On Infectious Intestinal Disease Surveillance using Social Media Content

Bin Zou
Department of Computer Science, University College London, United Kingdom
b.zou@cs.ucl.ac.uk

Vasileios Lampos
Department of Computer Science, University College London, United Kingdom
v.lampos@ucl.ac.uk

Russell Gorton
Field Epidemiology Services, Public Health England, United Kingdom
russell.gorton@phe.gov.uk

Ingemar J. Cox
Department of Computer Science, University College London, United Kingdom
and University of Copenhagen, Denmark
i.cox@ucl.ac.uk

ABSTRACT
This paper investigates whether infectious intestinal diseases (IIDs) can be detected and quantified using social media content. Experiments are conducted on user-generated data from the microblogging service, Twitter. Evaluation is based on the comparison with the number of IID cases reported by traditional health surveillance methods. We employ a deep learning approach for creating a topical vocabulary, and then apply a regularised linear (Elastic Net) as well as a nonlinear (Gaussian Process) regression function for inference. We show that like previous text regression tasks, the nonlinear approach performs better. In general, our experimental results, both in terms of predictive performance and semantic interpretation, indicate that Twitter data contain a signal that could be strong enough to complement conventional methods for IID surveillance.

Keywords
user-generated content; social media; Twitter; infectious intestinal disease; IID; disease surveillance; word embeddings

1. INTRODUCTION
A number of papers have demonstrated that online user-generated content (UGC) contains a significant amount of information about the actual offline behaviour or state of users, either at a collective or a more personalised level, [1, 7, 8, 10, 14, 19, 21, 23]. Several studies have also focused on the domain of health, developing applications that range from online disease surveillance [2, 4, 11, 12] to the assessment of health interventions [13], or health-related qualitative analyses [18, 25]. For disease surveillance, UGC has the main advantage of being a real-time data source, compared to traditional surveillance methods, where data may take days, weeks, or months to collect. In addition, it may also represent segments of the population that do not visit a medical facility, thereby providing health information on a complementary segment of the population. However, UGC data contains inaccurate and ambiguous information which makes interpretation challenging.

In this paper, we build on previous work and present the first effort to model infectious intestinal diseases (IIDs) from social media content. IIDs have a number of characteristics that are distinct from diseases that have been previously investigated using UGC data, such as influenza [3, 6, 12]. Specifically:

- IIDs originating from a single organism (virus, bacterium) are usually of a smaller prevalence in the population. As a result, their signal in social media is expected to be weaker and therefore, harder to detect.
- Most people who are affected by an IID do not seek medical attention [22, 24].
- Finally, self-diagnosis in UGC (e.g. as in “I am down with the flu”) is less frequent, resulting to sparser textual feature representations; for example, a feature as informative as the keyword ‘flu’ does not exist.

Laboratory confirmation of an IID may take several days. Hence, social media could play an important role in providing complementary as well as more timely information for an emerging IID outbreak.

2. DATA SET DESCRIPTION
Two data streams are used in our experiments: Twitter data and official health surveillance records obtained from Public Health England (PHE).

2.1 Twitter data
Tweets were retrieved using the Twitter API.1 Approximately 585 million tweets geolocated in England over a

1https://dev.twitter.com/overview/documentation
2.2 IID surveillance data

To train and evaluate our models, we use weekly IID surveillance reports from PHE. In particular, we focus on laboratory confirmed cases of (i) *Campylobacter* and (ii) *Norovirus* (the most recurrent organisms related to IIDs according to PHE reports). We also consider (iii) *food poisoning* notifications reported by registered medical practitioners across England. The laboratory confirmed data cover a period from 09/04/2012 to 14/06/2015 (166 weeks in total). The food poisoning notifications are from 09/04/2012 to 09/03/2014 (100 weeks in total).

3. METHODS

We first identify a set of keywords related to linguistic expression of IIDs. These keywords are used to formulate our feature space (n-grams) in a regression scenario, where health surveillance indicators are our target variable. Two learning approaches are applied, a regularised linear regressor, known as the Elastic Net [26], and a nonlinear one, based on the framework of the Gaussian Processes (GP) [20].

