Disease Surveillance using User-generated Content

Bin Zou

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Computer Science

University College London

February 3, 2019

I, Bin Zou, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Disease surveillance plays a crucial role in detecting or anticipating infectious disease outbreaks. It tracks health-related data from a population to identify and monitor early outbreaks of a disease. Traditional disease surveillance requires a widespread network of sentinel sites to track infections throughout the population. These networks are time and labour intensive to build and maintain, and this creates opportunities for utilizing online user-generated content. Compared to traditional data sources, online user-generated content is fast and cheap to obtain. It covers a larger population, and provides data on topics with little coverage from traditional sources. This can complement traditional disease surveillance systems. In this thesis, we focus on improving disease surveillance using online user-generated content, through machine learning and natural language processing techniques.

Our contributions are threefold. First, a feature selection method, which consists of a time series similarity filter and a topic filter, is proposed. The former filter ensures the selected features are good predictors, while the topic filter succeeds in eliminating features that may be highly correlated with disease rates, but are not referring to the target disease. Second, a multi-task learning framework for disease surveillance is proposed, where several disease surveillance models are jointly trained. Multi-task elastic net and multi-task Gaussian Processes are used for regression. The framework improves the generalization of a model by taking advantage of shared structures in the data. Third, a transfer learning framework is proposed for delivering accurate disease rate models without the existence of ground truth information for a target location. The framework consists of three steps: (1) learn a regularized regression model for a source country, (2) map the source queries

to target ones using semantic and temporal similarity metrics, and (3) re-adjust the weights of the target queries.

To support the theoretical derivations, extensive and repeatable experiments are carried out based on large-scale real-world data. Experimental results have demonstrated substantial improvement of the proposed solutions over strong baselines. In addition, we publish a website that reports real-time flu rate estimation in England (https://fludetector.cs.ucl.ac.uk/).

Impact Statement

In this thesis, we focus on improving disease surveillance using online user-generated content. Within academia, we made three contributions. First, we proposed a joint feature selection method, which consists of a time series similarity filter and a semantic filter. The former filter is based on Pearson correlation, and ensures the features remained are potentially good predictors. The later semantic filter is based on word embeddings, and succeeds in eliminating confounding features, i.e. queries that may be highly correlated with disease rates, but are not referring to the target disease. Second, we investigated the utility of multi-task learning techniques to disease surveillance from Web search data. Several related disease surveillance models from different geographies are jointly trained using linear (multi-task elastic net) and nonlinear (multi-task Gaussian Processes) models. The data structures are shared during joint training, which exploits the relatedness between tasks and improves the generalization of the model. Third, we proposed a transfer learning framework for delivering considerably accurate disease rate models without the existence of ground truth information for a target location. Our framework consists of three steps: (1) learn a regularized regression model for a source country, (2) map the source queries to target ones using semantic and temporal similarity metrics, and (3) re-adjust the weights of the target queries.

Outside of academia, this work improves the entire public health cycle. We estimate disease rates in real-time, and this complements traditional disease surveillance in monitoring the health of a population, i.e. the assessment in a public health cycle. In addition, our estimations are 2 to 3 weeks ahead of the official numbers published by established health agencies. This can provide an early warning before

epidemics happen and affect the public health policy development. Our work can also be used to evaluate the effectiveness of the policies.

Web-based disease surveillance systems can be used for both developed countries where well-established health systems exist and ground truth is sufficient, and low and middle income countries where such well-established health infrastructure is missing and ground truth partially and does not exist. For developed countries, although well-established health systems exist, there is usually several weeks delay on reporting. Web-based disease surveillance can complement traditional disease surveillance systems by providing accurate estimates of disease rates in (nearly) real-time. For low and middle income countries, where ground truth only partially or does not exist, multi-task and transfer-learning techniques can be used to provide accurate disease rate estimation.

Acknowledgements

First, I would like to express my sincere gratitude to my primary supervisor Prof. Ingemar J. Cox. I got to know Ingemar in October 2014. After a search on his research, I decided to apply his PhD at UCL. Fortunately, I got the offer and the full scholarships from Ingemar and started my PhD from February 2015.

During these 4 years of PhD with Ingemar, I have been feeling lucky that I have made the right choice to purse a PhD with him. In academic research, Ingemar is a careful and strict professor. He teaches me how to read papers, generate research idea, make experiments to validate hypothesis, and write papers logically. He also encourages me to attend summer schools, conferences and workshops to present my research to others. In life, Ingemar is a very nice and easy-going professor. He is humorous, and always encourages me when I feel frustrated. Without his generous support, I do not believe I could survive my PhD.

Second, I want to show my great appreciation to my secondary supervisor Prof. Sebastian Riedel. Sebastian discusses the research ideas with me and proposes many constructive suggestions to my PhD during my first year and transfer viva.

Third, I would like to thank my PhD viva examiners Prof. Nigel Collier (University of Cambridge) and Prof. Steve Hailes (University College London) for taking time to review my thesis and offer constructive comments and suggestions.

Fourth, I would like to thank my UCL research colleagues Dr. Vasileios Lampos, Prof. Emine Yilmaz, Prof. Shi Zhou, Prof. Jun Wang, Peter Hayes, Dr. Nikolaos Aletras, Jens K. Geyti, Dr. Zhaochun Ren, Dr. Shangsong Liang, David Guzman, Dr. Vincent Primault, Dr. Simon Moura, Dr. Jie Xiong, Dr. Weinan Zhang, Dr. Rui Yu, Dr. Ye Pan, Dr. Changwang Zhang, Rui Luo, Jie Li, Yixin Wu, Bowen

Zheng, Ying Wen, Zhen Tian and Qiang Zhang for their help and many insightful discussions.

Fifth, I would like to thank my research colleagues for providing the research data. They are Dr. Elad Yom-Tov from Microsoft Research, Dr. Richard Pebody and Dr. Russell Gorton from Public Health England, and Google for providing access to Google Health Trends API.

Moreover, I will not forget the help from my home university and industrial supervisors before my PhD study. They are Prof. Mads Nilsen, Dr. Lauge Sørensen, Dr. Sune Darkner, Prof. Christian Igel, Dr. Kang Li and Dr. Pengfei Diao from University of Copenhagen, Prof. Chengchen Hu, Prof. Lihua Tian and Prof. Ruifang Ma from Xi'an Jiaotong University, Dr. Erik Dam and Dr. Akshay Pai from Biomediq A/S.

Finally, I am very grateful to my wife Chu, and parents for their love and unconditional support during these 4 years. Without them, I cannot imagine how I could survive my PhD.

I dedicate this thesis to

my wife Chu

for her love and unconditional support.

I love you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Infectious diseases pose significant risks. Historically, the "Black Death" bubonic plague of the 14th century is estimated to have killed 75 to 100 million people, which was 30 to 60% of Europe's total population (Gottfried, 2010; Ziegler, 2013); the Spanish flu of 1918-1920 is estimated to have killed 50 to 100 million people, which is 3 to 5% of the world's population (Johnson and Mueller, 2002). The last century has witnessed great achievements in medical and pharmaceutical science, but infectious diseases remain as serious threats. An estimate of 31 to 35 million people in the world are HIV-affected (World Health Organization, 2006). Infectious diseases can spread rapidly, and threaten people worldwide. For example, the Ebola virus epidemic in 2014 has $28,502$ reported cases resulting in $11,312$ deaths in several months (World Health Organization, 2014). Various types of methods have been developed to reduce the risk of infectious disease outbreaks, including the development of new drugs, improvement of therapies, promotion of personal behavior, introduction of vaccination programs, hospital infection control, and disease surveillance (Wagner et al., 2011).

Disease surveillance plays a crucial role in detecting or anticipating disease outbreaks. It is the continuous, systematic collection, analysis, and interpretation of large volumes of health-related data.[1] Broadly speaking, disease surveillance has five goals: 1) evaluate the effectiveness of control and preventative health measures, 2) monitor changes in infectious agents, e.g. trends in development of antimicrobial

[1] http://www.who.int/topics/public_health_surveillance/

resistance, 3) support health planning and allocation of appropriate resources within the healthcare system, 4) identify high risk populations or areas to target interventions, and 5) provide a valuable archive of disease activity for future reference.[2] A key part of modern disease surveillance is the practice of disease case reporting. In order to make disease surveillance effective, the collection of surveillance data must be standardized on a national basis and be made available at local, regional, and national level.

