to the presence of overdispersion, a negative binomial distribution was modeled and that reduced the AIC greatly thereby improving the model.

Table 7. Model AIC

Model	AIC
Poisson Model	18459.05
Neg Binomial Model	6920.60
AR1 Model	6333.04

AIC values for Poisson, Negative Binomial, and Autoregressive models.

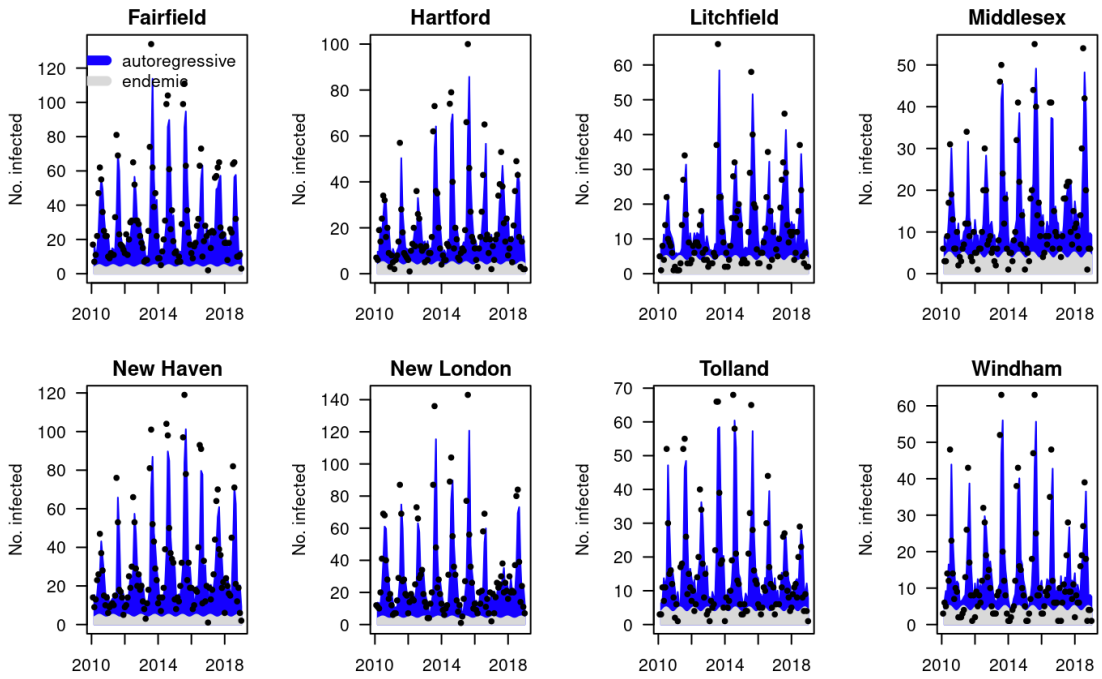

Figure 19. Temporal Analysis on Lyme Disease Spread.

Temporal analysis of Lyme disease spread in 8 counties of Connecticut.

The temporal analysis above showed the split of endemic and epidemic components of each county across the years. The occasional outbreak could be modeled from the epidemic component using this retrospective analysis.

