Syndromic Surveillance using Generic Medical Entities on Twitter

Pin Huang[◇], **Andrew MacKinlay**[◇♠] **and Antonio Jimeno Yepes**[◇]
[◇] IBM Research – Australia, Melbourne, VIC, Australia
[♠] Dept of Computing and Information Systems, University of Melbourne, Australia
pinhuang@outlook.com, {admackin,antonio.jimeno}@au1.ibm.com

Abstract

Public health surveillance is challenging due to difficulties accessing medical data in real-time. We present a novel, effective and computationally inexpensive method for syndromic surveillance using Twitter data. The proposed method uses a regression model on a database previously built using named entity recognition to identify mentions of symptoms, disorders and pharmacological substances over GNIP Decahose Twitter data. The result of our method is compared to the reported weekly flu and Lyme disease rates from the US Center of Disease Control and Prevention (CDC) website. Our method predicts the 2014 CDC reported flu prevalence with 94.9% Spearman correlation using 2012 and 2013 CDC flu statistics as training data, and the CDC Lyme disease rate for July to December 2014 with 89.6% Spearman correlation. It also predicts the prevalences for the same diseases and time periods using the Twitter data from the previous week with 93.31% and 86.9% Spearman correlations respectively.

1 Introduction

Real-time public health surveillance for tasks such as syndromic surveillance is challenging due to difficulties accessing medical data. Twitter is a social media platform in which people share their views, opinions and their lives. Data from Twitter is accessible in real-time and it could potentially be used for syndromic surveillance. Even if only a small portion of the tweets contains potentially information about the health of Twitter users (Jimeno-Yepes et al., 2015a), there is still a large volume of data that could be useful for public health surveillance.

Several approaches to predict flu prevalences from Twitter data already exist. These approaches either rely on topic modelling (e.g. Latent Dirichlet Allocation (LDA) (Blei et al., 2003)) (Paul and Dredze, 2012; Paul and Dredze, 2011) or rely on regression models on keyword frequency (Culotta, 2010a; Culotta, 2010b).

The topic modelling approach for flu prevalence prediction requires manually labelling a large number of tweets (e.g. 5,128 tweets) that are used to train a Support Vector Machine (SVM) (Joachims, 1999) classifier applied on 11.7 million messages. The predictions on the tweets are applied on a LDA based topic model to over millions of tweets (Paul and Dredze, 2012; Paul and Dredze, 2011). Regression approaches (Culotta, 2010a; Culotta, 2010b) require prior knowledge to develop a keyword list {*flu,cough,sore throat,headache*} that could identify tweets relevant to flu.

In this paper, we propose an effective and computationally efficient alternative for disease prevalence prediction based on an already existing database developed by (Jimeno-Yepes et al., 2015b). Our approach to predict disease prevalence does not require manual labelling of Twitter posts to determine whether the posts are related to a particular disease or not. Our training dataset uses aggregated weekly term frequencies, so it is less computationally expensive to train compared to other approaches trained on millions of tweets. In addition, compared with regression approaches (Culotta, 2010a; Culotta, 2010b), no prior knowledge was used to manually develop a list of keywords indicative of a disease. Overall, we used our method to effectively predict the prevalence of flu and Lyme disease one week ahead of reported CDC data.

Pin Huang, Andrew MacKinlay and Antonio Jimeno Yepes. 2016. Syndromic Surveillance using Generic Medical Entities on Twitter. In *Proceedings of Australasian Language Technology Association Workshop*, pages 35–44.

2 Modelling weekly syndromic rate

2.1 Dataset Introduction

The Twitter data for years 2012, 2013 and 2014 was obtained from the GNIP Decahose,[1] which provides a random 10% selection of available tweets. From here, only English tweets were considered and retweets were removed.

Each tweet was annotated with three types of medical named entities: disorders, symptoms and pharmacological substances (PharmSub) (Jimeno-Yepes et al., 2015b). These entity types are defined using the UMLS (Unified Medical Language System) semantic types (Bodenreider, 2004). Recognition of entities was performed using a trained conditional random field annotator. Statistics on the annotated entities by this classifier for the first half year of 2014 is available from (Jimeno-Yepes et al., 2015a). Annotation of pharmacological substances is complemented by using a dictionary based annotator using terms from the UMLS.

Since just a small portion of tweets contain declared location information, posts containing medical entities were automatically geolocated using the method presented in (Han et al., 2012). This geolocation has been used to select tweets from the USA, since our reference is US CDC.