Syndromic surveillance is a type of disease surveillance. It refers to the surveillance of a specific syndrome (a set of related symptoms). Syndromic surveillance uses case definitions that are based entirely on clinical features without any clinical or laboratory diagnosis (for example, collecting the number of cases of diarrhea rather than cases of cholera, or "rash illness" rather than measles). Without laboratory confirmation, syndromic surveillance is inexpensive and faster (Jamison et al., 2006). However, because of the lack of specificity, syndromic surveillance reports require more investigation from higher levels. For example, a "rash illness" could be anything from the relatively minor rubella to devastating hemorrhagic fevers. Also an increase in one disease causing a syndrome may mask an epidemic of another (for example, rotavirus diarrhea decreases at the same time cholera increases).

The rest of this chapter is structured as follows. In Section 1.1, we review different sources of data for disease surveillance and discuss their limitations. Then, we introduce Web-based disease surveillance in Section 1.2, which utilizes online user-generated content. In Section 1.3, we describe our research problems and contributions. Finally, we present the structure of the thesis in Section 1.4.

1.1 Sources of Data for Disease Surveillance

Traditionally, there are two main sources of data for disease surveillance: surveys and clinical records.

[2] http://www.hpsc.ie/abouthpsc/whatisdiseasesurveillance/

1.1.1 Surveys

Surveys have long been used for health study. There are several, large-scale surveys run on a regular basis to provide health-related data. Some surveys are based on telephone interviews. For example, Behavioral Risk Factor Surveillance Systems (BRFSS) collects data about United States (US) residents regarding their health-related risk behavior, chronic health conditions, and use of preventive services through health-related telephone surveys.[3] It runs annually and collects detailed data from more than $400,000$ people.

There are also surveys that rely on in-person interviews, such as the annual US National Survey on Drug Use and Health (NSDUH).[4] NSDUH is conducted every year to provide information on tobacco, alcohol, and drug use, mental health and other health-related issues in the US. The information is used to support prevention and treatment programs, monitor substance use trends, estimate the need for treatment, and inform public health policy.

Apart from telephone and in-person surveys, Web-based surveys are becoming popular due to their low cost (Cook et al., 2000; Eysenbach and Wyatt, 2002). An example is Flusurvey in the United Kingdom (UK).[5] Flusurvey is an online survey that monitors trends of influenza-like illness (ILI) in the community. Any member of the UK public can register the platform to report flu like symptoms they may experience during the winter months.

1.1.2 Clinical Records

The second traditional data source for disease surveillance is clinical records. An example is Influenza-Like Illness Surveillance Network (ILINet)[6] of Centers for Disease Control and Prevention (CDC)[7]. ILINet consists of more than $2,800$ enrolled outpatient healthcare providers in the US reporting more than 39 million patient visits each year. Each week, approximately $2,000$ outpatient healthcare providers around the country report data to CDC on the total number of patients

[3] **BRFSS**, https://www.cdc.gov/brfss/
[4] **NSDUH**, https://nsduhweb.rti.org/respweb/
[5] **Flusurvey**, https://flusurvey.net/
[6] **ILINet**, https://www.health.ny.gov/diseases/communicable/influenza/surveillance/ilinet_program/
[7] **CDC**, https://www.cdc.gov/

seen for any reason and the number of those patients with ILI, i.e. ILI rates. Similarly, the European Centre for Disease Prevention and Control (ECDC)[8] has developed European Influenza Surveillance Network (EISN) and publish weekly ILI rates through Flu News Europe.[9] These large-scale surveillance networks require significant coordination as they rely on active reporting from clinics.

1.1.3 Other Data Sources

While surveys and clinical records are the most common data sources, researchers are actively seeking new data sources for disease surveillance. These include monitoring sales of over-the-counter drug sales and pharmacy records (Heffernan et al., 2004; Magruder et al., 2004) to track gastrointestinal illness (Edge et al., 2004), utilizing absenteeism records of public or private schools to monitor influenza activity (Mook et al., 2007; Cheng et al., 2013), monitoring call records of emergency call centers (Yih et al., 2009; Hiller et al., 2013) for detection of ILI, and using insurance company billing records to track cardiovascular diseases (Lentine et al., 2009).

1.1.4 Limitations of Traditional Data Sources

Traditional data sources have their advantages. In general, the data is analyzed with biased corrected. Furthermore, many of these data sources date back many years, enabling us to make comparisons over time. However, traditional data sources have their limitations.

The phone surveys become less accurate over time, as more people do not use landline phones. This introduces the bias against low-income young people in survey results (Blumberg and Luke, 2007). In-person interviews are hard to conduct, especially when the survey size is large (Iannacchione, 2011).

Clinical records address some of these issues, but they are expensive to obtain, as large health monitoring systems need to be established (Wagner et al., 2011; Paul and Dredze, 2017). In addition, clinical records are in unstructured text, making them complex to analyze. Furthermore, clinical records can only cover certain top-

[8] **ECDC**, https://ecdc.europa.eu/
[9] **Flu News Europe**, https://flunewseurope.org/

ics, and many areas of health are hard to study because they lack sufficient data. For example, mental health disorders are still understudied using traditional data sources (Ofran et al., 2012; Yom-Tov et al., 2014).

1.2 Web-based Disease Surveillance

The limitations of traditional data sources creates opportunities for Web-based disease surveillance. Web-based disease surveillance utilizes online user-generated content as a data source. Online user generated content is electronic data created by users of an online system or service. They can be any form of content such as search queries log, blogs, wikis, discussion forums, posts, chats, tweets and other forms of media (Moens et al., 2014). A survey of Pew Research Centre found that more than 70% of US Internet users consult the Internet when they require medical information (Fox and Duggan, 2013). The user-generated content created by people who seek information on the Web, offers an unprecedented opportunity for building a new class of syndromic surveillance systems. Compared to other data sources, online user-generated content is fast, cheap, covers a larger population, and provides data on topics with little coverage from traditional sources. Traditional disease surveillance systems can be complemented with disease surveillance systems that utilize user-generated content. For the past decade, online user-generated content has been used in a variety of ways (Ginsberg et al., 2009; Lampos and Cristianini, 2010; Gomide et al., 2011; Paul et al., 2014; Fung et al., 2014; Eschler et al., 2015; Zou et al., 2016; McGough et al., 2017).

Wagner et al. (2011) defined three aspects to judge whether a data source is suitable for disease surveillance systems (1) information value, (2) availability, and (3) cost. Online user-generated content fulfills these three criteria. First, a number of works have shown that online user-generated content contains information of offline behavior (Heaivilin et al., 2011; Cobb et al., 2011; Paul and Dredze, 2013). Second, user-generated content is online and much of them can be accessed through an Application Programming Interface (API). Third, obtaining user-generated content is timely and inexpensive.

Despite the benefits of utilizing user-generated content for health research, concerns have been raised about the quality of online health information (Cline and Haynes, 2001). Firstly, user-generated content can be noisy and ambiguous. For example, tweet "I had Bieber fever" does not mean a person had a health problem. In addition, truthfulness of user-generated content has to be examined. Pelleg et al. (2012) explored personal topics such as body measurements, income and sexual behavior on Yahoo Answers, and found that Web users exhibit a low level of truthfulness on some topics, especially when the topics are personal and sensitive.

1.3 Research Problems and Contributions

In this thesis, we focus on improving disease surveillance systems using online user-generated content, through machine learning (ML) and natural language processing (NLP) techniques. The aim of disease surveillance is to infer disease rates as reported by established health agencies (e.g. CDC and ECDC). In this thesis, we make three contributions to disease surveillance.

1.3.1 Enhancing Feature Selection using Word Embeddings

Disease surveillance systems based on user-generated content often rely on the identification of textual markers (e.g. queries or a set of terms) 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 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 Chapter 3, we take advantage of current developments in statistical NLP and propose a feature selection method based on neural word embeddings and correlation. Word embeddings is used in an unsupervised manner to determine 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. We evaluate our method on two large-scale, practical, text regression tasks, specifically the estimation of

ILI rates and infectious intestinal disease rates from search query frequencies. Our empirical analysis shows that the proposed feature selection method provides significant performance gains under both linear and nonlinear regression models. Qualitative insights indicate that this is due to the topicality of the maintained textual features.