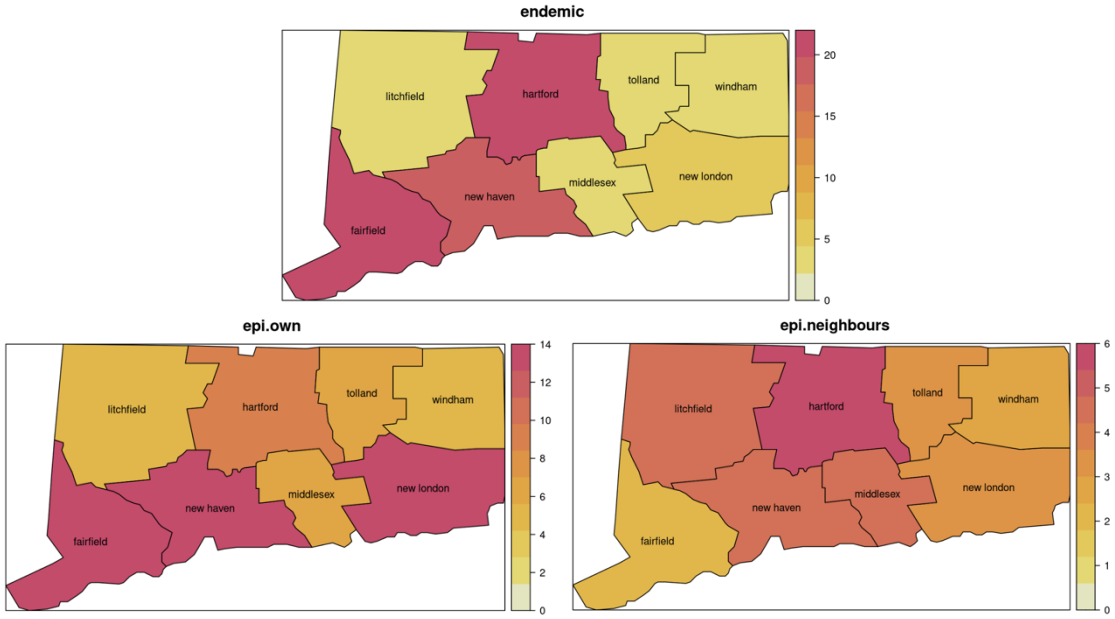

Figure 20. Endemic and Epidemic Rates for Connecticut.

Endemic rate and epidemic rates due to own and neighboring counties.

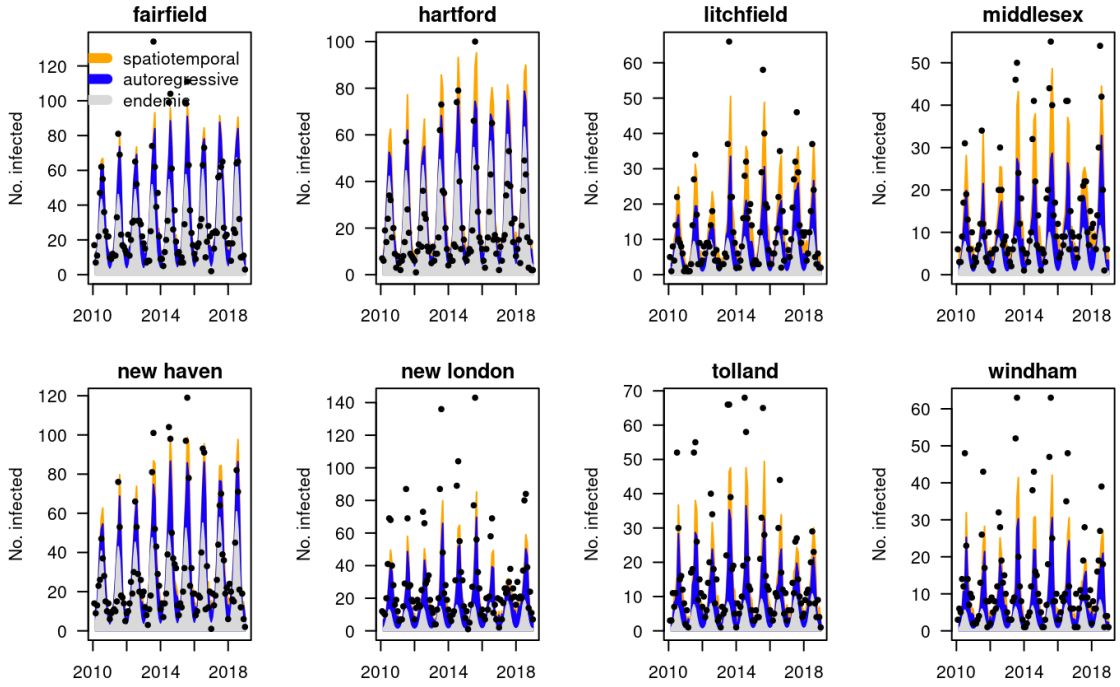

Figure 21. Spatio-Temporal Analysis on Lyme Disease Spread.

