Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance

Vasileios Lampos♠
v.lampos@ucl.ac.uk

Bin Zou♠
bin.zou.14@ucl.ac.uk

Ingemar J. Cox♠,♣
i.cox@ucl.ac.uk

♠ Department of Computer Science, University College London, London WC1E 6BT, United Kingdom
♣ Department of Computer Science, University of Copenhagen, Copenhagen 2200, Denmark

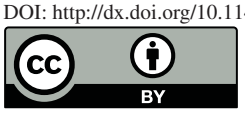
ABSTRACT

Health surveillance systems based on online user-generated content often rely on the identification of textual markers that are related to a target disease. Given the high volume of available data, these systems benefit from an automatic feature selection process. This is accomplished either by applying statistical learning techniques, which do not consider the semantic relationship between the selected features and the inference task, or by developing labour-intensive text classifiers. In this paper, we use neural word embeddings, trained on social media content from Twitter, to determine, in an unsupervised manner, how strongly textual features are semantically linked to an underlying health concept. We then refine conventional feature selection methods by a priori operating on textual variables that are sufficiently close to a target concept. Our experiments focus on the supervised learning problem of estimating influenza-like illness rates from Google search queries. A "flu infection" concept is formulated and used to reduce spurious —and potentially confounding— features that were selected by previously applied approaches. In this way, we also address forms of scepticism regarding the appropriateness of the feature space, alleviating potential cases of overfitting. Ultimately, the proposed hybrid feature selection method creates a more reliable model that, according to our empirical analysis, improves the inference performance (Mean Absolute Error) of linear and nonlinear regressors by 12% and 28.7%, respectively.

Keywords

Computational Health; Syndromic Surveillance; Influenza-Like Illness; User-Generated Content; Search Query Logs; Feature Selection; Word Embeddings; Regularised Regression; Gaussian Processes

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
DOI: http://dx.doi.org/10.1145/3038912.3052622



1. INTRODUCTION

Online user-generated content (UGC), primarily in the form of social media posts or search query logs, has been the focus of considerable research effort in recent years. It has facilitated methods, interpretations and inferences in various scientific areas, such as Computational Linguistics [19, 47], Behavioural Sciences [28, 55], Computational Social Science [3, 24] and Computational Health [8, 20, 37, 56], among many others.

A common paradigm, evident in many of these works, is the formulation of a supervised learning task based on a textual representation of UGC [10, 53]. This often involves a large number of features, but a moderate number of training samples, encouraging the application of statistical methods that are able to project the data to a lower dimensional space or maintain the most relevant predictors [33, 34, 36, 48]. A valid criticism of such approaches is that some of the selected features may have little or no semantic link to the regression task, increasing the possibility of overfitting, especially in situations where spurious correlations are observed. To alleviate this effect, methods in Natural Language Processing (NLP) have incorporated classification schemes or have routinely used lexical taxonomies, aiming to encourage a relatedness between the input information and the target thematic concept [2, 7, 11, 49]. However, these operations tend to require an extensive human effort, especially in obtaining a sufficient number of labelled outputs, and are limited to a specific task.

In this paper, we take advantage of current developments in statistical NLP and propose a method to address the aforementioned deficiencies. We form general textual concepts by adopting neural word embeddings [44], and then use them in conjunction with conventional feature selection methods to encourage a level of topicality in the selected predictors within a text regression task. This approach can be regarded as an unsupervised classification layer that favours textual features that belong to a target theme of interest. We use this method to improve feature selection for a large-scale, practical, and well-studied text regression task, specifically the inference of influenza-like illness (ILI) rates from time series of search query frequencies [20, 35, 59].

Monitoring disease rates from online activity can complement the existing health surveillance infrastructure, as it provides access to a larger part of the population including

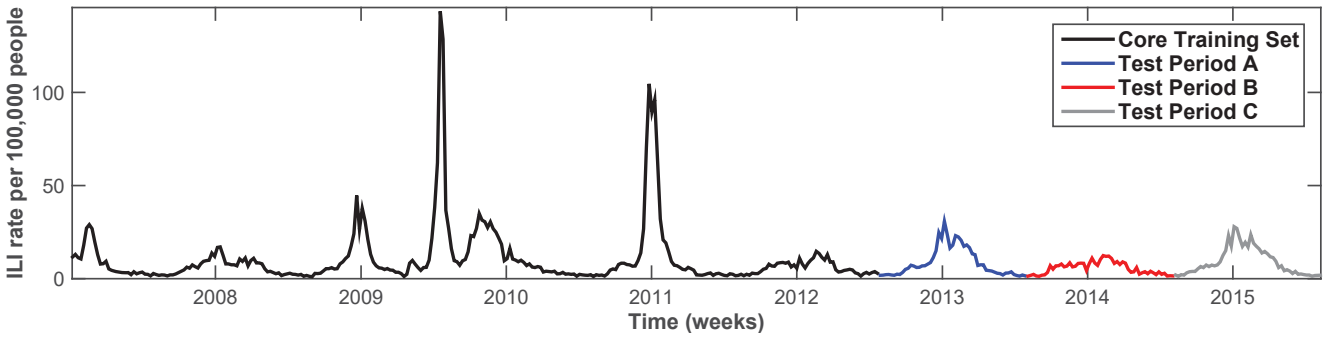


Figure 1: Weekly influenza-like illness (ILI) rates in England (per 100,000 people) from January 1, 2007 to August 9, 2015 obtained by RCGP and PHE. Training and test periods are denoted with different colourings.

individuals who do not visit a medical facility [35, 50]. Further advantages are the more timely and less costly disease rate estimates, as well as the ability to acquire information in geographical locations with less established healthcare systems [20, 33]. Whereas previous attempts to model ILI rates from search query logs [50, 20] have reportedly produced misleading outputs [38, 45], follow-up research has corroborated that this was due to inadequacy of the applied statistical framework [35, 59].

Here we report on findings that improve on the current state-of-the-art approaches, with a clear focus on the linguistic side of the task. We train word embeddings using microblogging text snippets from Twitter, so as to capture more direct and informal linguistic patterns that we assume to also be present in search queries. Supervised learning is based on official syndromic surveillance rates for ILI obtained by health agencies. Our empirical analysis shows that the proposed hybrid feature selection method provides significant performance gains (from 12% to 28.7% of relative improvement) under both linear and nonlinear regression functions. Qualitative insights indicate that this is due to the inherent topicality of the selected textual features as many spurious —and potentially confounding— queries are being automatically removed.

The paper's main contributions are the following:

1. We introduce a new **unsupervised approach for selecting textual features** that are relevant to a target concept without solely relying on statistical metrics, such as correlation or regression analysis.