1.3.2 Multi-Task Learning for Disease Surveillance

Disease surveillance models using user-generated content are predominantly based on single-task learning methods (Polgreen et al., 2008; Ginsberg et al., 2009; Culotta, 2010; Paul et al., 2014; Lampos et al., 2015; Yang et al., 2015). These models do not consider the relations of data and model across different geographies. They also do not consider the situation where ground truth (disease rates) is insufficient for training a model.

In Chapter 4, we focus on these two problems and investigate the utility of multi-task learning to disease surveillance using Web search data. Our motivation is twofold. First, we assess whether concurrently training models for various geographies can improve accuracy. Second, we test the ability of such models to assist health systems that are producing sporadic disease surveillance reports that reduce the quantity of available training data. We explore both linear and nonlinear regression models, namely multi-task elastic net (Lee et al., 2010) and multi-task Gaussian Processes (Bonilla et al., 2007), comparing them to their respective single task formulations. A case study on ILI rates estimation show that multi-task learning can improve regional and national models. Furthermore, in simulated scenarios, where only limited training data is available, we show that multi-task learning maintains a stable performance across all the affected locations.

1.3.3 Transfer Learning for Disease Surveillance

A considerable body of research has demonstrated that online search data can be used to complement current syndromic surveillance systems. The vast majority of previous work proposes solutions that are based on supervised learning paradigms, in which historical disease rates are required for training a model. However, for

many geographical regions this information is either sparse or not available due to poor health infrastructure. It is these regions that have the most to benefit from inferring population health statistics from online user search activity.

In Chapter 5, we address this issue and propose a statistical framework in which we first learn a supervised model for a region with adequate historical disease rates, and then transfer it to a target region, where no syndromic surveillance data exists. This transfer learning solution consists of three steps: (1) learn a regularized regression model for a source country, (2) map the source queries to target ones using semantic and temporal similarity metrics, and (3) re-adjust the weights of the target queries. Our solution is evaluated on the task of estimating ILI rates. We learn a source model for the US, and subsequently transfer it to three other countries, namely France, Spain and Australia. We use the existing ILI rates in the target countries only to evaluate our estimates. Overall, the transferred (unsupervised) models achieve strong performance in terms of Pearson correlation with the ground truth, and their mean absolute error does not deviate greatly from a fully supervised baseline.

1.3.4 Summarized Contributions

The scientific contributions of this thesis are threefold.

First, a feature selection method, which consists of a time series similarity filter and a semantic similarity filter, is proposed for disease surveillance. The former filter ensures the selected features are potentially good predictors, while the semantic filter succeeds in eliminating some confounding features, i.e. queries that may be highly correlated with disease rates, but are not referring to the target disease. The method is also applicable to more general text regression tasks.

Second, a multi-task learning framework is proposed for disease surveillance, where a number of disease surveillances models are jointly trained. In particular, multi-task elastic net and multi-task Gaussian Processes are applied to disease surveillance. The framework improves the generalization of a model by taking advantage of shared structures in the data.

Third, a transfer learning framework is proposed for delivering accurate dis-

ease rate models without the existence of ground truth information for a target location. After learning a regularized regression model for a source country, the framework maps the source queries to target ones using semantic and temporal similarity metrics, and finally re-adjusts the weights of the target queries.

Besides these significant scientific innovations, extensive and repeatable experiments on large-scale real-world data have been performed to verify the effectiveness of each proposed solution. The proposed solutions are included in an official report of Public Health England. More importantly, the proposed solutions are generalized solutions and applicable to other disease surveillance models.

In summary, the scientific and empirical contributions of this research are significant in terms of building disease surveillance systems using online user-generated content.

1.4 Thesis Structure

The rest of this thesis is organized as follows. In Chapter 2, we perform a literature review of related techniques and their connections to our research. In Chapter 3, we propose a feature selection method that considers not only the time series similarity, but also topicality of features. In Chapter 4, we propose a multi-task learning framework for disease surveillance. In Chapter 5, we propose a transfer learning framework for disease surveillance. Finally, we conclude this thesis and discuss future research work in Chapter 6.

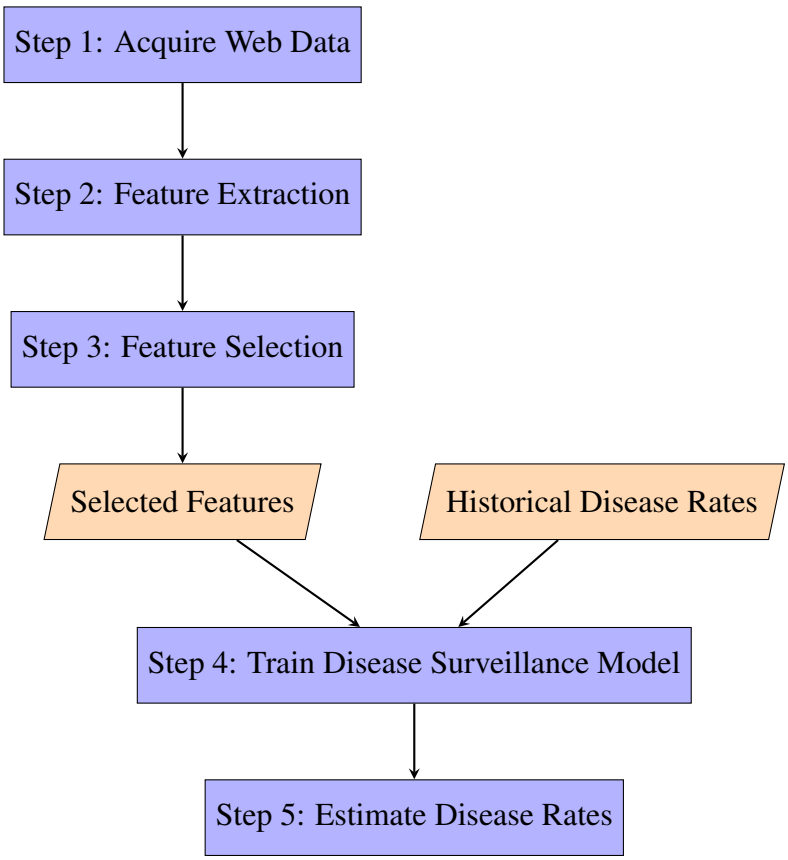
Chapter 2

Related Work

In this thesis, we focus on improving real-time disease surveillance systems using user-generated content from the Web, through machine learning and natural language processing techniques. The aim of disease surveillance is to infer disease rates as reported by established health surveillance systems. A common framework to tackle this challenge is presented in Figure 2.1. The framework consists of five steps: (1) acquire data from the Web, (2) extract features from the Web data, (3) select features from extracted features, (4) train supervised learning models, and (5) estimate disease rates and provide early warning before disease outbreak happens.

In the first step, we retrieve user-generated content from the Web. We discuss different kinds of user-generated content in Section 2.1. In the second step, we identify a set of textual markers that are related to a target disease, such as search queries or a contiguous sequence of n items from a given text (known as n-grams), and then extract frequencies of these textual markers from the user-generated content. Feature extraction techniques are discussed in Section 2.2. Given the high volume of available data, the number of extracted features is usually large. Therefore, in the third step, we perform feature selection to reduce the number of features. Section 2.3 review and discuss different feature selection methods. In the fourth step, we train supervised learning (regression) models using selected features and disease rates obtained from established health agencies. Linear and nonlinear regression models are discussed in Section 2.4. Finally, we make the inference of disease rates in the fifth step. Different evaluation metrics are reviewed in Section 2.5.

Figure 2.1: Web-based disease surveillance framework.