3.1 Formulating an IID vocabulary using deep learning

We learn Twitter-based word embeddings by training a skip-gram model [17] with hierarchical softmax sampling. We use a layer size of 256, the entirety of a tweet as our look-up window, and the *gensim* implementation. Through the application of a generalisation of the multiplicative cosine similarity proposed by Levy and Goldberg [15], we compute a similarity score $S$ between each keyword’s embedding $q$ and a topic $\tau$. A topic is defined by the embeddings of a small set of $N$ related 1-grams $\{g_1, \ldots, g_N\}$ in conjunction with a set of $M$ unrelated ones $\{z_1, \ldots, z_M\}$. The latter is used to refine the word selection. For example, ‘Bieber’ is often used in conjunction with ‘fever’ (i.e., ‘Bieber fever’) to refer to excitement surrounding the entertainer Justin Bieber, and not to a disease symptom. Thus, while the concept of fever as a disease symptom may be relevant to our purpose, the concept of excitement is not. The similarity score $S(q, \tau)$ is then defined by

$$S(q, \tau) = \frac{\cos(q, g_1) \times \cos(q, g_2) \times \ldots \times \cos(q, g_N)}{\cos(q, z_1) \times \ldots \times \cos(q, z_M) + \epsilon},$$

where $\epsilon = .001$ is used to prevent a division by zero, and cosine similarities are transformed to the interval $[0, 1]$ via $(x + 1)/2$ to avoid negative sub-scores. To define $\tau$, we use a set of IID symptoms, namely $g = \{\text{vomit}, \text{indigestion}, \text{heartburn}, \text{nausea}, \text{reflux}, \text{diarrhea}, \text{hiccups}\}$, and a few potentially helpful keywords in facilitating a disambiguation between IID and other relevant diseases, namely $z = \{\text{flu}, \text{cold}\}$. After computing $S$ for all the keywords in the processed Twitter corpus, the IID vocabulary is determined by the ones with the highest scores (see Section 4 for more details). A manual inspection may be required to determine the cutoff similarity score, i.e., the point where keywords begin to deviate a lot from the target topic.

3.2 Linear regression via the Elastic Net

For a set of $M$ n-grams and a set of $N$ time intervals, we form a matrix $X \in \mathbb{R}^{N \times M}$, which holds the frequency of these n-grams on the Twitter corpus for each time interval. Frequencies are computed by dividing the count of an n-gram with the total number of tweets (per time interval). For the same set of time intervals, we also obtain the target variable $y \in \mathbb{R}^N$ from PHE’s reports. In this regression task, for a single time interval $t$, our aim is to learn a set of weights $w$ such that

$$y_t = w^T x_t + \beta + \epsilon,$$

where $y_t$ and $x_t$ denote the values of $y$ and $X$ during $t$, $\beta \in \mathbb{R}$ is the regression’s intercept, and $\epsilon$ is independent, zero-centred noise. Previous work has shown the superiority of Elastic Net [26] in solving similar text regression tasks in comparison to other linear alternatives such as ridge regression or lasso [9, 13]. Elastic Net combines L1 and L2 norm regularisation, encouraging sparsity as well as avoiding model selection inconsistencies [5], and is defined by

$$\min_{w, \beta} \left( \sum_{t=1}^N (w^T x_t + \beta - y_t)^2 + \lambda_1 \sum_{j=1}^M |w_j| + \lambda_2 \sum_{j=1}^M w_j^2 \right),$$

where $\lambda_1$ and $\lambda_2$ are the corresponding regularisation coefficients. Lambdas are chosen using grid-search, and after presetting $\lambda_1 = 2\lambda_2$ to reduce the degrees of freedom.

---

1 https://radimrehurek.com/gensim/models/word2vec.html
3.3 Nonlinear regression via a Gaussian Process covariance function

To further explore potential nonlinearities in the relationship between the n-gram frequencies on Twitter and the target variable, we use Elastic Net’s positively weighted features in a GP [20], similarly to a recently proposed model for flu rate estimation from search query data [12].