2. The aforementioned approach is combined with conventional feature selection techniques, creating a **hybrid method that significantly improves model reliability** and, consequently, the inference performance under linear as well as nonlinear regressors.

3. From an applied perspective, we focus on an important health-related task, i.e. the **estimation of ILI rates in a population**.[1]

2. DATA SETS

We aim to infer influenza-like illness (ILI) rates as reported by the Royal College of General Practitioners (RCGP) and Public Health England (PHE).[2] RCGP/PHE estimates represent the number of doctor consultations reporting ILI symptoms per 100,000 people in England. Their weekly time series from January 1, 2007 to August 9, 2015 are displayed in Figure 1; different colourings denote training and testing periods (see Section 4 for a detailed reference).

Our input user-generated data set is a time series of search query frequencies. These are a non standardised version of the publicly available Google Trends outputs and were retrieved through a private Google Health Trends API, provided for academic research with a health-oriented focus. A query frequency expresses the probability of a short search session[3] for a specific geographical region and temporal resolution, drawn from a uniformly distributed 10%-15% sample of all corresponding Google search sessions.[4] We have used a set of 35,572 search queries (examples of which are presented in Table 1) and obtained their weekly frequency in England during an extensive period of 449 weeks ($\approx$ 8.6 years), from January 1, 2007 to August 9, 2015.

As we had no access to a raw, user-level corpus of search queries, we used a Twitter data set to learn word embeddings, aiming to capture more informal or direct ways of written expression. We collected tweets from users located in the United Kingdom (UK) to accommodate geographically constrained dialects and conversation themes. We also made an effort to maintain a user distribution that is proportional to regional UK population figures. The total number of tweets was approximately 215 million, dated from February 1, 2014 to March 31, 2016. We applied the `word2vec` neural embedding algorithm [43, 44] as implemented in the `gensim` library.[5] We have used a continuous bag-of-words representation, the entirety of a tweet as our window, negative sampling, and a dimensionality of 512. After filtering out the long tail of textual tokens with less than 500 occurrences (in the 215 million tweets) to eliminate potential spam expressions, we obtained an embedding corpus of 137,421

[1]The ILI estimates are showcased on a live web service, the Flu Detector (`fludetector.cs.ucl.ac.uk`) [30]

[2]PHE's weekly national flu reports, `gov.uk/government/statistics/weekly-national-flu-reports`

[3]A search session can be seen as a time window that may include more than one consecutive search queries from a user account. Therefore, a target search query is identified as a part of a potentially larger query set within a search session.

[4]The publicly available Google Trends (`google.com/trends`) represent a smaller sample.

[5]Python library `gensim`, `radimrehurek.com/gensim`

unigrams.[6] Note that we have not optimised `word2vec`'s settings for our task, but the above parametrisation falls within previously reported configurations [1, 51].

To capture and compare with more formal linguistic properties, we also used word embeddings trained on a Wikipedia corpus. The latter were obtained from the work of Levy and Goldberg [39] and have a dimensionality of 300.

3. METHODS

We first give an overview of the linear and nonlinear methods that we use for performing text regression. Then, we describe our approach in utilising word embeddings to create concepts that ultimately refine feature selection and ILI rate inference performance.

3.1 Linear and nonlinear text regression

In regression, we learn a function f that maps an input space $\mathbf{X} \in \mathbb{R}^{n \times m}$ (where n and m respectively denote the number of samples and the dimensionality) to a target variable $\mathbf{y} \in \mathbb{R}^n$. As described in the previous section, our input space $\mathbf{X}$ represents the frequency of m search queries during n (weekly) time intervals. In text regression, we usually operate on a high-dimensional, relatively sparse, textual feature space and a considerably smaller number of samples ($m \gg n$). To avoid overfitting and improve generalisation, a standard approach is to introduce a degree of regularisation during the optimisation of f [23].

In our experiments, we use Elastic Net as our linear regressor [63]. Elastic Net has been broadly applied in many research areas, including NLP [27, 36]. It can be seen as a generalisation of the L1-norm regularisation, known as the *lasso* [57], because it also applies an L2-norm, or *ridge* [25], regulariser on the inferred weight vector. The combination of the two regularisers encourages sparse solutions, thereby performing feature selection, and, at the same time, addresses model consistency problems that arise when collinear predictors exist in the input space [61]. Elastic Net is defined as:

$$\underset{\mathbf{w}^*}{\operatorname{argmin}} \left(\|\mathbf{X}^*\mathbf{w}^* - \mathbf{y}\|_2^2 + \frac{(1-a)\lambda}{2}\|\mathbf{w}\|_2^2 + a\lambda\|\mathbf{w}\|_1 \right), \quad (1)$$

where $\mathbf{w}^* = [\mathbf{w}\ w_0]^\mathsf{T}$, $\mathbf{X}^* = [\mathbf{X}\ \mathbf{1}]$ to incorporate the model's intercept, and a, λ control the level of regularisation.

For completeness, we also experiment with a nonlinear regression method formed by a composite Gaussian Process (GP). Numerous applications have provided empirical evidence for the predictive strength of GPs in Machine Translation tasks, as well as in text and multi-modal regression problems [4, 12, 13, 31, 35, 52]. One caveat is that GPs are not very efficient when operating in high dimensional spaces [9]. Thus, while we perform modelling with a nonlinear regressor, we rely on a pre-selected subset of features. As explained in the next paragraphs, these features are either selected based solely on a statistical analysis or using the hybrid selection approach introduced in this paper (see Section 3.2).

GPs are defined as random variables any finite number of which have a multivariate Gaussian distribution [54]. GP methods aim to learn a function $f : \mathbb{R}^m \to \mathbb{R}$ drawn from a GP prior. They are specified through a mean and a covariance (or *kernel*) function, i.e.