2.1 Web Data

In the last decade, there is an increasing trend in utilizing online user-generated content to study health issues. User-generated content is electronic data created by users of an online system or service. They can be any form of content such as search queries log, blogs, wikis, discussion forums, posts, chats, tweets and other forms of media (Moens et al., 2014). Different kinds of health issues have been studied using user-generated content, such as influenza surveillance (Polgreen et al., 2008; Ginsberg et al., 2009; Lampos and Cristianini, 2010; Culotta, 2010; Paul et al., 2014; Yang et al., 2015; Lampos et al., 2017; Zou et al., 2018), dengue (Gomide et al., 2011; Gluskin et al., 2014; Li et al., 2017), ebola virus diseases (Fung et al., 2014; Odlum and Yoon, 2015), Zika virus (Juric et al., 2017; Miller et al., 2017; McGough et al., 2017), cancer (Ofran et al., 2012; Eschler et al., 2015; Paul et al., 2016), diet and fitness (Abbar et al., 2015; Garimella et al., 2016; Zou et al., 2016). In gen-

eral, user-generated content is created by ordinary people, rather than professional writers or domain experts such as medical doctors.

User-generated content can be collected in two ways: explicitly and implicitly (Krumm et al., 2008). In the process of explicit data gathering, users interact with client controls and understand that they are inputting data (e.g. rating a video, uploading a picture on Facebook, and tweeting moods), while, in the process of implicit data gathering, users perform events that are tracked but they may not understand their actions are being monitored (e.g. using search engines, watching a video, and clicking a link). User-generated content comes in many forms. Different online platforms and websites exist for different audiences and different purposes, and different platforms may be better suited for particular health goals. In this section, we review some commonly used user-generated content. We first present the statistics on commonly used data sources, then we discuss their use for public health in details separately.

2.1.1 Statistics on Commonly used Web Data

In Table 2.1, we present the statistics of commonly used Web data sources. The statistics are obtained from surveys conducted by the Pew Research Center. The survey of search engine use was conducted by Purcell et al. (2012); the survey of Wikipedia use was conducted by Zickuhr and Rainie (2012); the survey of Twitter, Facebook, LinkedIn, Instagram, and Pinterest was conducted by (Greenwood et al., 2016). The statistics include

- Online adults use, which defines the percentage of online adults who use the data source.

- Gender, which includes men and women.

- Age, which includes 4 age groups, $18-29$, $30-49$, $50-64$, and 65+.

- Education background, which includes 3 categories, high school degrees or less, college, and college+.

- Income level, which includes less than \$30K/year, \$30K-\$49,999, \$50K-\$74,999, and \$75,000+.
- Developed environments, which includes urban, suburban and rural.
- Frequencies of using the data, which includes daily, weekly, and less often.
- Data format, including text or images.
- Accessibility, which includes whether data is public or private, data collection tools, and constraint on accessing the data.
- Ambiguity, which measures whether there is ambiguity in the data.

2.1.2 Search Data

Search queries are queries users enter into Web search engines to seek the information they need. A query in a search engine suggests an interest in a topic, and thus by analyzing what people are searching for, we can infer what people are interested in. Search data covers 91% of Internet users. Search is most popular among young adult internet users, those who have been to college, and those with the highest household incomes. But it is not biased much in terms of gender, age, education background or incomes. Search engines are widely used for seeking health information online. According to a survey from the Pew Research, 72% adult internet users in the US say they have searched online for information about a range of health issues, the most popular being specific diseases and treatments.

Search engines, such as Google[1], Bing[2], Baidu[3], and Yahoo[4], log the queries that are searched by users. Raw logs are private data, but some search engines make aggregate statistics about query volumes publicly available through services such as Google Trends. Figure 2.2 plots the weekly interest over time worldwide for query "Ebola". We can see that before August 2014, there is little interest in "Ebola" on Google. From September 2014 to December 2014, the search volume of query

[1] **Google**, https://www.google.com/
[2] **Bing**, https://www.bing.com/
[3] **Baidu**, http://www.baidu.com/
[4] **Yahoo Search**, https://www.yahoo.com/

Table 2.1: Statistics of commonly used Web data sources in US, obtained from surveys conducted by the Pew Research Center. Numbers in the table are in percentages. The statistics include percentage of online adults who use the data, gender, education background, income level, developed environments, frequency of using the data, format, accessibility, and ambiguity of the data. "n/a" means not available.

	Search Engines	**Twitter**	**Wikipedia**	**Facebook**	**LinkedIn**	**Instagram**	**Pinterest**
Online adults use	91	24	53	79	29	32	31
Men	90	24	56	75	31	26	17
Women	92	25	50	83	27	38	45
$18-29$	96	36	62	88	34	59	36
$30-49$	91	23	52	84	33	33	24
$50-64$	92	21	49	72	24	18	28
65+	80	10	33	62	20	8	16
High school degree or less	88	20	41	77	12	27	24
Some college	94	25	52	82	27	37	34
College+	95	29	69	79	50	33	34
<$30K/year	84	23	44	84	21	38	30
$30K-$49,999	93	18	49	80	13	32	32
$50K-$74,999	97	28	61	75	32	32	31
$75,000+	95	30	61	77	45	31	35
Urban	n/a	26	n/a	81	34	39	30
Suburban	n/a	24	n/a	77	30	28	34
Rural	n/a	24	n/a	81	18	31	25
Daily	54	42	n/a	76	18	51	25
Weekly	38	24	n/a	15	31	26	31
Less often	8	34	n/a	9	51	23	44
Data format	text	text	text	text and images	text	images	images
Access level	public	public	public	private	private	public	private
Constraint	only aggregated	number limited	no	friends only	friends only	number limited	friends only
Collection tool	Google Trends	Twitter API	Wikipedia dumps	Facebook API	Linkedin API	Instagram API	Pinterest API
Ambiguity	clear	ambiguous	clear	ambiguous	ambiguous	ambiguous	ambiguous

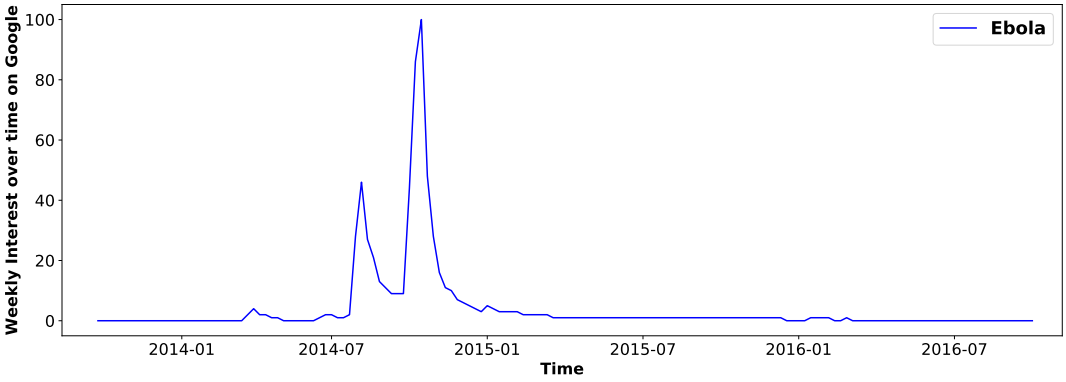


Figure 2.2: Weekly interest over time for query "Ebola" worldwide on Google from October 2013 to October 2016. The data was obtained from Google Trends on 26 Aug, 2018.

"Ebola" increases dramatically. This corresponds to the real-world event that Ebola virus diseases spread inside and outside Africa. After January 2015, the interest decreases to almost zero level, since Ebola virus diseases are under control.

Search data are one of the most popular data sources for disease surveillance. For example, Polgreen et al. (2008) utilized search data from Yahoo to predict influenza activity. Ginsberg et al. (2009) provided early detection of influenza outbreaks by monitoring search data on Google. The latter work became Google Flu Trends. However, Google Flu Trends has been criticized for poor predictive performance, underestimating or overstating the prevalence of flu (Lazer et al., 2014; Santillana et al., 2014b). Following Google Flu Trends, a significant amount of follow-up work has been conducted using search data (Xu et al., 2010; Cook et al., 2011; Dugas et al., 2012; Copeland et al., 2013; Yuan et al., 2013; Stefansen, 2014; Preis and Moat, 2014; Wang et al., 2015; Yang et al., 2015; Pollett et al., 2016; Shin et al., 2016). Apart from influenza surveillance, search data have also been used to study other diseases, such as chickenbox (Pelat et al., 2009; Valdivia and Monge-Corella, 2010), dengue (Gomide et al., 2011; Gluskin et al., 2014; Li et al., 2017), malaria (Ocampo et al., 2013), cancer (Ofran et al., 2012; Paul et al., 2016), asthma (Ram et al., 2015), urinary tract infection (Rossignol et al., 2013), and sexually transmitted infections (Johnson and Mehta, 2014).