Based on the annotated tweets in USA, Twitter terms' counts are aggregated into a weekly basis; then the terms' counts are normalized by the weekly total number of the tweets. For three years data, the sample size of the dataset used for the prevalence prediction is 156, because only about 52 weeks per year.

The weekly terms' frequencies data set is then mapped to the weekly CDC's data. Three years' data are available for the flu prevalence prediction. While year 2013 and 2014's CDC data is available for Lyme Disease prevalence prediction, therefore, the dataset for Lyme disease is with 104 sample size.

2.2 Overall Architecture

The proposed methodology is a predictive model which aims to achieve the following goals:

- Predict reported CDC flu and Lyme disease trend using weekly term frequencies to predict syndromic weekly rates.
- Predict reported CDC flu and Lyme disease trend one week in advance using weekly term frequencies to predict the following week syndromic rates.

[1] http://support.gnip.com/apis/firehose/overview.html

The overall architecture of the proposed methodology is shown in Figure 1. The first step is data preprocessing, followed by feature engineering and support vector machines (SVM) (Gunn and others, 1998) regression modelling.This regression model is trained to combine the engineered features from our Twitter database to perform syndromic prediction.

A major challenge of the first step is mapping Twitter terms with similar meanings from our database to a unique term. A mapping algorithm is proposed to map synonyms into a unique term.

After the synonyms mapping, a series of feature engineering methods are applied to engineer a final set of the most important features. Finally, prediction is made by using a trained SVM regression model on the final set of features.

Twitter Term	Concept	Entity Type
adrenal disease	adrenal disease	Disease
adrenal disorder	adrenal disease	Disease
adrenal gland disease	adrenal disease	Disease
adrenal gland disorder	adrenal disease	Disease
acne treatment	acne treatment	PharmSub
treatment acne	acne treatment	PharmSub
abdomen pain	abdominal pain	Symptom
abdominal pain	abdominal pain	Symptom
abdominal pains	abdominal pain	Symptom
gut pain	abdominal pain	Symptom

Table 1: A sample of Concept Mapping

EntityType	Found in UMLS	Not Found in UMLS
Disease	9162	19454
PharmSub	15891	23556
Symptom	2604	53142

Table 2: Unique Twitter entities found in UMLS

2.3 Twitter entity synonyms mapping

Terms in medical entities from our Twitter dataset may have the same meaning but different surface form, e.g. *vomit* and *throw up*. Treating these synonyms as different input features to a regression model may result in a performance bias. Aggregating weekly term counts for synonyms maximize the probability that each input feature is not highly correlated to each other.

Therefore, we propose a synonym mapping algorithm that uses the UMLS to map Twitter medical entity synonyms to a unique term. Table 2

Figure 1: Overall Architecture

shows the statistics of how many Twitter entities could be found or not in UMLS. The unique term is considered to be a concept term for synonyms. In UMLS, medical terms with the same meanings are associated with one concept ID. Twitter terms are mapped to the UMLS medical terms in order to find concept IDs for the Twitter synonyms. A C sharp program is developed to automate this task. Details of the algorithm is explained as below.

As already mentioned, Twitter terms are annotated with three types of medical entities: symptoms, disorders and pharmacological substances. So based on the medical entity associated with each Twitter medical term, these terms are segmented into three groups: Symptom Terms, Disorder Terms and Pharmacological Substance Terms. In UMLS, each concept ID is associated to a TUI (Type Unique Identifier), indicating the semantic type of the concept ID. Three types of TUI are used for the synonyms mapping: symptom, disorder and pharmacological substances. Each group of the Twitter medical terms mentioned above are mapped to three types of concept IDs in UMLS respectively.

If a Twitter medical term can be found in the UMLS dataset and it is mapped to only one UMLS concept ID, the concept ID will be used as reference for the term. If the Twitter term cannot be found in UMLS, the term will be the reference concept for itself.

An advantage of mapping three types of Twitter terms separately is that when a term is associated with more than one UMLS concept ID, the medical entity type associated to the term may help to determine the most suitable UMLS concept ID that is from the same type. For example, a concept ID in UMLS is associated with two semantic types: symptom and disorder; a Twitter term annotated with the disorder entity is mapped to this concept ID. The medical entity type of the Twitter term helps the algorithm to determine the most suitable concept ID for the term is the UMLS concept ID associated with the disorder semantic type. But it could still be possible that a term is mapped to more than one UMLS concept IDs. In this case, each UMLS concept ID is related one or more than one Twitter terms. The most appropriate concept ID for the Twitter term is the UMLS concept ID associated with a largest number of Twitter terms.

After determining the most suitable concept ID for each term, the algorithm continues to identify the best concept label for each concept ID. Using concept label instead of the ID helps us have a better understanding of the model outcomes.