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad (2)$$

where $\mathbf{x}$ and $\mathbf{x}'$ (both $\in \mathbb{R}^m$) denote rows of the input matrix $\mathbf{X}$. By setting $\mu(\mathbf{x}) = 0$, a common practice in GP modelling, we focus only on the kernel function. We use the Matérn covariance function [42] to handle abrupt changes in the predictors given that the experiments are based on a sample of the original Google search data. It is defined as

$$k_{\mathrm{M}}^{(\nu)}(\mathbf{x}, \mathbf{x}') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} r \right), \quad (3)$$

where K_ν is a modified Bessel function, ν is a positive constant,[7] ℓ is the lengthscale parameter, σ^2 a scaling factor (variance), and $r = \mathbf{x} - \mathbf{x}'$. We also use a Squared Exponential (SE) covariance to capture more smooth trends in the data, defined by

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\frac{r^2}{2\ell^2} \right). \quad (4)$$

We have chosen to combine these kernels through a summation. Note that the summation of GP kernels results in a new valid GP kernel [54]. An additive kernel allows modelling with a sum of independent functions, where each one can potentially account for a different type of structure in the data [18]. We are using two Matérn functions ($\nu = 3/2$) in an attempt to model long as well as medium (or short) term irregularities, an SE kernel, and white noise. Thus, the final kernel is given by

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{2} \left(k_{\mathrm{M}}^{(\nu=3/2)}(\mathbf{x}, \mathbf{x}'; \sigma_i, \ell_i) \right) \\ + k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}'; \sigma_3, \ell_3) + \sigma_4^2 \delta(\mathbf{x}, \mathbf{x}'), \end{aligned} \quad (5)$$

where δ is a Kronecker delta function. Parameters (7 in total) are optimised using the Laplace approximation (under a Gaussian likelihood), as detailed in related literature [4, 35, 54].

The choice of this kernel structure was not arbitrary, but based on some initial experimentation as the combination that provided a better fit to the training data according to the negative log-marginal likelihood metric. More advanced kernels, operating on structured subsets of the feature space (e.g. as in the work by Lampos et al. [35]), may have obtained better performance estimates. However, their application would not have been helpful in the comparative assessment of the feature selection operation as meta-structures (e.g. query clusters) may vary for different selected feature sets.

3.2 Concept formulation and feature selection using word embeddings

Neural word embeddings have been used as an input in various models and tasks in the recent years [5, 21, 26]. Here we are formulating a method based on word embedding similarities to encourage a more topical selection of features. This approach is unsupervised, overcoming the burden of obtaining labels for training a topic classifier.

[6] The UK Twitter word embeddings can be obtained from `dx.doi.org/10.6084/m9.figshare.4052331.v1`

[7] When $\nu \to \infty$, we obtain the SE covariance function.

Table 1: A set of concepts ($\mathcal{C}$) with their defining positive and negative context n-grams, as well as the top most similar search queries (obtained by applying the similarity function defined in Eq. 7). Concepts $\mathcal{C}_1$ to $\mathcal{C}_6$ are based on Twitter content, whereas $\mathcal{C}_7$ is based on Wikipedia articles. Reformulations of a search query with the inclusion of stop words or a different term ordering are not shown.

ID	Concept	Positive context	Negative context	Most similar search queries
$\mathcal{C}_1$	flu infection	#flu, fever, flu, flu medicine, gp, hospital	bieber, ebola, wikipedia	cold flu medicine, flu aches, cold and flu, cold flu symptoms, colds and flu, flu jab cold, tylenol cold and sinus, flu medicine, cold sore medication, cold sore medicine, flu, home remedy for sinus infection, home remedies for sinus infection, cold flu remedies
$\mathcal{C}_2$	flu infection	flu, flu fever, flu symptoms, flu treatment	ebola, reflux	flu, flu duration, flu mist, flu shots, cold and flu, how to treat the flu, flu near you, 1918 flu, colds and flu, sainsburys flu jab, flu symptoms, cold vs flu symptoms, cold vs flu, cold flu symptoms, flu jab, avian flu, bird flu, flu jabs, flu jab cold, influenza flu
$\mathcal{C}_3$	flu infection	flu, flu gp, flu hospital, flu medicine	ebola, wikipedia	flu aches, flu, colds and flu, cold and flu, cold flu medicine, flu jab cold, flu jabs, flu stomach cramps, flu medicine, sainsburys flu jab, flu stomach pain, cold flu symptoms, baby cold sore, gastric flu, cold sore medication, stomach flu, flu jab, flu mist
$\mathcal{C}_4$	infectious disease	cholera, ebola, flu, hiv, norovirus, zika	diabetes	cholera, cholera outbreak, norovirus outbreak, ebola outbreak, norovirus, virus outbreak, ebola virus, ebola, swine flu outbreak, flu outbreak, haiti cholera, outbreak, swine flu virus, measles outbreak, flu virus, virus, measles virus, influenza a virus
$\mathcal{C}_5$	health	doctors, health, healthcare, nhs	cinema, football	vaccinations nhs, nhs dental, nhs sexual health, nhs nurses, nhs doctors, nhs appendicitis, nhs pneumonia, physiotherapy nhs, nhs prescriptions, nhs physiotherapist, nhs prescription, ibs nhs, health diagnosis, nhs diagnose, nhs medicines, nhs vaccination, mrsa nhs
$\mathcal{C}_6$	gastrointestinal disease	diarrhoea, food poisoning, hospital, salmonella, vomit	ebola, flu	tummy ache, nausea, feeling nausea, nausea and vomiting, bloated tummy, dull stomach ache, heartburn, feeling bloated, aches, belly ache, stomach ache, feeling sleepy, spasms, stomach aches, stomach ache after eating, ache, feeling nauseous, headache and nausea
$\mathcal{C}_7$	flu infection (Wikipedia)	fever, flu, flu medicine, gp, hospital	bieber, ebola, wikipedia	flu epidemic, flu, dispensary, hospital, sanatorium, fever, flu outbreak, epidemic, flu medicine, doctors hospital, flu treatment, influenza flu, flu pandemic, gp surgery, clinic, flu vaccine, flu shot, infirmary, hospice, tuberculosis, physician, flu vaccination

As explained in Section 2, we have used `word2vec` [44] to obtain 512-dimensional embeddings for a set of approximately 137K Twitter tokens. Search queries are projected into the same space by using these embeddings. The underlying assumption is that the informal, direct, and dense language observed in tweets can capture similar characteristics present in search queries.