Search data can also be analyzed from domain-specific websites, such as

PubMed (Mosa et al., 2015).[5] However, these data are often accessed through private services not publicly obtainable. For example, Santillana et al. (2014a) utilized the search data of UpToDate, which is a disease database used by clinicians, to predict influenza prevalence in the US.[6]

One advantage of search data is that it can cover the situation when users have serious and stigmatizing health conditions (e.g. HIV and sexually transmitted diseases). However, search data often misses the reason behind the search. With a query of several words, it is hard to investigate further. In addition, search data publicly available is often aggregated across users and locations. This makes analysis at the user level difficult.

2.1.3 Wikipedia

Wikipedia is a public source of browsing. Education level is the strongest predictor of Wikipedia use. The collaborative encyclopedia is most popular among internet users with at least a college degree, 69% of whom use the site. Additionally, Wikipedia is generally more popular among those with annual household incomes of at least $\$50,000$, as well as with young adults: 62% of internet users under the age of 30 using the service, compared with only 33% of internet users aged 65 and older.

Page view statistics of Wikipedia can be obtained from the Wikipedia dumps website.[7] This data can be used to measure the levels of interest in articles such as "Influenza" (McIver and Brownstein, 2014). Figure 2.3 shows the number of page views for the "Influenza" article on Wikipedia, we can see that the high volume appears between November to February each year. This corresponds to flu activity in the real world. Generous et al. (2014) and Priedhorsky et al. (2017) also used Wikipedia to investigate multiple diseases. However, Wikipedia logs do not contain information about the locations of the readers, which makes location-focus studies hard to undertake.

[5] **PubMed**, https://www.ncbi.nlm.nih.gov/pubmed/
[6] **UpToDate**, https://www.uptodate.com/
[7] **Wikipedia dumps**, https://dumps.wikimedia.org/

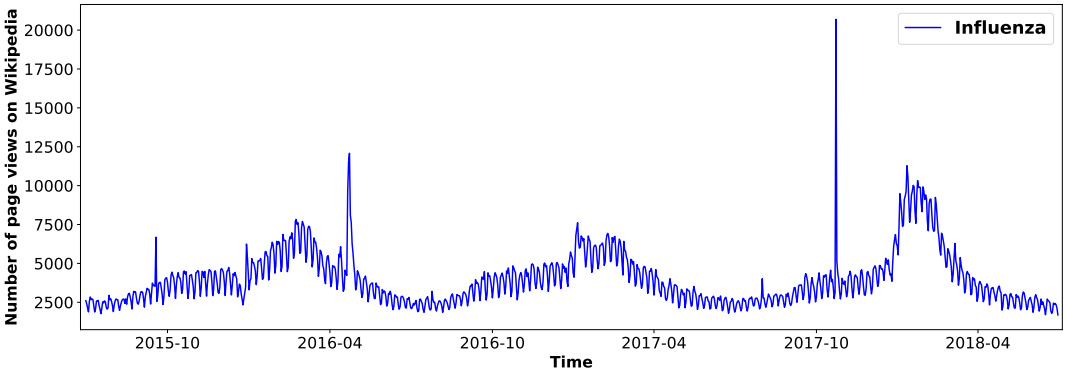


Figure 2.3: Daily number of page views for the "Influenza" on Wikipedia from July 2015 to July 2018.

2.1.4 Microblogs

Microblogs, such as Twitter[8] and Sina Weibo[9], are social media platforms where users share "status updates". Twitter enables users to send and read short 140-character "tweets". Twitter can provide information of users such as their current activities, locations, feelings, thoughts, and social surroundings. Microblogs are designed for broadcasting information to a public audience. Hence, content on these platforms is often public. For example, Twitter data are publicly accessible through the Twitter API.[10] By aggregating the words used by millions of Twitter users, we can approximately infer the level of interest for a topic. Note that Twitter streaming API only allows quite low volume (1%) tweets extraction, but keyword match is available in Twitter Advanced Search.

According to Table 2.1, younger people are more likely than older to be on Twitter. Some 36% of online adults ages $18-29$ are on Twitter, more than triple the share among online adults ages 65 and older (just 10% of whom are Twitter users). Twitter is also more popular among the highly educated: 29% of internet users with college degrees use Twitter, compared with 20% of those with high school degrees or less.

Twitter is also a popular data source for disease surveillance. It has been used to study influenza surveillance (De Quincey and Kostkova, 2009; Chew and

[8] **Twitter**, https://twitter.com/
[9] **Weibo**, https://www.weibo.com/
[10] **Twitter API**, https://developer.twitter.com/

Eysenbach, 2010; Culotta, 2010; Lampos and Cristianini, 2010; Signorini et al., 2011; Achrekar et al., 2012; Lamb et al., 2013; Velardi et al., 2014; Paul et al., 2015; Sun et al., 2016), ebola virus diseases (Fung et al., 2014; Odlum and Yoon, 2015), infectious intestinal diseases (Diaz-Aviles and Stewart, 2012; Zou et al., 2016), Zika virus (Juric et al., 2017; Miller et al., 2017), asthma rates (Zhang et al., 2016; Dai et al., 2017), diabetes (Liu et al., 2016b), human immunodeficiency virus (HIV) (Han et al., 2016; Young et al., 2017), cardiac diseases (Bosley et al., 2013), suicide (Robinson et al., 2016), vaccination (Wagner et al., 2017). Weibo is mainly used for studying health conditions from Chinese content (Zhang et al., 2014a; Sun et al., 2014; Feng and Hossain, 2016; Li and Hu, 2016; Sun et al., 2017).

Twitter and Weibo are good sources for identifying common, real-time trends. Topics like influenza are often discussed in the population at large, so it is well represented on Twitter. Microblogs also enables the study at user level, where search data cannot. However, Twitter is unsuitable to study the serious and stigmatizing health conditions. For instance, people are more likely to search for "I've been diagnosed with HIV", than to tweet it. Furthermore, tweets can be ambiguous. For example, tweet "I had Bieber fever" does not mean a person had a health problem. In contrast, search queries, which usually consist of several terms, are topical and less ambiguous.

2.1.5 Other Social Media Data

Social networks, such as Facebook and Linkedin, allow users to connect with each other. Different from microblogs, where information is usually broadcasted, information on social networks is only available to limited audience, such as friends. Social networks are designed for maintaining relationships, and data are often private. For these reasons, they are less commonly used for disease surveillance.

Some media sharing platforms, such as Instagram[11] and Pinterest[12], are primarily used for sharing images and videos. Some specific behavior can be studied using this data. For example, De Choudhury et al. (2016) studied dietary choice,

[11] **Instagram**, https://www.instagram.com/
[12] **Pinterest**, https://www.pinterest.com/

nutrition, and language in food desserts via Instagram. Morgan et al. (2010) studied alcohol consumption, inebriated behavior, and recreational marijuana use using images and videos from MySpace[13] and YouTube[14]. Social media sharing platforms are often private, where data is difficult to access. In addition, information from images and videos are harder to extract, compared to text data from microblogs and search data.

User reviews are a specific type of social media, where users write reviews of services and products. This data can also be used to study diseases. For example, Yates and Goharian (2013) detected adverse drug effects from drug review social media sites (askapatient.com, drugs.com, and drugratingz.com). Harrison et al. (2014) monitored a restaurant review website, Yelp, to detect food poisoning outbreaks.[15] These kinds of domain-specific social media, are suitable for in-depth studies of a specific health condition. Given a topic or product, usually in-depth discussions from different users are available. This is especially good for topics that are not common in the general population. Furthermore, this data may contain years of data, making longitudinal study possible.

2.2 Feature Extraction

In disease surveillance, features are frequencies of keywords (e.g. a query or n-grams) obtained from Web data. The core question is the choice of keywords. In the following part, we review three methods for identifying these keywords.