Spatio-Temporal analysis of Lyme disease spread in 8 counties of Connecticut.

The spatio-temporal analysis for the counties showed the influence of neighboring counties in orange color. A county is considered to have high disease incidence if there are 10 cases per 100,000 people per year. Using spatio-temporal analysis, it could be inferred that a borderline low incidence county could become a high incidence county because of the influence of neighboring counties, particularly if the neighbors had higher incidence rates. 'Moran I' test for summer and winter months with spatio-temporal data for Connecticut counties was performed to evaluate spatio-temporal correlation between tweet counts and Lyme cases. The 'Moran I' values are better (> 0.2) than the Tulloch et al. value of 0.172 for most of the summer months. However, the spatio-temporal correlations between tweet counts and Lyme cases were not significant because the p-values were above 0.05.

Surveillance as a predictive model Predictive model is a time series forecasting model built with the help of a surveillance package using tweet counts as exogenous variables along with CDC lyme cases counts.Using the model built, the last 5 months of case counts are predicted. Predicted model is validated using various statistical measures to observe the significance of the exogenous variable.

Calibration Test This stat test helps to validate the samples are poisson distributed after the addition of exogenous variables. Observed p-value in our case is 2.2e-16 which is < 0.001 and thus we reject the null hypothesis and accept the alternate hypothesis that the samples follow the poisson/negative binomial distribution.

Validation metric (score) The AIC values for the model with just CDC cases and the model with CDC and tweet counts are as follows. The AIC for the model built on CDC data alone looks better.

Table 8. AIC For Predictive Models.

	AIC
CDC Data alone	6005.4
CDC+Twitter Data	8346.7

AIC values for predictive models.

However, Czado et al. (2009) and Paul and Held (2011) suggested AIC and BIC as validation metric can be problematic because of random effects rather they suggest using logarithmic score (logs) or the ranked probability score (rps). The scores for the model with just CDC cases and the model with CDC and tweet counts are as follows.

Table 9. Logs and RPS Scores for Predictive Models.

	logs	**rps**
CDC Data alone	3.55	6.03
CDC+Twitter Data	4.78	6.49

Logs and RPS scores for the surveillance predictive models.

Looking at the above metrics, there was no improvement in the hhh4 predictive model by adding Twitter as an exogenous variable.

Comparison of the Incidence Rate Models

Overall, tweet data did not present significant improvement to Lyme disease incidence rates predictions on top of the disease cases data. With KNN and Decision Tree classification models, tweet data was considered to have comparable predictive power as Lyme disease data. However, ARIMA time series forecasting and hhh4 Spatio-Temporal predictive models revealed inconclusive results due to inconsistency in the significance levels of tweet data for all the counties. Even though no evidence was found to support the feasibility of using Twitter data to improve Lyme disease incidence rate predictions, the experiment demonstrated that Twitter could potentially be used as a proxy for historical Lyme disease data to give early warning signals on high disease cases onset.

Due to limited availability of monthly Lyme disease data, the time series models built could be biased, and the results were not truly representative for all counties in the United States. Other factors such as inaccurate tweet counts and under-reported Lyme disease cases, could all contribute to the uncertainty of whether tweet data was significant to additionally explain the variance in Lyme disease case counts. Further experiments using data of higher quality and more geolocations are needed for comprehensive assessment of Twitter.

Chapter V.

Conclusion and Future Direction

Approximately 1.3 million tweets were successfully collected and processed to isolate the most relevant tweets about Lyme disease, where many contained geolocations as well. These tweets were labeled as either relevant or irrelevant to Lyme disease with a chosen set of keywords. The tweets were then used to train and test a couple of different NLP models (i.e., logistic regression, XGBoost, BERT, and BERTweet) to classify tweets as either relevant or irrelevant to Lyme disease. Ultimately, BERTweet was chosen as the best classification model given the highest accuracy, F1, precision, and recall scores.