We consider a search query $\mathcal{Q}$ as a set of t textual tokens, $\{\xi_1, \ldots, \xi_t\}$, where standard English stop words are ignored.[8] The embedding of $\mathcal{Q}$, $\mathbf{e}_\mathcal{Q}$, is estimated by averaging across the embeddings of its tokens, that is

$$\mathbf{e}_\mathcal{Q} = \frac{1}{t}\sum_{i=1}^{t}\mathbf{e}_{\xi_i}\,, \tag{6}$$

where $\mathbf{e}_{\xi_i}$ denotes the Twitter-based embedding of a search query token ξ_i. Using word embeddings we also form themes of interest, and we refer to them as concepts. A concept $\mathcal{C}(\mathcal{P},\mathcal{N})$ consists of a set $\mathcal{P}$ of related or *positive* n-grams, $\{P_1,\ldots,P_k\}$, and a set $\mathcal{N}$ of non related or *negative* ones, $\{N_1,\ldots,N_z\}$. $\mathcal{P}$ and $\mathcal{N}$ are also referred to as positive and negative context, respectively. For context n-grams with $n \geq 2$, we retrieve the average embedding across the n 1-grams (in our experiments, we have restricted $n \leq 2$). We then compute a similarity score, $S(\mathcal{Q},\mathcal{C})$, between query embeddings and the formulated concept, using the following similarity function:

$$S(\mathcal{Q},\mathcal{C}) = \frac{\sum_{i=1}^{k}\cos(\mathbf{e}_\mathcal{Q},\mathbf{e}_{P_i})}{\sum_{j=1}^{z}\cos\left(\mathbf{e}_\mathcal{Q},\mathbf{e}_{N_j}\right)+\gamma}\,. \tag{7}$$

The numerator and denominator of Eq. 7 are sums of cosine similarities between the embedding of the search query and each positive or negative concept term respectively. All cosine similarities (θ) are transformed to the interval $[0,1]$ through $(\theta+1)/2$ to avoid negative sub-scores, a $\gamma = 0.001$ is added to the denominator to prevent division with zero, and we always set $k > z$ so that the positive similarity part is more dominant than the negative. Eq. 7 combines the notion of the additive similarity with the multiplicative one as it chooses to divide instead of subtracting with the negative context [40, 41]. However, we note that the extension applied here has not received a dedicated evaluation in the literature, something hard given its unconstrained nature, i.e. the use of multiple positive and negative context terms.

Table 1 lists the concepts we formed and experimented with in our empirical analysis together with the most similar (according to Eq. 7) search queries. We provide more insight in Section 4. After deriving a concept similarity score (S) for each search query, we begin filtering out queries that are below the mean score (μ_S), and refine this further using standard deviation steps (σ_S). Essentially, this creates an unsupervised query topic classifier, where the only driver is a few contextual keywords that may need to be manually

[8]We use a standard English language stop word list as defined in the NLTK software library (`nltk.org`).

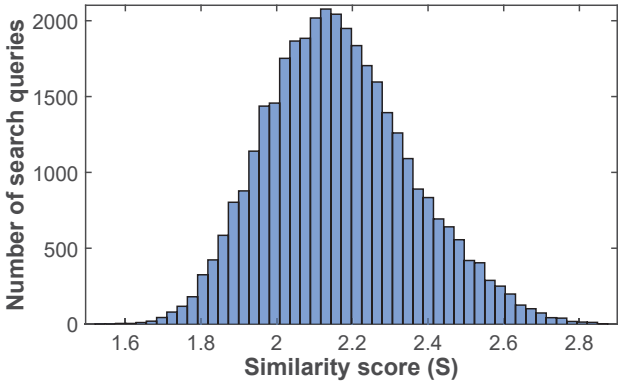


Figure 2: Histogram presenting the distribution of the search query similarity scores (S; see Eq. 7) with the flu infection concept $\mathcal{C}_1$.

decided, perhaps with the assistance of an expert. As described in the following sections, the optimal performance is obtained when a broad version of this similarity based filter is combined with more traditional feature selection methods.

4. EXPERIMENTS

We first assess the predictive capacity of the **word embedding based feature selection** method in inferring ILI rates in England, using Elastic Net. We then present strong performance baselines obtained by selecting the input features to Elastic Net based on their bivariate Pearson correlation with the target variable. We use the term **correlation based feature selection** to refer to this combination of bivariate linear correlation and Elastic Net regression. Finally, we propose a **hybrid combination** of the above approaches, showcasing significant performance gains. The selected features from the various investigated feature selection approaches are also tested under the GP regressor described in Section 3.

We evaluate performance based on three metrics: Pearson correlation (r_y),[9] Mean Absolute Error (MAE), and Mean Absolute Percentage of Error (MAPE) between the inferred and target variables. We assess predictive performance on three flu seasons (2012/13, 2013/14, 2014/15; test periods A, B, and C respectively), each one being a year-long period (see Fig. 1). We train on past data (all weeks prior to a flu season), emulating a realistic evaluation setup. To train an Elastic Net model, we set $a = 0.5$,[10] and decide the value of λ automatically by validating it on a held-out stratified subset ($\approx 7\%$) of the training set.

4.1 Feature selection using word embeddings

The first row of Table 1 describes concept $\mathcal{C}_1$, which we refer to as *flu infection*, that was chosen as the main concept for our experimental evaluation. The rational behind $\mathcal{C}_1$ is straightforward: the search queries that are relevant to our task should be about the topic of flu, with a certain focus on content that is indicative of infection. Hence, the positive context is formed by strongly topical keywords, such as *flu*, the Twitter hashtag *#flu* or the 2-gram *flu medicine*, as well as more general ones, such as a major symptom (*fever*) and the need for medical attention (*gp*[11] and *hospital*). Likewise, the negative context tries to disambiguate from other infectious diseases (*ebola*), spurious contextual meanings (*bieber* as in 'Bieber fever') and the general tendency of information seeking (*wikipedia*). The most similar search queries to $\mathcal{C}_1$ are indeed about ILI, and relevant symptoms or medication (e.g. *cold flu medicine*, *flu aches* and so on). Alternative concept formulations and their potential impact are explored in Section 4.3.

Table 2: Linear regression (Elastic Net) performance estimates for the word embedding based feature selection. NA (last row) denotes that no word embedding based feature selection has been applied.

$S > \mu_S$	$\lvert\mathcal{Q}\rvert$	r_{train}	$\lvert\mathcal{Q}_{\text{en}}\rvert$	r_y	MAE	MAPE
$+0$	14,798	$-.036$	246	.742	6.791	138.69
$+\sigma_S$	5,160	.106	91	**.897**	3.807	101.74
$+2\sigma_S$	1,047	.599	233	.887	3.182	65.35
$+2.5\sigma_S$	303	.752	33	.867	**3.006**	**61.05**
$+3\sigma_S$	69	.735	56	.784	4.043	77.51
$+3.5\sigma_S$	7	.672	6	.721	6.271	110.80
NA	35,572	.018	174	.800	4.442	112.01

Figure 2 shows the unimodal distribution of the similarity scores (Eq. 7) between $\mathcal{C}_1$ and the embeddings of all search queries in our data set. We use the mean similarity score, $\mu_S = 2.165$, and products of the standard deviation, $\sigma_S = 0.191$, to define increasingly similar subsets of search queries. We evaluate the predictive performance of each subset using Elastic Net; the results are presented in Table 2. The last row of the table shows the performance of Elastic Net when all search queries are candidate features, i.e. when embedding based feature selection is omitted. Columns $|\mathcal{Q}|$ and $|\mathcal{Q}_{\text{en}}|$ denote the average number of candidate and selected (by receiving a nonzero weight) search queries in the three test periods. We use r_{train} to denote the average aggregate[12] correlation of the data with the ground truth in the training set prior to performing regression. This indicator can be used as an informal metric for the goodness of the unsupervised, word embedding based feature selection. As the feature selection becomes more narrow, i.e. for higher similarity scores, we observe strongly positive correlations which illustrate that the formulated concept succeeds in capturing the target variable.