A UMLS concept ID may be associated to more than one Twitter medical terms. The best label for a concept ID is a Twitter term that appears the most in the Twitter database. Table 1 shows a sample of the concept mapping.

2.4 Feature Engineering

After mapping synonyms to unique terms, the dataset contained 112,690 unique terms. A series of feature engineering steps are conducted to improve the computational efficiency and the predictive performance of our methodology. With the feature engineering, a set of the most important features is selected and mathematically reduced using Partial Least Square and Recursive Feature Elimination with SVM. These engineered features are the input features for the final regression model that is trained to predict weekly disease rates.

The architecture of the feature engineering is shown in Figure 2. We use a nested cross-validation strategy. The outer division is two-fold or three-fold, into a training dataset and a separate validation set. We then apply 10-fold CV to the training set of each outer fold.

Non frequent and irrelevant features are first removed. Partial Least Square (PLS) regression (Abdi, 2003) is then applied to reduce the number of dimensions. Different number of PLS components are computed from PLS. A dimension reduction technique of selecting the optimal number of PLS components is proposed in later subsection. With the optimal number of components selected from PLS, each feature's 'Variable Importance of Project (VIP)' (Wold and others, 1995) is

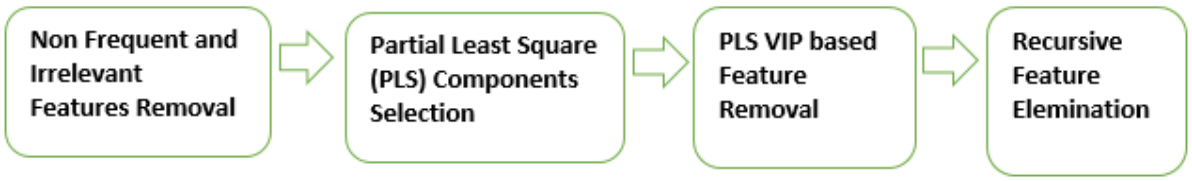

Figure 2: Feature Engineering Workflow.

Algorithm 1 Pseudocode of Selecting The Optimal Number of PLS Components

```
N = Maximum Number of Components resulted from PLS
MC = Maximum Validation Correlation
BN = Selected Optimal Number of Components
for n = 1, n = n+1, n <= N do
    Let Validation Correlation be the correlation for the outer validation dataset
    Validation Correlation = Cor(Predicted Result,CDC Rates of validation set)
   if MC < Validation Correlation then
      MC = Validation Correlation
      BN = n
   end if
end for
Return MC, BN
```

calculated. Wold and others (1995) suggest that features with very low VIP are unimportant and can be removed. A PLS VIP based feature removal technique is proposed to further remove non important features. Recursive Feature Elimination using Support Vector Machines (SVM) (Guyon et al., 2002; Gunn and others, 1998) is then used to retrieve the final set of the most important features.

2.4.1 Non Frequent and Irrelevant Features Removal

Non frequent unique terms tend to have zero variance in the dataset, which do not significantly impact the prediction outcome. Therefore, unique terms were removed if they appeared in less than 30 tweets in our dataset. This threshold was selected based on the examination of a histogram to determine the cutoff point to exclude the "long tail" of terms while still retaining important terms likely to be useful for our modelling process. Applying this cutoff the number of unique concept terms is reduced from 112,690 to 8,525. However, the number of the features is still far more than the number of samples in the reference CDC dataset (8,525 features vs 52 weeks per year each year in our study). Recent studies have shown that PLS is able to deal with datasets with more features than the sample size (Li and Zeng, 2009), therefore PLS is our first preferred algorithm to train the dataset. It has been shown that PLSs predictive performance will be improved if the irrelevant features are removed beforehand (Li et al., 2007). Our approach to determine irrelevant features is different to Li et al. (2007). In this paper, the PLS's predictive performance is considered to the correlation between the predictive CDC weekly rates and the actual CDC weekly rates. We apply Pearson correlation to determine irrelevant features. Pearson correlation measures linear relationship between two sets of variables. For each input feature, the correlation between the feature's weekly frequency and the CDC's weekly rates is calculated in the training set. If the correlation is less than 0.1, it is assumed that the linear relationship between the feature and the CDC rates is very weak, so the feature is regarded as irrelevant and removed. The remaining set of relevant features are used as input features for the next step.