An untouched set of tweets with geolocation were then successfully classified as either relevant or irrelevant to Lyme disease. These classified tweets were then used to correlate or predict actual Lyme disease incidence rates using KNN, Decision Tree, ARIMA, and hhh4 spatio-temporal models. Ultimately, none of these models provided conclusive evidence that they could reliably be used to correlate or predict Lyme disease incidence rates. For example, KNN and Decision Tree models can reliably predict Lyme disease incidence rates in counties that have historically high or low incidence rates, but had lower predictive performance in counties with historically mixed incidence rates. Likewise, the hhh4 spatio-temporal model indicates some level of correlation between tweet and incidence rates as indicated by the Moran I score greater than 0.22 just like the Tulloch, et. al. paper (Tulloch et al., 2019), but the correlation was not significant enough based on the p-value greater than 0.05.

As mentioned in previous sections, the tweets have many limitations that can impact the ability to generate a good correlation with the Lyme disease incidence rates. As a future direction, it

would be worthwhile to identify and label tweets that are about *confirmed cases* of Lyme disease rather than just any tweet that is about Lyme disease to obtain more accurate correlations across all US counties.

References

Basch, C.H., Mullican, L.A., Boone, K.D., Yin, J., Berdnik, A., Eremeeva, M.E. and Fung, I.C.-H. (2017). Lyme Disease and YouTubeTM: A Cross-Sectional Study of Video Contents. Osong Public Health and Research Perspectives, 8(4), pp.289–292. doi:10.24171/j.phrp.2017.8.4.10.

Bureau, U.C. (n.d.). County Adjacency File. [online] Census.gov. Available at: https://www.census.gov/geographies/reference-files/2010/geo/county-adjacency.html [Accessed 5 May 2022].

CDC (2021a). Lyme Disease. [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/lyme/index.html [Accessed 18 Jan. 2022].

CDC (2021b). Lyme disease surveillance and available data. [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/lyme/stats/survfaq.html [Accessed 17 Jan. 2022].

CDC. (2019). Lyme Disease Charts and Figures: Historical Data. [online] Available at: https://www.cdc.gov/lyme/stats/graphs.html.

Centers for Disease Control and Prevention. (2019). Lyme Disease Maps: Historical Data. [online] Available at: https://www.cdc.gov/lyme/stats/maps.html.

CT.gov - Connecticut's Official State Website. (n.d.). Lyme Disease Statistics. [online] Available at: https://portal.ct.gov/DPH/Epidemiology-and-Emerging-Infections/Lyme-Disease-Statistics [Accessed 25 Feb. 2022].

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [online] arXiv.org. Available at: https://arxiv.org/abs/1810.04805.

Di Martino, S., Romano, S., Bertolotto, M., Kanhabua, N., Mazzeo, A. and Nejdl, W. (2017). Towards Exploiting Social Networks for Detecting Epidemic Outbreaks. Global Journal of Flexible Systems Management, 18(1), pp.61–71. doi:10.1007/s40171-016-0148-y.

ekontowicz (2021). Lyme-disease-Elastic-Net-regression-Nowcasting. [online] GitHub. Available at: https://github.com/ekontowicz/Lyme-disease-Elastic-Net-regression-Nowcasting [Accessed 18 Jan. 2022].

Garzon-Alfonso, C.C. and Rodriguez-Martinez, M. (2018). Twitter Health Surveillance (THS) System. 2018 IEEE International Conference on Big Data (Big Data). doi:10.1109/bigdata.2018.8622504.

geopy.readthedocs.io. (n.d.). GeoPy's documentation. [online] Available at: https://geopy.readthedocs.io/en/stable/ [Accessed 25 Jan. 2022].

Höhle, M., Meyer, S., Paul, M., Held, L., Burkom, H., Correa, T., Hofmann, M., Lang, C., Manitz, J., Riebler, A., Bové, D.S., Salmon, M., Schumacher, D., Steiner, S., Virtanen, M., Wei, W., Wimmer, V. and R), R.C.T. (A few code segments are modified versions of code from base (2022). surveillance: Temporal and Spatio-Temporal
Modeling and Monitoring of Epidemic Phenomena. [online] R-Packages. Available at: https://cran.r-project.org/web/packages/surveillance/index.html [Accessed 5 May 2022].