After applying Elastic Net, the best performing subset includes queries with similarity scores greater than 2.5 standard deviations from the mean. The relative performance improvement as opposed to using all search queries as candidate features in Elastic Net (last row of Table 2) is equal to 32.33% (in terms of MAE), a statistically significant difference according to a t-test ($p = .0028$). This indicates that selecting features via a semantically informed manner is better than solely relying on a naïve statistical approach.

[9] We use r_y to denote a correlation with the target variable y and to disambiguate between other uses of r.

[10] This results into a 1:2 balance between the regularisation factors of the L2-norm and L1-norm of $\mathbf{w}$, respectively.

[11] *gp*, in this context, is an abbreviation for General Practitioner.

[12] Represents the mean frequency of all search queries.

Table 3: Performance results for linear regression (Elastic Net) by applying a correlation based or a hybrid feature selection.

Correlation based feature selection						Hybrid feature selection				
$r >$	$\lvert\mathcal{Q}\rvert$	$\lvert\mathcal{Q}_{\text{en}}\rvert$	r_y	MAE	MAPE	$\lvert\mathcal{Q}^S\rvert$	$\lvert\mathcal{Q}^S_{\text{en}}\rvert$	r_y	MAE	MAPE
0	15,942	127	.560	5.864	134.41	2,275	168	.899	2.772	48.36
.10	3,238	128	.841	4.639	103.95	669	121	.918	2.206	44.57
.20	719	214	.811	3.861	78.39	256	42	.897	2.122	41.70
.30	279	121	.891	2.199	47.49	168	50	**.913**	**1.880**	**36.23**
.40	165	65	.876	2.137	47.15	118	53	.906	2.119	39.08
.50	104	80	.888	2.245	39.74	72	43	.905	2.347	39.90
.60	61	38	.850	2.577	45.47	40	18	.828	2.962	49.52
.70	26	9	.863	3.853	67.11	20	10	.863	3.855	67.17

However, while the obtained performance is quite strong, the correlation based feature selection outperforms it, as we report in the next section.

4.2 Hybrid feature selection using statistical learning and word embeddings

In supervised learning, a common approach for filtering out irrelevant features is performed by checking their bivariate correlation with the target variable [22]. This is often applied prior to training a regression model, as a procedure that can reduce overfitting and offer performance gains (which we also report below). This form of feature selection has been used in the task of ILI rate modelling from social media or search queries [15, 20, 35]. However, a correlation filter is not always successful in removing spurious features (e.g. it usually fails to remove search queries that reflect on seasonal activities, such as *skiing*), and conversely, when a strict correlation threshold is enforced potentially useful predictors may be lost.

To mitigate this effect, we combine correlation based and word embedding based feature selection, creating a hybrid approach. Features selected based on correlation are passed into the embedding based feature selector and only features that exceed a similarity threshold with the target concept are retained. After some preliminary experimentation with the data, a broad similarity threshold was found to provide better results, given that otherwise the number of features becomes relatively small. Thus, in the experiments below, word embedding feature selection maintains queries with a similarity score that is greater than one standard deviation from the mean similarity score (i.e. $S{>}\mu_S{+}\sigma_S$).

Table 3 presents the performance outcomes under Elastic Net for correlation based and hybrid feature selection. The left part enumerates the results for a number of correlation thresholds ($r{>}\rho$, $\rho{\in}[0,1)$), whereas on the right we report the corresponding results using a combination of a correlation and similarity threshold ($r{>}\rho \cap S{>}\mu_S{+}\sigma_S$, $\rho{\in}[0,1)$). Correlation based feature selection improves the performance estimates as opposed to using all the features (last row of Table 2), yielding its best performance, in terms of MAE, for $r{>}.40$. This supports similar findings in the literature [35]. It also outperforms the estimates obtained when the similarity filter is applied alone, something expected given that a correlation is a statistical determinant based on the actual time series of the data, and not just on the textual content of a search query. Focusing on the right side of Table 3, where, based on the hybrid approach, queries that may be sufficiently correlated, but dissimilar to the specified concept are automatically omitted, we observe that the performance is decidedly enhanced, reaching a relative improvement of 12.03% (from 2.137 to 1.880 in terms of MAE, alas not a statistically significant difference according to a t-test). As the correlation filter becomes more strict ($r{>}.50$), the number of features (denoted by $|\mathcal{Q}|$ or $|\mathcal{Q}^S|$) becomes quite small, and the performance drops, regardless of the feature selection method.

Table 4: Non repetitive examples of search queries that are filtered out using the optimal hybrid feature selection settings, together with their prior weights as well as a weight percentage (ratio over the highest positive or lowest negative weight). Test periods are defined in Fig. 1. Redacted text is in *italics*.

Test period	Examples of filtered search queries
A	prof. *surname* (.0536; 70.3%), *name surname* (.0208; 27.2%), heal the world (.0167; 21.9%), heating oil (.0162; 21.2%), *name surname* recipes (.0160; 21.0%), the white company (−.0152; 35.3%), tlc diet (.0102; 13.3%), blood game (.0093; 12.3%), night vision (.0086; 20.1%), swine flu vaccine side effects (.0055; 7.2%), spine world (−.0031; 7.2%), face login (−.0012; 2.7%)
B	flu season (.0164; 22.4%), broken sword (−.0153, 38.3%), size conversion (.0104; 14.2%), *name surname* fat (.0085; 11.6%), i touch (−.0079; 19.7%), special k (.0063; 8.5%), snow rock (.0048; 6.6%), beat it (.0028; 3.8%), gas homecare (.0024; 3.2%), north face (−.0018; 4.4%), low cost flights (−.0014; 3.5%), love and other drugs (−.0014; 3.4%), all saints (−.0011; 2.8%)
C	*name surname* (.1070; 100%), *name surname* (.0511; 47.7%), florence and the machine lungs (−.0229; 38.1%), acia berry (.0223; 20.8%), testicular cancer symptoms (.0165; 15.4%), pleurisy symptoms (−.0161; 26.8%), flu vaccine nhs (.0077; 7.2%), normal temperature (.0066; 6.2%), swine flu vaccination (.0030; 2.8%), jerks (.0024; 2.3%), boots sale (−.0016; 2.6%)