2.4.2 PLS Components Selection

With a set of relevant features obtained from previous step, the PLS algorithm is applied to the train-

Algorithm 2 Pseudocode of VIP Based Feature Removal

```
T = {T_1, ..T_i, ...T_n} as the collection of VIP Threshold
T_1 =0.02, T_n=1
MaxValCor = Maximum Validation Correlation among each T_i in T
BestVIPThreshold = VIP threshold associated with MaxValCor
for T_i = 0.02, T_i = T_i+ 0.02, T_i <=1 do
   Remove features with VIP < T_i
   Run PLS on the dataset, and let MC be the 'Maximum Validation
  Correlation among PLS Components'
   MC = result of running Pseudocode of Selecting The Number of
  PLS Components
   if MaxValCor < MC then
      MaxValCor = MC
      BestVIPThreshold = T_i
   end if
end for
```

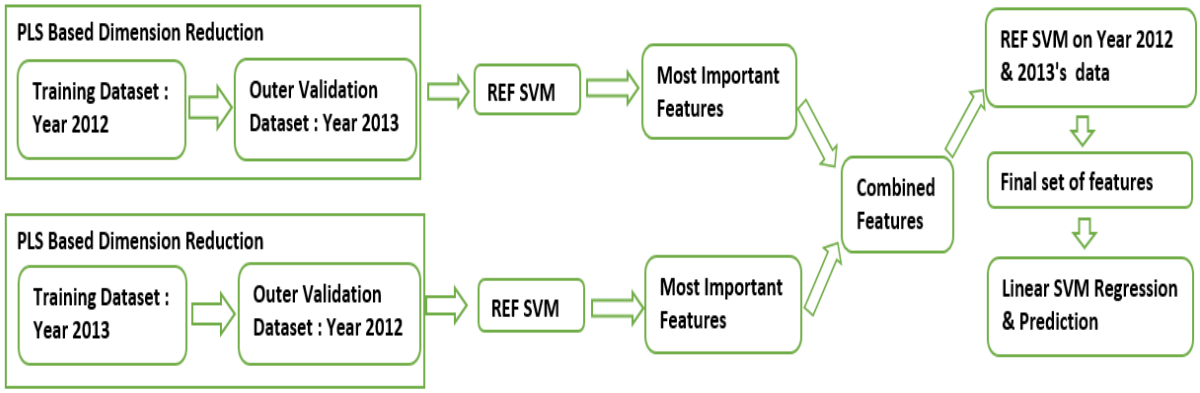

Figure 3: Flu Trend Prediction Experiment Work Flow

ing dataset with a 10-fold cross validation (Results of this are not shown in the paper). Different number of components are created by applying PLS. In order to select the optimal number of components from PLS, the "outer" validation set (from the outer CV and separate to training) is used to validate the predictive performance of applying different number of PLS components. The term 'outer validation set' is used in the later sections to refer to the validation set that is separate to the training set. The optimal number of PLS components is selected based on the maximum correlation among all components on the validation set.

Algorithm 1 shows pseudocode of a selection process to determine the optimal number of PLS components with the best predictive performance on the validation set. A loop of calculating the correlation for the validation set by using all number of PLS components is used in Algorithm 1.

2.4.3 PLS VIP based Feature Removal

With the selected number of PLS components, each input feature's VIP is calculated. Features' VIP values are only valid for the selected set of PLS components; they would be different if a different set of PLS components was selected. Each feature's VIP value is related to the feature's weights for each latent component and the variance explained by each latent component. Formula for the j_{th} feature's VIP calculation is shown below (Wold and others, 1995; Mehmood et al., 2011), where N is the number of features, m is the number of PLS latent components, w_{mj} is the PLS weight of the j_{th} feature for the m_{th} latent component, P_m is the percentage of the response factor (in our experiment, it is CDC weekly disease rate) explained by the m_{th} latent component:

$$VIP_j = \sqrt{\frac{N}{\sum_{m=1}^{M} .P_m} \sum_{m=1}^{M} w_{mj}^2 .P_m}$$

Features' VIP values are used to determine whether the feature should be removed or not. If a feature's VIP is less than a particular threshold, this feature is removed before applying PLS again to train the dataset.

We set the VIP threshold using the following methodology: Values from 0.02 to 1 are considered. We use 1 as the maximum possible, as heuristically anything greater indicates that the feature is important (Cassotti and Grisoni, 2012). The optimal VIP threshold is determined by running a loop, in which different VIP threshold values ranging from 0.02 to 1 are all used to remove features. Let T = T_1,..T_i,...T_n be a collection of VIP thresholds, n is the number of thresholds, T_i is the i_{t_h} threshold in T. For each T_i ($1 \leq i \leq n$) in T, features with VIP less than T_i are removed, then PLS is used to train on the rest of features, which results in different number of PLS components. The optimal number of PLS components is selected if it has the maximum value of correlation for the outer validation dataset. This outer validation set is the same dataset used in previous step. These components are the representation for the best result produced by removing features with VIP lower than T_i. The optimal VIP threshold in T is the one that yields the maximum correlation for the outer validation dataset. Pseudocode for VIP Based Feature Removal is shown in Algorithm 2. Any features with VIP less than the selected optimal VIP threshold are not included for the next step.