In the first method, a keyword dictionary is manually defined, and an exact match method is used to count the frequencies of these predefined keywords. For example, Polgreen et al. (2008) extracted search queries that contain the terms "influenza" or "flu", but do not contain the terms "bird", "avian", or "pandemic". Lampos and Cristianini (2010) defined a set of 41 n-grams expressing flu symptoms or relevant terminology, e.g. "fever", "temperature", "sore throat", "infection", "headache" and so on, then extracted the frequencies of these keywords from

[13] **MySpace**, https://myspace.com/
[14] **YouTube**, https://www.youtube.com/
[15] **Yelp**, https://www.yelp.com

tweets collected from the streaming API as features. This method has also been used in (Chew and Eysenbach, 2010; Culotta, 2010, 2013; Broniatowski et al., 2013; Paul et al., 2014; Jin et al., 2014; Towers et al., 2015; Wong et al., 2017). Keyword and phrase-based matching is thought to be especially effective for search queries, which are typically very short and direct, compared to longer text, like social media messages (Carmel et al., 2014). However, it is limited because it does not distinguish between different contexts in which words or phrases appear. For example, not all tweets that mention "fever" indicate that the user is sick with fever; a tweet might mean a user is very interested in a star (for example, "Bieber fever") that is irrelevant to disease surveillance.

The second method extends the first method using correlation. A dictionary of keywords that are relevant to a target disease is firstly defined, then queries that are highly correlated (in terms of frequencies) or co-occur with dictionary terms are extracted as candidate queries. For example, Yang et al. (2015); Lampos et al. (2015, 2017); Zou et al. (2018) utilized Google Correlate to identify queries that are relevant to the flu topic.[16] Zou et al. (2016) utilized the co-occurrence method to identify terms that are correlated with gastrointestinal-related terms on tweets.

In the third method, topic modeling is used to find keywords relevant to a topic. Topic modeling is a type of unsupervised learning. They are statistical models that treat text documents as if they are composed of underlying "topics", where each topic is defined as a probability distribution over words and each document is associated with a distribution over topics. Topics models cluster together words into topics, which then allows documents with similar topics to be clustered. Topic modeling methods have been used in (Brody and Elhadad, 2010; Paul and Dredze, 2011; Prier et al., 2011; Wang et al., 2014; Paul and Dredze, 2014; Chen et al., 2016).

[16] **Google Correlate**, https://www.google.com/trends/correlate

2.3 Feature Selection

A prevalent paradigm, evident in disease surveillance, is the formulation of a supervised learning task based on a textual representation of user-generated content (Choi and Varian, 2012; Rao et al., 2010). This often involves a large number of features, but a moderate number of training samples. In such tasks, it is common to rely on applying a statistical method to maintain the most relevant features (Lampos et al., 2013; Park et al., 2015). In this section, we review relevant feature selection methods.

According to (Liu and Motoda, 2012; Guyon and Elisseeff, 2003; Chandrashekar and Sahin, 2014), feature selection methods can be classified into three categories: filter methods, wrapper methods and embedded methods. Filter methods select subsets of variables as a pre-processing step, independently of the chosen features. Wrappers utilize the learning machine of interest as a black box to select subsets of variables according to their predictive capacity. Embedded methods perform variable selection in the processing of training and are usually specific to given learning machines. In disease surveillance, only filter methods and wrapper methods have been used, so we only discuss these two.

2.3.1 Filter Methods

Filter methods use variable ranking techniques as the principal criteria for variable selection. Variable ranking makes use of a scoring function computed from the features and the ground truth (disease rates). This score function measures the feature's usefulness in predicting the target. Usually a high score is indicative of a valuable variable and that we sort variables in decreasing order. Variable ranking is usually a pre-processing step for selecting features (Kohavi and John, 1997).

Variance based methods are commonly used as filters. Given a threshold value, a feature with a variance lower than the threshold will be removed, since it is not informative. In disease surveillance, Pearson correlation is a widely used method for feature selection. A feature that has a high correlation with target disease rates (ground truth) is considered to be a good feature. The correlation method has been used in many disease surveillance works, such as influenza (Culotta, 2010; Paul

et al., 2014; Lampos et al., 2015; Yang et al., 2015; Lampos et al., 2017; Zou et al., 2018), and infectious intestinal diseases (Zou et al., 2016). However, correlation can only measure linear dependency.

Information theoretic metrics are also popular variable ranking criteria for feature selection. One such approach relies on empirical estimates of the mutual information the variable and the target. This criterion is a measure of dependency between the density of variable features and the density of the target. The probability densities can be estimated from frequency counts or an approximating method such as Parzen windows (Torkkola, 2003). The advantage of information theoretic criteria is that it can detect not only the linear, but also the nonlinear dependencies between the variable and the target. It has been successfully used to detect the causality between social media and stock prices in (Souza and Aste, 2016). However, information theoretic metrics are rarely used in disease surveillance. Because this method ignores the time dependency of sample points when estimating the probability density.

2.3.2 Embedded Methods

Embedded methods incorporate feature selection as part of the training process, usually in a supervised learning setting. Embedded methods are more efficient since they reach a solution faster by avoiding retraining a feature from scratch for every variable subset investigated (Guyon and Elisseeff, 2003).

In disease surveillance, Lasso and elastic net are two embedded methods that incorporate feature selection in the model training process. By including the ℓ^1 norm in the optimization function, weights of some features can shrunk to zero during training, making the model able to select features. In section 2.4, we review and discuss this method in more detail.

2.4 Inference

Inference models are the core part of the disease surveillance framework (see Figure 2.1). In this thesis, we focus on regression models, which map interest scores to ground truth (gold standard) values from existing surveillance systems. In this

section, we first introduce the disease rates, which are ground truth, and then review and discuss different linear and nonlinear regression models.

2.4.1 Mathematical Notations

In disease surveillance, our aim is to infer disease rates as reported by an established health surveillance system using the frequencies of text markers we extract from Web data (e.g. search queries and *n*-grams). We formulate this as a regression task, where we learn a function

$$f : \mathbf{X} \rightarrow \mathbf{y} \tag{2.1}$$

that maps the input space $\mathbf{X} \in \mathbb{R}^{n \times p}$ to the target variable $\mathbf{y} \in \mathbb{R}^n$; n denotes the number of samples and p is the size of our feature space, i.e. the number of unique text markers we consider. $\mathbf{X}$ contains normalized frequencies of text markers for a specified time interval and $\mathbf{y}$ has the disease rates for the same time intervals as reported by the health agency. A normalized frequency is defined as the count of a text marker divided by the total number of queries or tweets during a fixed time interval, e.g. one week.

2.4.2 Ground Truth

Disease surveillance models utilize supervised learning techniques, and disease rates are used as ground truth for training the model. Different health agencies have different definitions for disease rates. We take influenza as an example, and review ILI rates defined in different countries.

In the US, ILI rates are published by CDC. These rates represent the average percentage of all outpatient visits to health care providers normalized by the respective regional population figures and are records by CDC's ILI surveillance network, ILINet. ILI rates are weekly. CDC also publishes ILI rates for 10 regions defined by Department of Health and Human Services (HHS).

In the UK, ILI rates are reported by the Royal College of General Practitioners (RCGP)[17] and Public Health England (PHE)[18]. ILI rates represent the number

[17] **RCGP**, http://www.rcgp.org.uk/
[18] **PHE**, https://www.gov.uk/government/organisations/public-health-england

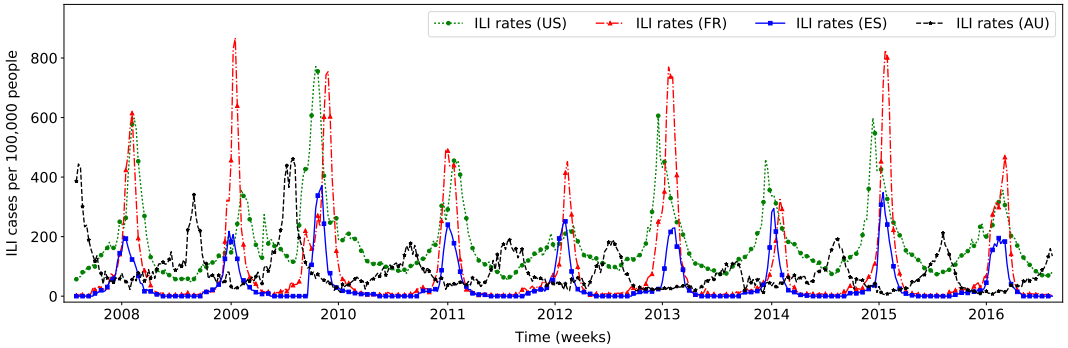


Figure 2.4: ILI rates of US, France, Spain, and Australia from September 2007 to September 2016.

of doctor consultations reporting ILI symptoms per 100,000 people. The data is weekly, and PHE also report regional ILI rates for England, Wales, Scotland, and Northern Ireland.