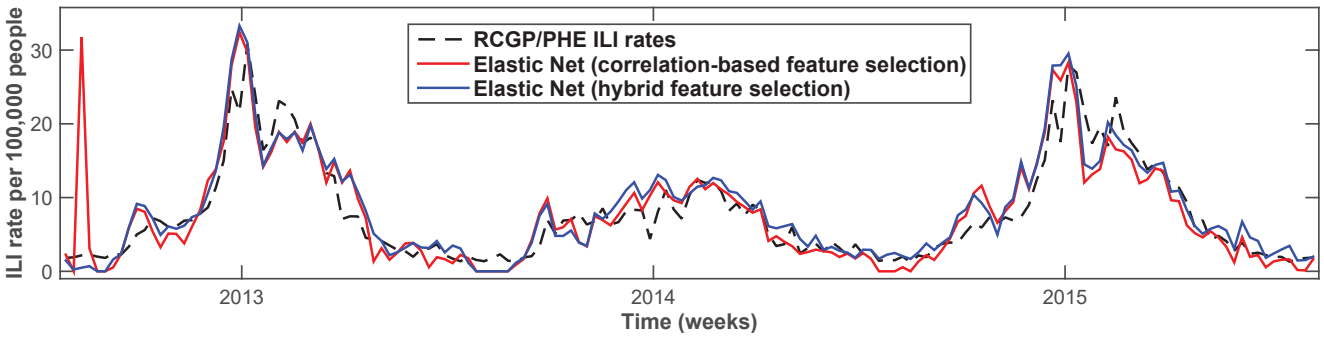


Figure 3: Comparative plot of the optimal models for the correlation based and hybrid feature selection under Elastic Net for the estimation of ILI rates in England (RCGP/PHE ILI rates denote the ground truth).

Table 4 shows a few characteristic examples of potentially misleading queries that are filtered by the hybrid feature selection approach, while previously have received a nonzero regression weight. Evidently, there exist several queries irrelevant to the target theme, referring to specific individuals and related activities, different health problems or seasonal topics. The regression weight that these queries receive tends to constitute a significant proportion of the highest weight, in the positive or the negative space. Whereas some filtered queries are indeed about flu, at the same time, they are more likely seeking for information about the disease (e.g. 'flu season') or relevant vaccination programmes, which usually take place well before the flu season emerges. Hence, from a qualitative perspective, we can deduce that the proposed feature selection is contributing towards a more semantically reliable model, where some of the spurious predictors are being omitted.

Figure 3 compares the best-performing models, under Elastic Net, for the two approaches of performing feature selection (r>.40 vs. r>.30 $\cap$ S>μ_S+σ_S). It is evident that the correlation based approach makes some odd inferences at certain points in time, whereas the hybrid one seems to accommodate more stable estimates. For example, a confusing query about a celebrity is responsible for the over-prediction on the third week of the 2012/13 flu season, with an estimated 47.52% impact on that particular inference. This query is discarded by the hybrid feature selection model as it is irrelevant to the concept of flu.

To evaluate the proposed feature selection approach with the nonlinear GP regression model, we focus on the linear regression setups (correlation based or hybrid feature selection), where the dimensionality is tractable (< 300), and a reasonable performance has been obtained. We also separately test the features that have received a nonzero weight after applying Elastic Net. The results are enumerated in Table 5 and point again to the conclusion that the hybrid feature selection yields the best performance. The best performing GP regression model (r>.30 $\cap$ S>μ_S+σ_S) amounts to the statistically significant (via a t-test) improvements —in terms of MAE— of:

1. 28.7% against the best nonlinear correlation based performance outcome ($p = .0091$), and
2. 16.6% against the best linear model ($p = .026$).

Interestingly, when the word embedding based feature selection is not applied, the nonlinear model can seldom exceed the performance of the corresponding linear model, providing an indirect indication for the inappropriateness of the selected features.

Figure 4 draws a comparison between the inferences of the best nonlinear and linear models, both of which happen to use the same feature basis (r>.30 $\cap$ S>μ_S+σ_S). The GP model provides more smooth estimates and an overall better balance between stronger and milder flu seasons. It is also more accurate in inferring the peaking moments of a flu season as the linear model repeatedly arrives to that conclusion one or more weeks before the actual occurrence (as reported in the RCGP/PHE ILI rate reports).

Table 5: Nonlinear regression (GP) performance estimates, where S>μ_S+σ_S. Check marks indicate the applied feature selection method(s). Their application sequence follows the left to right direction of the table columns.

r >	$\cap S$	Elastic Net	r_y	MAE	MAPE
.10	-	✓	.568	5.344	80.98
	✓	✓	.912	2.057	36.17
.20	-	✓	.814	4.015	63.68
	✓	✓	.920	1.892	33.08
.30	-	-	.857	2.858	54.22
	-	✓	.891	2.686	48.63
	✓	-	**.942**	**1.567**	**25.81**
	✓	✓	.928	1.696	30.30
.40	-	-	.864	2.475	45.76
	-	✓	.895	2.347	40.13
	✓	-	.913	2.110	33.65
	✓	✓	.934	2.030	33.96
.50	-	-	.887	2.197	34.17
	-	✓	.921	2.308	35.88
	✓	-	.908	2.267	35.48
	✓	✓	.926	2.292	36.55
.60	-	-	.819	2.742	43.66
	-	✓	.851	2.598	44.65
	✓	-	.865	2.614	44.36
	✓	✓	.831	2.880	52.56

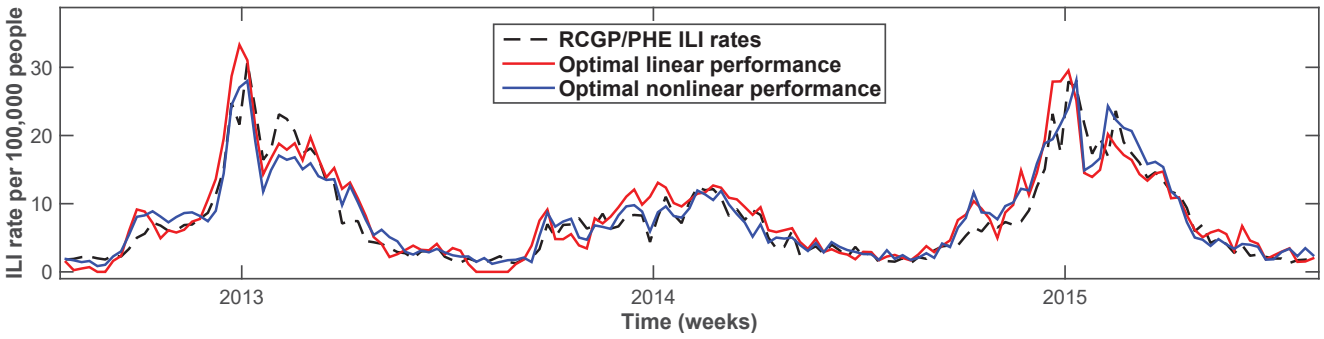


Figure 4: Comparative plot between the optimal nonlinear and linear models (both using hybrid feature selection) for the estimation of ILI rates in England (RCGP/PHE ILI rates denote the ground truth).