2.4.4 Recursive Feature Elimination (RFE)

In terms of the computational cost, if hundreds of features resulted from the previous step are input features for RFE, it might take too much time for RFE to present results. Therefore, if the number of features is greater than 200, features with VIP less than 0.2 are removed before applying RFE. The reduced number of features are then used as input features for linear SVM based RFE (Guyon et al., 2002; Gunn and others, 1998). Five times ten fold cross validation is used for RFE with SVM. Features selected from RFE with SVM is the final set of features, which are the input feature for the next step.

2.5 Linear SVM Regression and Prediction

After feature engineering, an SVM regression model with a linear kernel function (Gunn and others, 1998) is trained on the most important features selected from previous RFE. The final prediction is made using this SVM regression model.

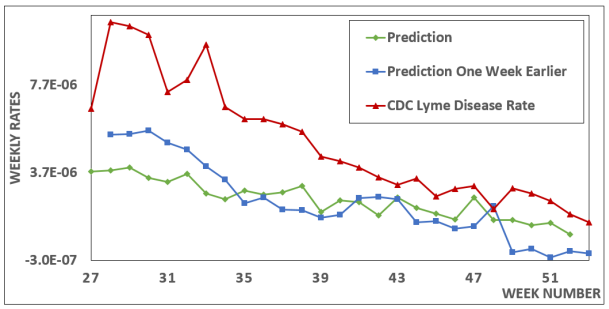

Figure 5: Second Half Year of 2014 predicted weekly Lyme Disease rates versus US CDC

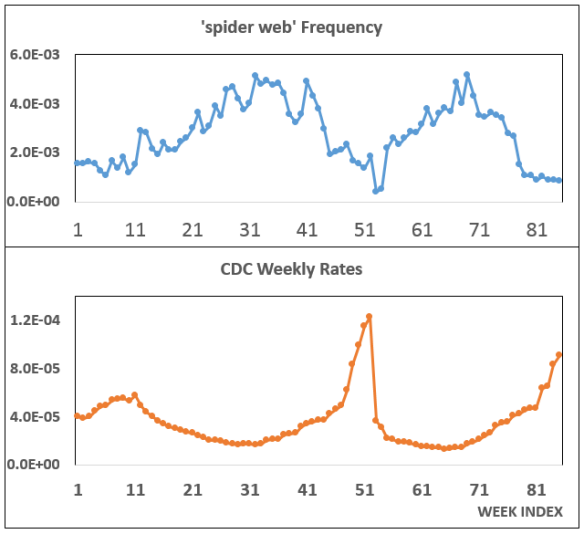

Figure 6: Weekly frequency of *spider web* versus US CDC weekly flu rate

3 Experimental Results

In this section, we show results for flu prevalence prediction for the year 2014, as well as Lyme disease prevalence prediction for the second half year of 2014, based both on the current week's data as well as using posts from a week in advance to evaluate the possibility of getting a signal earlier. Data from 2012 and 2013 is used as the training set, year 2014's weekly data is used to predict the weekly flu rates of the year 2014. For Lyme disease prediction, only 18 months of data (from 2013 to the first half of 2014) is available as the training set for Lyme disease prevalence prediction.

Two fold cross validation is used for feature tuning for flu prevalence. Figure 3 shows details of

	Stage	Training Set	Validation / Test Set	Features For PLS	VIP Feature Removal	VIP > 0.2	RFE SVM
Flu, current	1st	2012	2013	5353	412	103	61
	2nd	2013	2012	5353	992	154	74
	Final	2012, 2013	2014	–	–	–	22
Flu, one week in advance	1st	2012	2013	5060	141	–	141
	2nd	2013	2012	5060	78	–	77
	Final	2012, 2013	2014	–	–	–	21
Lyme Disease, current	1st	2013 H1, 2013 H2	2014 H1	6139	693	103	22
	2nd	2013 H1, 2014 H1	2013 H2	6139	65	–	35
	3rd	2013 H2, 2014 H1	2013 H1	6139	627	105	24
	Final	2013, 2014 H1	2014 H2	–	–	–	41
Lyme Disease, one week in advance	1st	2013 H1, 2013 H2	2014 H1	6076	63	–	21
	2nd	2013 H1, 2014 H1	2013 H2	6076	61	–	36
	3rd	2013 H2, 2014 H1	2013 H1	6076	167	–	67
	Final	2013, 2014 H1	2014 H2	6076	–	–	54