In France, ILI rates are published by French GPs Sentinelles Network (SN).[19] As for the UK, ILI rates in France are weekly, and represent the number of ILI cases seen in General Practices per 100,000 inhabitants. ILI rates in France are also available at a regional level for 13 metropolitan regions.

In Spain, ILI rates are published by the Spanish Influenza Sentinel Surveillance System (SISSS).[20] The data is weekly, and represent the number of cases with ILI symptoms per 100,000 people. Note that SISSS only report ILI rates during flu seasons (from November to March next year, around 20 weeks). The data is only available nationally.

In Australia, ILI rates are published by Australian Sentinel Practices Research Network (ASPREN).[21] The data is weekly, and represent the number of ILI cases per 10,000 consultations. The data is only available nationally.

To better introduce the disease rates, in Figure 2.4 we plot ILI rates of some countries. All ILI rates in the figure were converted to the same scale. Note that the use of weekly average reporting are quite sensible in the light of reporting fluctuations, e.g. weekends/holidays/sporting events. Web-based disease surveillance can complement this.

[19] **SN**, https://websenti.u707.jussieu.fr/sentiweb/
[20] **SISSS**, http://www.eng.isciii.es/ISCIII/
[21] **ASPREN**, https://aspren.dmac.adelaide.edu.au/

2.4.3 Linear Regression

Given the input features $\mathbf{X}$ and disease rates $\mathbf{y}$, the aim of linear regression is to find a weight vector $\mathbf{w} \in \mathbb{R}^p$, such that

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \beta\,, \tag{2.2}$$

where β an intercept term. When the number of feature is 1, $\mathbf{X}$ is a vector, and the model is simple linear regression; when the number of feature is bigger than 1, $\mathbf{X}$ is a matrix, and the model is multiple linear regression.

After $\mathbf{w}$ and β are fixed, given a new observation $\mathbf{x}_*$, the estimation can be made using

$$y_* = \mathbf{x}_*\mathbf{w} + \beta\,. \tag{2.3}$$

In the following part, we review and discuss different methods for learning weight vector $\mathbf{w}$ and intercept term β.

2.4.3.1 Ordinary Least Squares

Ordinary least squares is a type of linear least squares method for estimating the unknown parameters in linear regression models. It is defined as

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left\{ \|\mathbf{X}\mathbf{w} + \beta - \mathbf{y}\|_2^2 \right\}, \tag{2.4}$$

where $\|\cdot\|_2$ is ℓ^2-norm. Given a vector $\mathbf{x}$, ℓ^2-norm is defined as

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{p} |x_i|^2}\,. \tag{2.5}$$

Equation (2.4) minimizes the squares of the difference between the predictions $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} + \beta$ and the target values $\mathbf{y}$. We define

$$\begin{aligned} \mathbf{e} &= \hat{\mathbf{y}} - \mathbf{y} \\ &= \mathbf{X}\mathbf{w} + \beta - \mathbf{y}\,, \end{aligned} \tag{2.6}$$

where **e** is called the residual. Ordinary least squares minimizes the vertical distance between the predicted points and the observed points.

In disease surveillance, ordinary least squares has been used in (Polgreen et al., 2008) to predict influenza outbreaks. Ginsberg et al. (2009) also employed a linear regression model. Compared to the work in (Polgreen et al., 2008), a logit function is used on input and output. Given input features $\mathbf{X}$ and disease rates $\mathbf{y}$, their formulation is

$$\text{logit}(\mathbf{y}) = \beta + \text{logit}(\mathbf{X})\mathbf{w}, \tag{2.7}$$

where the logit function is defined as

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \tag{2.8}$$

where p is a number between 0 and 1. Ordinary least squares are then used for learning $\mathbf{w}$ and β. The logit function ensures the inferred values are always positive and improves the performance when the range of features is huge. The logit function can also be seen as applying a nonlinear transformation on the the input and output.

2.4.3.2 Ridge Regression

The analytical solution of the weight vector $\mathbf{w}$ to Equation (2.4) is

$$\mathbf{w} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}. \tag{2.9}$$

When multicollinearity (also known as collinearity) exists in $\mathbf{X}$, i.e. one prediction variable in $\mathbf{X}$ can be linearly predicted from the others with a substantial degree accuracy, $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ may not exist or be close to zero. This causes the solution to be highly unstable. The problem can be tackled by adding an ℓ^2-norm on $\mathbf{w}$ during optimization. This is called ridge, and is given by

$$\underset{\mathbf{w},\beta}{\text{argmin}} \left\{ \|\mathbf{X}\mathbf{w} + \beta - \mathbf{y}\|_2^2 + \lambda_2 \|\mathbf{w}\|_2^2 \right\}, \tag{2.10}$$

where $\|\cdot\|_2$ is ℓ^2-norm, and λ_2 is a parameter that controls the level of regularization. Ridge regression can shrink some weights close to zero. Culotta (2010) utilized ridge regression for predicting influenza epidemics.

2.4.3.3 Lasso

Ridge regression can handle multicollinearity, but it cannot perform feature selection. Least absolute shrinkage and selection operator (Lasso) added an ℓ^1-norm to the optimization, and performs feature selection when learning $\mathbf{w}$ and β. Lasso is defined as

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left\{ \|\mathbf{X}\mathbf{w} + \beta - \mathbf{y}\|_2^2 + \lambda_1 \|\mathbf{w}\|_1 \right\}, \tag{2.11}$$

where $\|\cdot\|_1$ is ℓ^1-norm, and λ_1 is a parameter that controls the level of regularization. ℓ^1-norm is defined as

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{p} |x_i| . \tag{2.12}$$

Different from Ridge regression that uses ℓ^2-norm as a regularizer, Lasso regularizes the coefficients $\mathbf{w}$ using ℓ^1-norm, which is the sum of coefficient absolute values. This can shrink some weights to be zero, and consequently perform feature selection. Lampos and Cristianini (2010) utilized Lasso for inferring ILI rates in the UK.

2.4.3.4 Elastic Net

Lasso performs feature selection, but it cannot make a consistent selection of the true model, when collinear predictors are present in the data. A more robust generalization of Lasso, namely elastic net, can be employed. Given $\mathbf{X}$ and $\mathbf{y}$, the aim of elastic net is to find a weight vector $\mathbf{w} \in \mathbb{R}^p$, such that $\mathbf{y} = \mathbf{X}\mathbf{w} + \beta$. The weight vector is obtained by minimizing the following function

$$\underset{\mathbf{w},\beta}{\operatorname{argmin}} \left\{ \|\mathbf{y} - \mathbf{X}\mathbf{w} - \beta\|_2^2 + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2 \right\}, \tag{2.13}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2^2$ are ℓ^1-norm and ℓ^2-norm, λ_1 and λ_2 are parameters that control the level of regularization. The first term of Equation (2.13) minimizes the residuals

between our predictions $\mathbf{X}\mathbf{w} + \beta$ and the target values $\mathbf{y}$, while the second and third terms of Equation (2.13) add penalty on the weight vector $\mathbf{w}$.

Elastic net can be seen as a generalization of the ℓ^1-norm regularization, known as the Lasso (Tibshirani, 1996), because it also applies an ℓ^2-norm, or ridge (Hoerl and Kennard, 1970), regularizer on the inferred weight vector. The combination of the two regularizers 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 (Zhao and Yu, 2006). Elastic net has been used in influenza surveillance (Lampos et al., 2015, 2017; Zou et al., 2018) and infectious intestinal surveillance (Zou et al., 2016).

2.4.4 Nonlinear Regression

Linear regression models may ignore the presence of possible nonlinearities in the data (Lampos et al., 2015). Thus, nonlinear regression models have been explored for disease surveillance.

