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

[1]https://dev.twitter.com/overview/documentation

Digital Health 2016 Montreal, Canada

DOI: 00.000/000_0

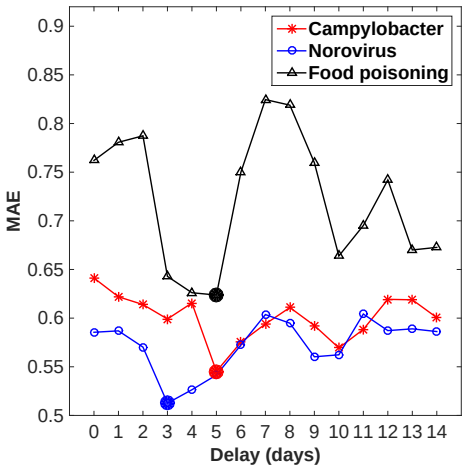

Figure 1: Average inference MAE for the GP model for different time shifts of the Twitter content (days of delay from 0 to 14).

period of 166 weeks from 09/04/2012 to 14/06/2015 were collected. Geolocation was performed either by geocoding the user's profile information or by taking advantage of the exact user geo-coordinates, when they were available. After removing retweets and tweets with links (since these types of expression are rarely used to phrase a health problem), the final Twitter data set contained approx. 410 million tweets (denoted by $\mathcal{T}_1$). A different Twitter data set has been used to train word embeddings (see Section 3.1). This data set ($\mathcal{T}_2$) is an aggregation of all tweets (approx. 140 million) posted by a set of 100,000 UK users from 01/02/2014 to 22/03/2015 (retweets have been filtered out). These users were randomly selected, keeping their numbers proportionally distributed to the regional population figures of the UK.

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.[2] 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 τ. 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 $\mathrm{S}(q, \tau)$ is then defined by

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

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 τ, we use a set of IID symptoms, namely $g = \{$*vomit*, *indigestion*, *heartburn*, *nausea*, *reflux*, *diarrhea*, *hiccups*$\}$, and a few potentially helpful keywords in facilitating a disambiguation between IID and other relevant diseases, namely $z = \{$*flu*, *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 $\mathbf{X} \in \mathbb{R}^{N \times M}$, 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 $\mathbf{y} \in \mathbb{R}^N$ 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

$$y_t = \mathbf{w}^\intercal \mathbf{x}_t + \beta + \epsilon\,,$$

where y_t and $\mathbf{x}_t$ denote the values of $\mathbf{y}$ and $\mathbf{X}$ during t, $\beta \in \mathbb{R}$ is the regression's intercept, and ϵ 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

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left(\sum_{t=1}^{N} (\mathbf{w}^\intercal \mathbf{x}_t + \beta - y_t)^2 + \lambda_1 \sum_{j=1}^{M} |w_j| + \lambda_2 \sum_{j=1}^{M} w_j^2 \right),$$

where λ_1 and λ_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.

[2] https://radimrehurek.com/gensim/models/word2vec.html

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

		Elastic Net			Gaussian Process		
IID target	Optimal delay	μ(MAE)	$\mu(r)$	Aggr. r	μ(MAE)	$\mu(r)$	Aggr. r
Campylobacter	5 days	.572 (.132)	.625 (.177)	.701	.545 (.114)	.633 (.190)	.724
Norovirus	3 days	.554 (.168)	.596 (.142)	.677	.513 (.169)	.607 (.158)	.730
Food poisoning	5 days	.700 (.180)	.702 (.123)	.727	.624 (.004)	.711 (.141)	.771