Table 3: Number of Input Features After Each Dimension Reduction

	Testing Period	Pearson Correlation	Spearman Correlation	R^2	RMSE
Flu, current	2014	92.4%	94.9%	85.3%	1.51E-05
	2014 H1	96.3%	96.6%	92.7%	1.06E-05
	2014 H2	94.8%	92.3%	89.8%	1.90E-05
Flu, one week in advance	2014	91.3%	93.3%	83.3%	1.55E-05
	2014 H1	91.6%	94.6%	84.0%	1.41E-05
	2014 H2	96.0%	92.3%	92.1%	1.69E-05
Lyme Disease, current	2014 H2	86.6%	89.6%	75%	3.41E-06
Lyme Disease, one week in advance	2014 H2	90.32%	86.9%	81.6%	3.02E-06

Table 4: Flu and Lyme Disease trend prediction results

the flu trend prediction experiment. When data from 2012 is used for training, data from 2013 is used as an outer validation set, and vice versa. After PLS based dimension reduction, RFE with SVM is applied to obtain the most important features from each fold. Another round of RFE with SVM is applied to train on the year 2012 and 2013's data with all unique input features selected from the previous step. This results in a final input feature set, and then a regression based SVM with linear kernel function is trained using 2012 and 2013 data. Finally, prediction of weekly flu rates of the year 2014 is made from the trained SVM.

For Lyme disease, we have only two years of CDC data (2013 and 2014) which overlap with our dataset of NER-tagged tweets. We set aside 2013 and the first 6 months of 2014 as for training and feature tuning, keeping the final six months for testing. We use three-fold cross-validation for feature tuning. With each fold, six months of data is used as an outer validation set, with the remainder used as a training set. Similar to the flu trend experiment procedure shown in Figure 3, important features selected from each cross-validation round are all included for another round of feature reduction by using RFE with SVM. With the final feature set determined by RFE with SVM, an SVM with a linear kernel function is then trained on the training set to make the final prediction for the Lyme disease trend for the second half of 2014.

3.1 Results of Flu and Lyme Disease Trend Prediction and Detection

Table 3 shows number of features being reduced after each step of dimension reduction. After VIP based feature removal, if the number of features exceeds 200, only features with VIP greater than 0.2 are selected. Otherwise, these features are the input features for the next step.

Experimental results for flu and Lyme disease trend prediction are presented in Table 4. Both Pearson and Spearman correlations are included. Pearson correlation measures the linear relationship between two sets of variables, while Spearman correlation measures correlation between two set of ranked variables, which is used to check whether one variable increases, the other increases or not. Therefore, Spearman correlation is used in this paper as an alternative measurement to ex-

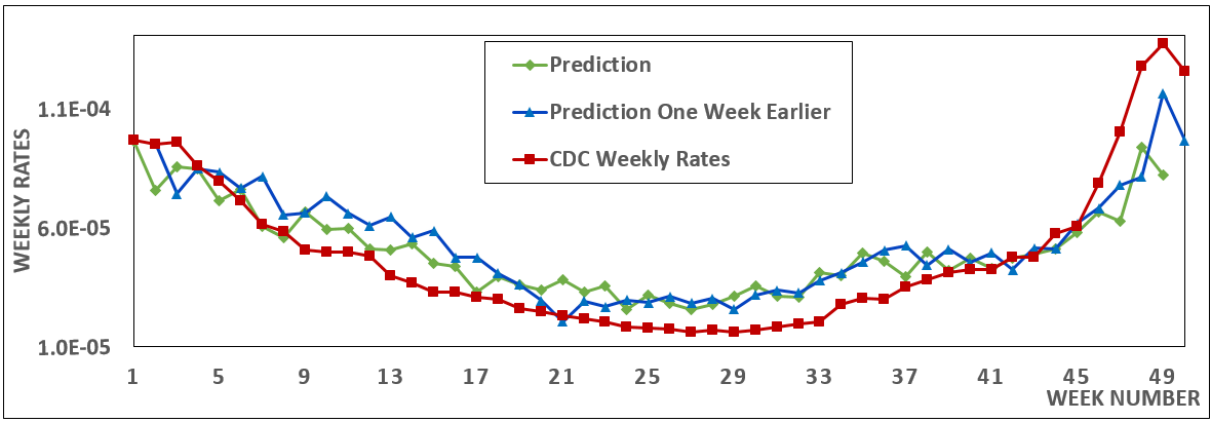

Figure 4: Year 2014 predicted weekly flu rates versus the US CDC weekly flu rates.