4.3 How are inferences affected by the choice of a different concept?

The main human intervention[13] in the proposed feature selection process is the choice of positive and negative n-grams for the formation of a concept. A reasonable question would be how the choice of these n-grams affects the feature selection and the inference performance. To provide more insight on this, we have experimented with a number of different concepts (see Table 1). $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ are variations of the flu infection topic, $\mathcal{C}_4$ and $\mathcal{C}_5$ capture the general subjects of infectious diseases and health, respectively, and $\mathcal{C}_6$ describes a different type of infection (gastrointestinal). Finally, $\mathcal{C}_7$ is a replication of $\mathcal{C}_1$ (without the Twitter hashtag *#flu*), but it is based on word embeddings trained on Wikipedia articles.

Table 6 enumerates the best obtained performance (under Elastic Net) for all investigated concepts for variants of the hybrid feature selection method ($r{>}\rho \cap S{>}\mu_S{+}\sigma_S$, $\rho{\in}[0,1)$). As we are drifting away from the flu infection topic, the performance declines (in terms of MAE or MAPE), and when the focus is drawn on a different disease (gastrointestinal; $\mathcal{C}_6$), the inference error increases significantly, providing further proof-of-concept for our approach. Yet, while remaining on the flu infection topic, we are obtaining similar (for $\mathcal{C}_2$) or slightly superior performance (for $\mathcal{C}_3$). This robustness could be justified by the average percentage of common features (∼98%) with the ones formed by using $\mathcal{C}_1$ (column '$\cap\ \mathcal{C}_1(\%)$'). Finally, the Wikipedia word embeddings produce more formal features (as it has been already indicated by Table 1), which end up providing inferior performance to the ones trained on Twitter.

[13]It could be automated by using a knowledge base.

Table 6: Optimal performance estimates after applying the hybrid feature selection method ($S{>}\mu_S{+}\sigma_S$) for varying concepts ($\mathcal{C}_1$ to $\mathcal{C}_7$) under Elastic Net. The concepts are defined in Table 1.

ID	$S \cap r >$	$\cap\ \mathcal{C}_1(\%)$	r_y	MAE	MAPE
$\mathcal{C}_1$	.30	100%	.913	1.880	36.23
$\mathcal{C}_2$	.30	98.6%	.914	1.864	37.09
$\mathcal{C}_3$	.30	98.4%	.913	**1.788**	**31.20**
$\mathcal{C}_4$	.30	87.5%	**.920**	2.084	41.51
$\mathcal{C}_5$	.30	43.1%	.891	2.237	44.19
$\mathcal{C}_6$	.20	8.3%	.616	5.217	96.45
$\mathcal{C}_7$	.30	94.2%	.909	2.116	41.88

5. RELATED WORK

Regularisation for feature selection has been routinely applied in supervised learning NLP tasks [36, 46, 60]. Word embeddings have also facilitated a number of text regression approaches, such as extending a financial lexicon for modelling risk [58], improving the inference of movie revenues based on textual reviews [5], or establishing a better feature extraction for the modelling of infectious intestinal diseases from social media content [62]. Notably, during initial experimentation we determined that dimensionality reduction, performed by using the search query embeddings directly as features in a regression model,[14] significantly reduced the inference performance. A similar in nature result has been reported in [35], when instead of raw search queries, search query n-grams have been deployed.

GP models for text regression have provided solutions in NLP applications [6, 32, 52]. For flu surveillance from search queries, more advanced regression models that accounted for potential internal structure (e.g. sub-clusters of search queries) or embedded autoregressive components have been proposed [35, 59]. Here, we use a straightforward GP kernel that is more suitable for directly assessing the predictive capacity of the selected features.

Finally, many works have focused on disease text disambiguation by training various forms of classifiers [14, 16, 17], or developing laborious, task dependent NLP schemes [2, 29]. In contrast, we have described an unsupervised, potentially task-independent, approach for quantifying, and therefore assessing, the semantic relationship between textual predictors and a target concept.

6. CONCLUSIONS

We have presented a hybrid feature selection method for digital syndromic surveillance that employs neural word embeddings to improve the topicality of the selected features. Our approach can be seen as an unsupervised filter for a

[14]The high-dimensional space of search queries is being reduced (compressed) to the dimensionality of the embedding.

target thematic concept that can be easily applied in conjunction with current feature selection techniques. Using social media content to learn word embeddings, our regression experiments were conducted on an 8-year-long data set of search queries, with the aim to infer flu rates in a population. We have shown that the proposed hybrid feature selection method generates a more reliable regression model that can significantly outperform competitive approaches (by 12% or more). Future work will focus on further generalisations of the reported outcomes by exploring different infectious diseases, focusing on different locations, or even expanding to other application domains.

7. ACKNOWLEDGEMENTS

This work has been supported by the grant EP/K031953/1 (EPSRC) and a Google Research Sponsorship. The authors would also like to acknowledge PHE and RCGP (in particular Richard Pebody and Simon de Lusignan, respectively) for providing syndromic surveillance data, and Google for providing access to the Google Health Trends API. Finally, we thank Jens K. Geyti for leading the development of Flu Detector (`fludetector.cs.ucl.ac.uk`).

8. REFERENCES

[1] S. Amir, R. Astudillo, W. Ling, et al. INESC-ID: A Regression Model for Large Scale Twitter Sentiment Lexicon Induction. In *Proc. of SemEval '15*, pages 613–618, 2015.

[2] E. Aramaki, S. Maskawa, and M. Morita. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In *Proc. of EMNLP '11*, pages 1568–1576, 2011.

[3] P. Barberá. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Polit. Anal.*, 23(1):76–91, 2014.

[4] D. Beck, T. Cohn, C. Hardmeier, and L. Specia. Learning Structural Kernels for Natural Language Processing. *Trans. Assoc. Comput. Linguist.*, 3:461–473, 2015.