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 $\mathbf{x}$ and $\mathbf{x}'$ (both $\in \mathbb{R}^Q$, where Q here denotes the number of positively weighted n-grams in the Elastic Net output), we want to learn a function f: $\mathbb{R}^Q \rightarrow \mathbb{R}$ such that $f \sim \mathcal{GP}(\mu(\mathbf{x}), C(\mathbf{x}, \mathbf{x}'))$, 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 $\mathcal{T}_1$; 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 $\mathcal{T}_2$, we form a vocabulary $\mathcal{S}_{\text{IID}}$ of 597 1-grams that have the highest multiplicative cosine similarity with the predefined IID topic (including the positive keywords used to formulate the similarity score). We use a smaller subset of $\mathcal{S}_{\text{IID}}$, $\mathcal{S}^*_{\text{IID}}$ (the top 212 1-grams), which contains fewer extraneous words, to perform keyword matching with 2- and 3-grams from $\mathcal{T}_1$.[4] This process produces the final set of textual features used in our experiments, containing 597 1-grams, 928 2-grams, and 122 3-grams. Weekly term counts are normalised using the total number of tweets published in a week.

We evaluate our model via k-fold cross validation, dividing the data into k consecutive time periods (using a week as our main time unit). We set $k = 8$ for the campylobacter and norovirus experiments, and $k = 2$ when modelling food poisoning cases, given the smaller time span of the data. When applying Elastic Net, we use the same values for the regularisation parameters (λ_1, λ_2) in all folds, and in each fold's training set we pre-filter features by applying a soft linear correlation threshold with the corresponding ground truth.[5] The GP model is applied on the positively weighted features selected by the Elastic Net (per fold). We use Mean Absolute Error (MAE) and Pearson correlation (r) to measure the performance of the models; note that $\mathbf{y}$ has been standardised (zero mean, standard deviation of 1) throughout our experiments, so that the MAEs for the different target variables are comparable with each other. We also separately compute MAE for the 'peaking' periods (peak-MAE), where the ground truth is greater than its mean value, to assess the performance of the models during periods of increased incidence of an IID. Given that all ground truth indications embody a delay of days, we repeat our experimental process considering a delay d from 0 to 14 days (Twitter data are shifted back d days). We then determine the optimal d based on the minimum average MAE derived from the cross-validation.

Average MAEs for all the investigated delays under the GP are presented in Figure 1. Initially the MAE decreases as the delay increases, until an optimal performance is reached (3 to 5 days). Notably, the Elastic Net model points to the exact same optimal delays as the GP model. Table 1 enumerates the results for the two methods at their respective optimal delays. The GP method outperforms Elastic Net; the difference in their mean performance (using MAE) is statistically significant according to a Kolmogorov-Smirnov test ([16]; $p < .05$).[6] For campylobacter, norovirus and food poisoning, the average MAE between inferences and standardised target values is .545, .513 and .624, whereas their

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

Table 2: Comparison of the inference performance (average MAE) methods using the optimal delay (see Table 1), when the IID activity is above its mean value.

	μ(MAE)	
IID target	**Elastic Net**	**Gaussian Process**
Campylobacter	.623 (.214)	.562 (.186)
Norovirus	.790 (.227)	.732 (.271)
Food poisoning	.927 (.287)	.802 (.045)

[4]Both cutoff thresholds (597 and 212 top terms) have been decided through manual inspection.

[5]This threshold is configured dynamically so that an adequate number of features is kept each time.

[6]Given the small sample (2 folds only) in the experiment for modelling food poisoning, we could not assess its statistical significance.