Flu, Current	'stomach flu' 'pneumonia' 'bronchitis' 'coughing' 'sick' 'cough medicine' 'cold sore' 'cough syrup' 'sickness' 'cold' 'red nose' 'fever' 'sinus infection' 'ear infection' 'body ache' 'blush' 'spider web' 'throat hurt' 'aching' 'strep throat' 'alcide' 'gelato'
Flu, In Advance	'stomach flu' 'pneumonia' 'bronchitis' 'coughing' 'sick' 'cough medicine' 'red nose' 'cough syrup' 'cold sore' 'cold' 'sickness' 'sinus infection' 'aloe' 'ear infection' 'fever' 'spider web' 'sleepy' 'aching' 'body ache' 'sore' 'seeing double'
Lyme Disease, Current	'coughing' 'bronchitis' 'pneumonia' 'runny nose' 'stuffy nose' 'cold' 'stomach flu' 'sick' 'throat hurt' 'sinus infection'
Lyme Disease, In Advance	'cold' 'coughing' 'pneumonia' 'bronchitis' 'runny nose' 'stuffy nose' 'sick' 'stress' 'aloe' 'stomach flu' 'sinus infection' 'caffeine' 'shaking' 'snoring' 'fart' 'concussion' 'throw up' 'migraine' 'dizzy' 'sore throat'

Table 5: Important Features for Flu and Lyme Disease Trend Prediction

amine similarities among downward or upward movements of the predicted trend and the CDC trend.

When making flu prevalence predictions for the first half, second half and the whole year of 2014, Spearman correlations are 96.6%, 92.3% and 94.9% respectively. The first half year's Spearman correlation is higher than the second half year. When the proposed methodology is used to predict flu trend one week earlier, the first half year's Spearman correlation (with 94.6%) is higher than the second half (with 92.3%). This means the final set of the most important features selected tends to represent more for the first half year's flu prevalence than the second half of year 2014. The Spearman correlation for predicting flu trend one week before CDC for the year 2014 is 93.5%, which indicates that the proposed methodology has some advance predictive power ahead of the CDC data, which is inherently less timely due to delays in collection. Figure 4 illustrates predicted flu prevalence against current CDC data as well as one week before.

For Lyme disease, as shown in Table 4, the Pearson correlation between the predicted prevalence and CDC weekly rates is 86.6%, while the Spearman correlation is higher,as 89.6%. A few weeks at the end of the year are predicted with negative rates, contributing to a relatively low Pearson correlation; without considering the last five weeks, the Pearson correlation increases to 93.3%. The relatively high Spearman correlation for Lyme disease has indicated that the upward or downward trends are well predicted. The Spearman correlation of detecting Lyme disease trend one week before CDC is 86.9%, which is lower than for the current week but still shows that a useful signal is being predicted. Predicted Lyme disease prevalence and CDC-reported Lyme disease weekly rates are shown in Figure 5.

The most important features selected by the proposed methodology for flu and Lyme disease trend predictions are presented in Table 5. Most of the features for flu prevalence prediction are reasonable, such as *coughing*, *cold* and *fever*, which are flu symptoms. However, *spider web* has been ranked as one of the features for flu prediction which appears in our database because *spider web* appears as a pharmacological substance in the UMLS. The weekly term frequency for *spider web*

is highly negatively correlated to CDC weekly flu rates as shown in Figure 6, due to many spider webs being observed in the Northern hemisphere in September, close to the low point of the flu season. *Gelato* is also detected as relevant for a similar reason, due to an coincidental (negative) correlation with the flu season. *Gelato* has been wrongly annotated by our system as a pharmacological substance since in the UMLS it refers *gelato sodium fluoride* instead of *ice cream*.

Table 5 shows the most important features for Lyme disease prevalence prediction. Many features selected are very similar to flu symptoms, in line with many symptoms of Lyme disease matching those of flu;[2] in addition, *dizzy* matches a Lyme disease symptom. However, overall the term list for Lyme disease is less convincing than for flu, with more symptoms of Lyme disease missed and more terms included with no immediately obvious relationship to the disease. It seems that the relative rarity of Lyme disease is leading to noisier signal in tweets about its symptoms.