Kapitány-Fövény, M., Ferenci, T., Sulyok, Z., Kegele, J., Richter, H., Vályi-Nagy, I. and Sulyok, M. (2018). Can Google Trends data improve forecasting of Lyme disease incidence? Zoonoses and Public Health, 66(1), pp.101–107. doi:10.1111/zph.12539.

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

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

Leighton, P.A., Koffi, J.K., Pelcat, Y., Lindsay, L.R. and Ogden, N.H. (2012). Predicting the speed of tick invasion: an empirical model of range expansion for the Lyme disease vectorIxodes scapularisin Canada. Journal of Applied Ecology, [online] 49(2), pp.457–464. doi:10.1111/j.1365-2664.2012.02112.x.

Meyer, S., Held, E.-N. and Höhle, M. (2017). hhh4: Endemic-epidemic modeling of areal count time series. [online] Available at: https://cran.r-project.org/web/packages/surveillance/vignettes/hhh4_spacetime.pdf.

Nguyen, D., Vu, T. and Nguyen, A. (2020). BERTweet: A pre-trained language model for English Tweets. [online] Available at: https://arxiv.org/pdf/2005.10200.pdf.

Paul, M. and Meyer, S. (2016). hhh4: An endemic-epidemic modelling framework for infectious disease counts. [online] Available at: https://cran.r-project.org/web/packages/surveillance/vignettes/hhh4.pdf.

Paul, M.J., Dredze, M. and Broniatowski, D. (2014). Twitter Improves Influenza Forecasting. PLoS Currents, 6. doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

Pesälä, S., Virtanen, M.J., Sane, J., Mustonen, P., Kaila, M. and 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), p.e86. doi:10.2196/publichealth.8306.

Rees, E., Ng, V., Gachon, P., Mawudeku, A., McKenney, D., Pedlar, J., Yemshanov, D., Parmely, J. and Knox, J. (2019). Risk assessment strategies for early detection and prediction of infectious disease outbreaks associated with climate change. Canada Communicable Disease Report, 45(5), pp.119–126. doi:10.14745/ccdr.v45i05a02.

Sadilek, A., Hswen, Y., Bavadekar, S., Shekel, T., Brownstein, J.S. and Gabrilovich, E. (2020). Lymelight: forecasting Lyme disease risk using web search data. npj Digital Medicine, 3(1). doi:10.1038/s41746-020-0222-x.

Scheerer, C., Rüth, M., Tizek, L., Köberle, M., Biedermann, T. and 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), p.e18581. doi:10.2196/18581.

Schwartz, A.M., Kugeler, K.J., Nelson, C.A., Marx, G.E. and 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), pp.499–507. doi:10.3201/eid2702.202728.

Seifter, A., Schwarzwalder, A., Geis, K. and Aucott, J. (2010). The utility of 'Google Trends' for epidemiological research: Lyme disease as an example. Geospatial health, 4(2), p.135. doi:10.4081/gh.2010.195.

Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., Lin, B., An, X., Feng, D. and Tong, Y. (2017). Dynamic Forecasting of Zika Epidemics Using Google Trends. PLOS ONE, 12(1), p.e0165085. doi:10.1371/journal.pone.0165085.

Tulloch, J.S.P., Vivancos, R., Christley, R.M., Radford, A.D. and Warner, J.C. (2019). 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, [online] 100, p.100060. doi:10.1016/j.yjbinx.2019.100060.

United States Census Bureau (2020). Census Data. [online] Census.gov. Available at: https://data.census.gov/cedsci/ [Accessed 25 Feb. 2022].

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

Yousefinaghani, S., Dara, R., Poljak, Z., Bernardo, T.M. and Sharif, S. (2019). The Assessment of Twitter's Potential for Outbreak Detection: Avian Influenza Case Study. Scientific Reports, 9(1). doi:10.1038/s41598-019-54388-4.