GPs are sets of random variables, any number of which have a multivariate Gaussian distribution. In GP regression, given the inputs \( x \) and \( x' \) (both \( \in \mathbb{R}^n \), where \( Q \) here denotes the number of positively weighted n-grams in the Elastic Net output), we want to learn a function \( f: \mathbb{R}^n \to \mathbb{R} \) such that \( f \sim GP(\mu(x), C(x, x')) \), where \( \mu(\cdot) \) and \( C(\cdot, \cdot) \) denote the mean and covariance (or kernel) functions respectively.

Following the approach in [13], we attempt to capture the potentially distinctive semantics of the n-gram categories \((1 \leq n \leq 3)\) using a different kernel. Given the small number of 3-grams selected by the Elastic Net, we group the 2-grams and 3-grams and only separate them from 1-grams. We define the mean and covariance functions similarly to Lampos et al. in [12] using a squared exponential kernel as our main component (instead of the rational quadratic kernel function).

4. EXPERIMENTAL RESULTS

To create vector space representations of the Twitter corpus, we first extract all n-grams \((1 \leq n \leq 3)\) from \( T_1 \); to form an n-gram, we filter out a list of common English stop words,\(^3\) and then use a look ahead window equal to the length of each tweet (i.e. many n-grams are formed by tokens that were nearby, but not next to each other inside a tweet). We filter low-volume information by keeping n-grams that appear more than 700 times. This yields 47,049 1-grams, 390,593 2-grams, and 152,329 3-grams. After applying the procedure described in Section 3.1 on \( T_2 \), we form a vocabulary \( S_{ID} \) of 597 1-grams that have the highest multiplicative cosine similarity with the predefined IID topic\(^3\).

The applied list of English stop words was a concatenation of various lists available online.

Table 1: Performance indicators for the IID indicator inference task from Twitter content in England. Parentheses include the standard deviation of the estimated mean.

<table>
<thead>
<tr>
<th>IID target</th>
<th>Optimal delay</th>
<th>Elastic Net</th>
<th>Gaussian Process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu(MAE) )</td>
<td>( \mu(r) )</td>
<td>Aggr. ( r )</td>
</tr>
<tr>
<td>Campylobacter</td>
<td>5 days</td>
<td>.572 (.132)</td>
<td>.625 (.177)</td>
</tr>
<tr>
<td>Norovirus</td>
<td>3 days</td>
<td>.554 (.168)</td>
<td>.596 (.142)</td>
</tr>
<tr>
<td>Food poisoning</td>
<td>5 days</td>
<td>.700 (.180)</td>
<td>.702 (.123)</td>
</tr>
</tbody>
</table>

\(^3\)Both cutoff thresholds (597 and 212 top terms) have been decided through manual inspection.

\(^4\)This threshold is configured dynamically so that an adequate number of features is kept each time.

\(^5\)Given the small sample (2 folds only) in the experiment for modelling food poisoning, we could not assess its statistical significance.
linear correlation is .633, .607 and .711, according to the better-performing GP model. Figures 2, 3, and 4 present the GP inferences in all the folds for the three case studies.

We also estimate an aggregated correlation by concatenating the inferences of all folds. This yields correlations that are greater than .7 (up to .77) for all target variables under the GP model. Looking at the average peak-MAE performance figures (Table 2), we see that the performance gap between Elastic Net and GP models increases, emphasising the value of a nonlinear approach when the IID signal gains a significant presence.

5. CONCLUSIONS AND FUTURE WORK

We have presented a basic regression framework for inferring IID occurrences (reported by PHE) from Twitter in England. In contrast to previous work, the original set of features (vocabulary of n-grams) was created using a deep learning approach. The nonlinear regression method (Gaussian Process) outperformed a strong linear alternative (Elastic Net). Overall, we observed good predictive accuracy in all case studies with average linear correlations (between the inferred and target variables) ranging from .607 to .711. We also determined the optimal delay (3-5 days) between Twitter postings and ground truth, indicating that social media may be capable of issuing an earlier warning for an emerg-
6. ACKNOWLEDGMENTS

This research has been supported by the EPSRC IRC grant EP/K031953/1 ("Early-Warning Sensing Systems for Infectious Diseases"). We would like to thank Jens K. Geyti as well as the Twitter data sets, and Public Health England for providing surveillance data.

7. REFERENCES