4 Discussion

The proposed methodology is an effective approach to predict prevalences for influenza and Lyme disease based on social media posts. It predicts flu prevalences for 2014 with Pearson correlations range from 92.4% to 96.3%. Similar results have been reported with other existing approaches for flu prevalence prediction: Paul and Dredze (2012) and Paul and Dredze (2011) predicted flu rate from August 2009 to May 2010 with Pearson correlations of 95.8% and 93.4% respectively; Culotta (2010a) made predictions for flu rate from September 2009 to May 2010 with 95% Pearson correlation. However, our method has some advantages over these, as they require labour-intensive manual labelling of tweets and significant computational resources to train their system using millions of data samples, in contrast with the method proposed here, where the only computationally-intensive step is a one-off step (reusable for other diseases and other kinds of analytics) of applying an NER tagger to a large Twitter corpus. In addition, Culotta (2010a) presented a method that requires prior knowledge to manually identify flu-related key words. Here, a manually pre-built keyword list is not required as the most important features related to flu are automatically selected based on the data. We also show that our method can predict disease prevalence with some reliability in a small time window ahead of the reported CDC figures, which has potential utility for real-time disease monitoring and alerts.

Our method is somewhat generalisable, with roughly the same approach achieving good correlations against CDC data for Lyme disease. An existing approach to track Lyme disease (Seifter et al., 2010) requires knowledge to select key words from Google trends, but there is no evaluation provided. To our knowledge there is relatively little other work on Lyme disease surveillance so this application is somewhat novel. However, accuracy for Lyme disease was weaker than for flu, in terms of raw correlations as well as basic plausibility checks on the most important indicative terms – canonical indicators such as the erythema migrans rash did not make the list. An important factor is probably the lower overall prevalence of the disease (an average of 1500 reported to the CDC per week in our test set versus 15,000 for flu), there are fewer instances of Twitter users experiencing the disease and the relevant symptoms, which they could then tweet about.

This hints a limitation of the method we have developed. Terms showing a natural seasonal fluctuation but are not indicative of disease (such as *gelato*) happen to coincide with a disease may accidentally come out as important terms in the analysis. One way to mitigate this would be to improve the accuracy of the named entity tagging in the source data.

Future extensions may also be generalisable to other regions with a high number of English-language tweets.

5 Conclusion

We have presented an effective methodology which produce predictions for flu and Lyme disease prevalences with strong or moderate correlations with current CDC figures for the whole of the US, and those of a week later and we expect the approach to be somewhat generalisable across diseases and regions.

[2] http://www.cdc.gov/lyme/signs_symptoms/

References

Hervé Abdi. 2003. Partial least square regression (pls regression). *Encyclopedia for research methods for the social sciences*, pages 792–795.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Matteo Cassotti and Francesca Grisoni. 2012. Variable selection methods: an introduction.

Aron Culotta. 2010a. Detecting influenza outbreaks by analyzing twitter messages. *arXiv preprint arXiv:1007.4748*.

Aron Culotta. 2010b. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM.

Steve R Gunn et al. 1998. Support vector machines for classification and regression. *ISIS technical report*, 14.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Bo Han, Timothy Baldwin, and Paul Cook. 2012. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012: Technical Papers*, pages 1045–1062.

Antonio Jimeno-Yepes, Andrew MacKinlay, and Bo Han. 2015a. Investigating public health surveillance using twitter. *ACL-IJCNLP 2015*, page 164.

Antonio Jimeno-Yepes, Andrew MacKinlay, Bo Han, and Qiang Chen. 2015b. Identifying diseases, drugs, and symptoms in twitter. *Studies in health technology and informatics*, 216:643–647.

Thorsten Joachims. 1999. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 19(4).

Guo-Zheng Li and Xue-Qiang Zeng. 2009. Feature selection for partial least square based dimension reduction. In *Foundations of Computational Intelligence Volume 5*, pages 3–37. Springer.

Guo-Zheng Li, Xue-Qiang Zeng, Jack Y Yang, and Mary Qu Yang. 2007. Partial least squares based dimension reduction with gene selection for tumor classification. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pages 1439–1444. IEEE.

Tahir Mehmood, Harald Martens, Solve Sæbø, Jonas Warringer, and Lars Snipen. 2011. A partial least squares based algorithm for parsimonious variable selection. *Algorithms for Molecular Biology*, 6(1):1.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272.

Michael J Paul and Mark Dredze. 2012. A model for mining public health topics from twitter. *Health*, 11:16–6.

Ari Seifter, Alison Schwarzwalder, Kate Geis, and John Aucott. 2010. The utility of google trends for epidemiological research: Lyme disease as an example. *Geospatial health*, 4(2):135–137.

S Wold et al. 1995. Pls for multivariate linear modeling. *Chemometric methods in molecular design*, 2:195.