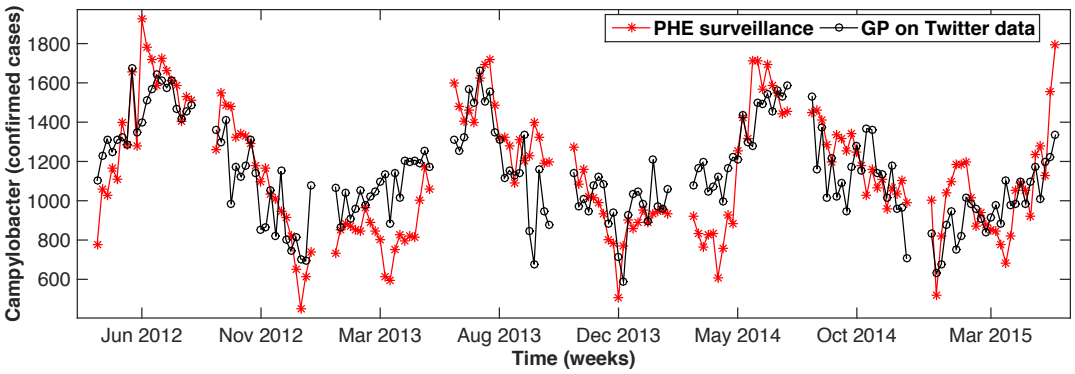

Figure 2: Comparative plot between laboratory confirmed campylobacter cases in England (reported by PHE) and the indication inferred from Twitter content based on the GP model. The gaps separate the folds in the 8-fold cross validation process.

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 emerging IID outbreak. Given that our approach is based on the textual analysis of tweets, it can also be extended to other forms of textual UGC, such as mobile phone communications, different social networks or search query logs.

Our work has the same limitations as most research efforts operating on social media data. The information may be noisy and the underlying semantic notions may not have been disambiguated properly. Furthermore, our experiments cover a small period of 166 weeks (100 weeks for food poisoning cases), which may not be adequate enough to make solid conclusions. Moreover, the evaluation needs to be improved by obtaining more representative ground truth, as the current experiments are based on either laboratory confirmed data (which are generally sparse) or food poisoning cases.

In the future, we plan to incorporate more data from different sources, such as search query logs, to enhance our user-generated signal. We will also extend our focus, by looking at specific IID outbreaks in England, to determine whether these were also evident in and predictable from UGC. Finally, we aim to develop an unsupervised learning technique based on a more thorough, and deep understanding of natural language. This will reduce any biases introduced by the potentially inaccurate ground truth, and will also assist in creating an independent, complementary sensor for infectious diseases.

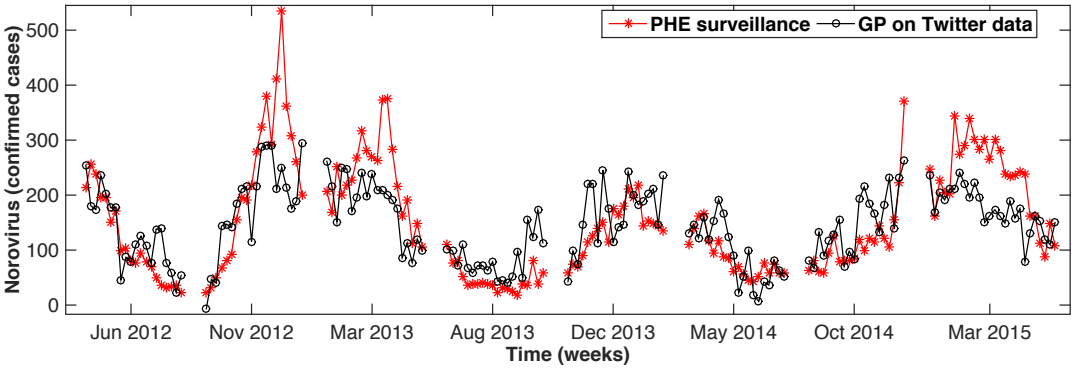

Figure 3: Comparative plot between laboratory confirmed norovirus cases in England (reported by PHE) and the indication inferred from Twitter content based on the GP model. The gaps separate the folds in the 8-fold cross validation process.

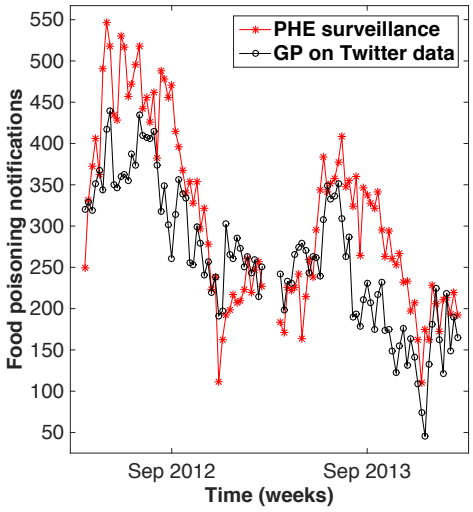

Figure 4: Comparative plot between food poisoning cases in England (reported by PHE) and the indication inferred from Twitter content based on the GP model. The gaps separate the folds in the 2-fold cross validation process.