[5] Z. Bitvai and T. Cohn. Non-Linear Text Regression with a Deep Convolutional Neural Network. In *Proc. of ACL '15*, pages 180–185, 2015.

[6] Z. Bitvai and T. Cohn. Predicting Peer-to-Peer Loan Rates Using Bayesian Non-Linear Regression. In *Proc. of AAAI '15*, pages 2203–2209, 2015.

[7] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *J. Comput. Sci.*, 2(1):1–8, 2011.

[8] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. Digital Disease Detection — Harnessing the Web for Public Health Surveillance. *N. Engl. J. Med.*, 360(21):2153–2157, 2009.

[9] A. D. Bull. Convergence Rates of Efficient Global Optimization Algorithms. *J. Mach. Learn. Res.*, 12:2879–2904, 2011.

[10] H. Choi and H. Varian. Predicting the Present with Google Trends. *Economic Record*, 88:2–9, 2012.

[11] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting Depression via Social Media. In *Proc. of ICWSM '13*, pages 128–137, 2013.

[12] T. Cohn, D. Preoţiuc-Pietro, and N. Lawrence. Gaussian Processes for Natural Language Processing. In *Proc. of ACL '14: Tutorials*, pages 1–3, 2014.

[13] T. Cohn and L. Specia. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proc. of ACL '13*, pages 32–42, 2013.

[14] N. Collier, N. T. Son, and N. M. Nguyen. OMG U got flu? Analysis of shared health messages for bio-surveillance. *J. Biomed. Semant.*, 2(5):1–10, 2011.

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

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

[17] S. Doan, L. Ohno-Machado, and N. Collier. Enhancing Twitter Data Analysis with Simple Semantic Filtering: Example in Tracking Influenza-Like Illnesses. In *Proc. of the 2nd International IEEE Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 62–71, 2012.

[18] D. Duvenaud. *Automatic Model Construction with Gaussian Processes*. Ph.D. Thesis, University of Cambridge, 2014.

[19] K. Ganchev, K. Hall, R. McDonald, and S. Petrov. Using search-logs to improve query tagging. In *Proc. of ACL '12*, pages 238–242, 2012.

[20] J. Ginsberg, M. H. Mohebbi, R. S. Patel, et al. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[21] Y. Goldberg. A Primer on Neural Network Models for Natural Language Processing. *arXiv Preprint arXiv:1510.00726*, 2015.

[22] I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

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

[24] W. R. Hobbs, M. Burke, N. A. Christakis, and J. H. Fowler. Online social integration is associated with reduced mortality risk. *Proc. Natl. Acad. Sci. U.S.A.*, 113(46):12980–12984, 2016.

[25] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[26] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *Proc. of ACL '15*, pages 1681–1691, 2015.

[27] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie Reviews and Revenues: An Experiment in Text Regression. In *Proc. of NAACL '10*, pages 293–296, 2010.

[28] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. U.S.A.*, 111(24):8788–8790, 2014.

[29] A. Lamb, M. J. Paul, and M. Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Proc. of NAACL '13*, pages 789–795, 2013.

[30] V. Lampos. Flu Detector: Estimating influenza-like illness rates from online user-generated content. *arXiv Preprint arXiv:1612.03494*, 2016.

[31] 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 '16*, pages 689–695, 2016.

[32] V. Lampos, N. Aletras, D. Preoţiuc-Pietro, and T. Cohn. Predicting and Characterising User Impact on Twitter. In *Proc. of EACL '14*, pages 405–413, 2014.

[33] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the Social Web. In *Proc. of the 2nd International Workshop on Cognitive Information Processing*, pages 411–416, 2010.

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

[35] 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.

[36] V. Lampos, D. Preoţiuc-Pietro, and T. Cohn. A user-centric model of voting intention from Social Media. In *Proc. of ACL '13*, pages 993–1003, 2013.

[37] 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.

[38] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205, 2014.

[39] O. Levy and Y. Goldberg. Dependency-Based Word Embeddings. In *Proc. of ACL '14*, pages 302–308, 2014.

[40] O. Levy and Y. Goldberg. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proc. of CoNLL '14*, pages 171–180, 2014.

[41] O. Levy, Y. Goldberg, and I. Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Trans. Assoc. Comput. Linguist.*, 3:211–225, 2015.

[42] B. Matérn. *Spatial Variation*. Springer, 1986.

[43] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. In *Proc. of the ICLR, Workshop Track*, pages 1–12, 2013.

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

[45] D. R. Olson, K. J. Konty, M. Paladini, et al. Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales. *PLoS Comput. Biol.*, 9(10), 10 2013.

[46] O. Owoputi, B. O'Connor, C. Dyer, et al. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proc. of NAACL '13*, pages 380–390, 2013.

[47] M. Paşca and B. Van Durme. Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. In *Proc. of ACL '08*, pages 19–27, 2008.

[48] G. Park, H. A. Schwartz, J. C. Eichstaedt, et al. Automatic personality assessment through social media language. *J. Pers. Soc. Psychol.*, 108(6):934–952, 2015.

[49] M. J. Paul and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proc. of ICWSM '11*, pages 265–272, 2011.

[50] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using Internet Searches for Influenza Surveillance. *Clin. Infect. Dis.*, 47(11):1443–1448, 2008.

[51] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras. An analysis of the user occupational class through Twitter content. In *Proc. of ACL '15*, pages 1754–1764, 2015.

[52] 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.

[53] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying Latent User Attributes in Twitter. In *Proc. of the 2nd International Workshop on Search and Mining User-generated Contents*, pages 37–44, 2010.

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

[55] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, et al. Personality, gender, and age in the language of social media: The Open-Vocabulary approach. *PLoS ONE*, 8(9):e73791, 2013.

[56] A. Signorini, A. M. Segre, and P. M. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, 6(5):e19467, 2011.

[57] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58(1):267–288, 1996.

[58] M.-F. Tsai and C.-J. Wang. Financial Keyword Expansion via Continuous Word Vector Representations. In *Proc. of EMNLP '14*, pages 1453–1458, 2014.

[59] S. Yang, M. Santillana, and S. C. Kou. Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc. Natl. Acad. Sci. U.S.A.*, 112(47):14473–14478, 2015.

[60] T. Yano, N. A. Smith, and J. D. Wilkerson. Textual Predictors of Bill Survival in Congressional Committees. In *Proc. of NAACL '12*, pages 793–802, 2012.

[61] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[62] B. Zou, V. Lampos, R. Gorton, and I. J. Cox. On Infectious Intestinal Disease Surveillance Using Social Media Content. In *Proc. of the 6th International Conference on Digital Health*, pages 157–161, 2016.

[63] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, 67(2):301–320, 2005.