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 for assisting in the preparation of Twitter data sets, and Public Health England for providing surveillance data.

7. REFERENCES

[1] D. J. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating Gender on Twitter. In *Proc. of EMNLP*, pages 1301–1309, 2011.

[2] A. Culotta. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In *Proc. of the 1st Workshop on Social Media Analytics*, pages 115–122, 2010.

[3] A. Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Lang. Resour. Eval.*, 47(1):217–238, 2013.

[4] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting Influenza Epidemics using Search Engine Query Data. *Nature*, 457(7232):1012–1014, 2009.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[6] A. Lamb, M. J. Paul, and M. Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Proc. of NAACL-HLT*, pages 789–795, 2013.

[7] V. Lampos. *Detecting Events and Patterns in Large-Scale User Generated Textual Streams with Statistical Learning Methods*. Ph.D. Thesis, University of Bristol, 2012.

[8] V. Lampos, N. Aletras, J. K. Geyti, B. Zou, and I. J. Cox. Inferring the Socioeconomic Status of Social Media Users based on Behaviour and Language. In *Proc. of ECIR*, 2016.

[9] V. Lampos, N. Aletras, D. Preoţiuc-Pietro, and T. Cohn. Predicting and Characterising User Impact on Twitter. In *Proc. of EACL*, pages 405–413, 2014.

[10] V. Lampos and N. Cristianini. Nowcasting Events from the Social Web with Statistical Learning. *ACM Trans. Intell. Syst. Technol.*, 3(4):72:1–72:22, 2012.

[11] V. Lampos, T. De Bie, and N. Cristianini. Flu Detector: Tracking Epidemics on Twitter. In *Proc. of ECML PKDD*, pages 599–602, 2010.

[12] V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen. Advances in Nowcasting Influenza-like Illness Rates using Search Query Logs. *Sci. Rep.*, 5(12760), 2015.

[13] V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. Assessing the Impact of a Health Intervention via User-generated Internet Content. *Data Min. Knowl. Discov.*, 29(5):1434–1457, 2015.

[14] T. Lansdall-Welfare, V. Lampos, and N. Cristianini. Nowcasting the Mood of the Nation. *Significance*, 9(4):26–28, 2012.

[15] O. Levy and Y. Goldberg. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proc. of CoNLL*, pages 171–180, 2014.

[16] J. R. Massey and J. Frank. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.*, 46(253):68–78, 1951.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS 26*, pages 3111–3119, 2013.

[18] M. J. Paul and M. Dredze. Discovering Health Topics in Social Media Using Topic Models. *PLoS ONE*, 9(8):e103408, 2014.

[19] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE*, 10(9):e0138717, 2015.

[20] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[21] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proc. of WWW*, pages 851–860, 2010.

[22] C. C. Tam et al. Longitudinal Study of Infectious Intestinal Disease in the UK (IID2 Study): Incidence in the Community and Presenting to General Practice. *Gut*, 61(1):69–77, 2012.

[23] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proc. of ICWSM*, pages 178–185, 2010.

[24] J. G. Wheeler et al. Study of Infectious Intestinal Disease in England: Rates in the Community, Presenting to General Practice, and Reported to National Surveillance. *BMJ*, 318(7190):1046–1050, 1999.

[25] E. Yom-Yov, L. Fernandez-Luque, I. Weber, and S. P. Crain. Pro-Anorexia and Pro-Recovery Photo Sharing: A Tale of Two Warring Tribes. *J. Med. Internet Res.*, 14(e151), 2012.

[26] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *J Roy. Stat. Soc. B Met.*, 67(2):301–320, 2005.