2.4.4.1 Kernel Methods

Kernel methods, especially Gaussian Processes (GP), have been used for disease surveillance. Gaussian Processes is a family of statistical distributions. In a Gaussian Process, each point in the input space is considered to be a random variable that follows the Gaussian distribution, and the finite collection of these random variables follow a multivariate Gaussian distribution (Rasmussen and Williams, 2006). Gaussian Processes can be utilized as a prior probability distribution over functions in Bayesian inference for prediction. In the regression framework, Gaussian Processes observed the coordinates $\mathbf{X}$, and the vector of values $\mathbf{y}$ is just one sample from a multivariate Gaussian distribution. For the inputs $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$ (both expressing rows of the input features $\mathbf{X}$), our aim is to learn a function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}$ that is drawn from a $\mathcal{GP}$ prior,

$$\mathbf{f} \sim \mathcal{GP}(\mu(\mathbf{x}),\, C(\mathbf{x}, \mathbf{x}')), \tag{2.14}$$

which means the random function $\mathbf{f}$ is distributed as a $\mathcal{GP}$ with mean function $\mu(\mathbf{x})$ and covariance function $C(\mathbf{x}, \mathbf{x}')$.

Gaussian Processes have been used to study influenza surveillance (Lampos et al., 2015, 2017; Zou et al., 2018), and infectious intestinal diseases surveillance (Zou et al., 2016).

2.4.4.2 Neural Networks

In the last decade, neural networks, in particular deep neural networks, have proved to perform well in many ares of computer science, such as computer vision, natural language processing and speech recognition (Goodfellow et al., 2016). Neural networks can be used for regression and classification tasks. However, for infectious disease surveillance, only weekly ground truth are available, making deep neural networks hard to train, since the number of parameters in deep neural networks is much larger than the number of samples.

2.4.5 Forecasting Models

The linear and nonlinear regression models discussed above, focus on estimating the disease rates for current week, i.e. they predict the present; in regression models, the features are from the Web. Other features can also be included in regression models. For example, a good feature is the trend itself: the previous weeks's value is a good feature for the current week. This is known as forecasting, which is a process of making predictions of the future based on past and present data and most commonly by analysis of trends.

Autoregressive models have also been used to forecast disease rates. Different from regression models, an autoregressive moving average (ARMA) model expresses the conditional mean of y_t as a function of both past observations, $y_{t-1}, \ldots, y_{t-a}$, and past residual terms, $\varepsilon_{t-1}, \ldots, \varepsilon_{t-b}$. The number of past observations that y_t depends on, a, is the autoregressive (AR) degree. The number of past innovations that y_t depends on, b, is the moving average MA degree. The model is denoted as ARMA(a,b), and is formulated as

$$\begin{aligned} y_t &= c + \alpha_1 y_{t-1} + \ldots + \alpha_a y_{t-a} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_b \varepsilon_{t-b} \\ &= c + \varepsilon_t + \sum_{i=1}^{a} \alpha_i y_{t-i} + \sum_{i=1}^{b} \theta_i \varepsilon_{t-i} \end{aligned} \tag{2.15}$$

where ε_t is an uncorrelated white noise error term with mean zero and constant variance.

The autoregressive moving average model with exogenous inputs (ARMAX) model (Box et al., 2015) refers to the model with a autoregressive terms, b moving average terms and d exogenous inputs terms. This model contains the AR(a) and MA(b) models and a linear combination of the last d terms of a known and external time series X_t. It is given by:

$$y_t = c + \varepsilon_t + \sum_{i=1}^{a} \alpha_i y_{t-i} + \sum_{i=1}^{b} \theta_i \varepsilon_{t-i} + \sum_{i=1}^{d} \eta_i X_{t-i}\,. \tag{2.16}$$

ARMA models have been used for influenza prediction from social media (Achrekar et al., 2012; Paul et al., 2014).

A commonly used extension to the linear autoregressive model is the autoregressive integrated moving average (ARIMA) model, which assumes an underlying smooth behavior in the time series. ARIMA models are applied in the cases where data is non-stationary, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity. ARIMA models have also been used for predicting influenza prevalence (Dugas et al., 2013; Preis and Moat, 2014; Broniatowski et al., 2015).

2.5 Evaluation Metrics

To evaluate the effectiveness of disease surveillance models, different evaluation metrics are applied in the literature, including Pearson correlation, coefficient of determination, mean absolute error, mean squared error, and root mean squared error. In this section, we review and discuss these evaluation metrics.

2.5.1 Pearson Correlation

The Pearson correlation coefficient r, also known as Pearson product-moment correlation coefficient, is a measure of the linear correlation between two variables. It has a value between -1 and 1, where 1 is a total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. It was devel-

oped by Galton (1886) and Pearson (1895). Denote variable $\mathbf{y}$ as the disease rates obtained from health agencies, and $\hat{\mathbf{y}}$ as the predicted disease rates by inference models, Pearson correlation is defined as

$$\begin{aligned} r &= \frac{\text{cov}(\mathbf{y}, \hat{\mathbf{y}})}{\sigma_y \sigma_{\hat{y}}} \\ &= \frac{S_{y\hat{y}}}{\sqrt{S_{yy}}\sqrt{S_{\hat{y}\hat{y}}}} \\ &= \frac{\sum_{i=1}^{n}(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_{\hat{y}})^2}}, \end{aligned} \tag{2.17}$$

where cov is the covariance, σ_y and $\sigma_{\hat{y}}$ are the standard deviations of $\mathbf{y}$ and $\hat{\mathbf{y}}$, μ_y and $\mu_{\hat{y}}$ are the mean of $\mathbf{y}$ and $\hat{\mathbf{y}}$, S_{yy} and $S_{\hat{y}\hat{y}}$ are sum of squares for $\mathbf{y}$ and $\hat{\mathbf{y}}$.

2.5.2 Coefficient of Determination

The coefficient of determination, denoted as R^2, is the proportion of the variance in the dependent variable that is predictable from the independent variables, i.e. the fraction of the total variance that is explained by the linear relation between the observed disease rates $\mathbf{y}$ and predicted disease rates $\hat{\mathbf{y}}$. It is defined as

$$\begin{aligned} R^2 &= 1 - \frac{S_{y\hat{y}}}{S_{yy}} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu_y)^2}, \end{aligned} \tag{2.18}$$

where $S_{y\hat{y}}$ is the residual sum of squares between $\mathbf{y}$ and $\hat{\mathbf{y}}$ (Weisberg, 2005). We investigate R^2 further and have

$$\begin{aligned} R^2 &= \frac{S_{yy} - S_{y\hat{y}}}{S_{yy}} \\ &= \frac{S_{yy} - \left(S_{yy} - \frac{S_{y\hat{y}}^2}{S_{\hat{y}\hat{y}}}\right)}{S_{yy}} \\ &= \frac{S_{y\hat{y}}^2}{S_{yy}S_{\hat{y}\hat{y}}} \\ &= r^2, \end{aligned} \tag{2.19}$$

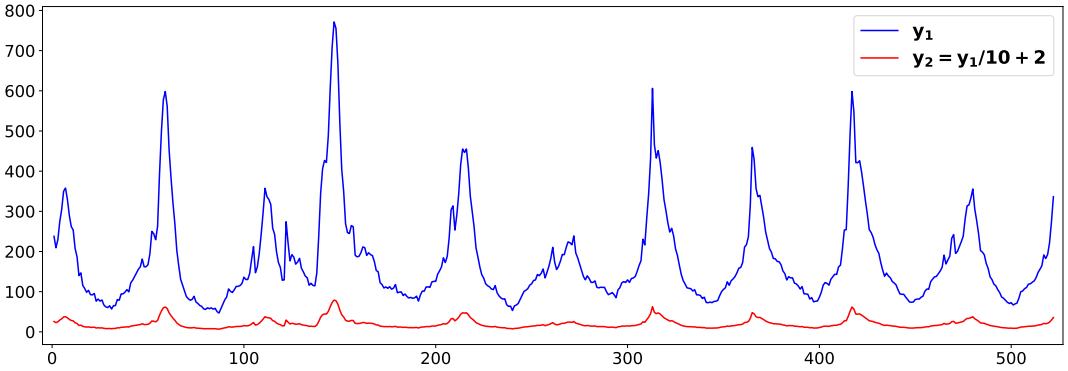


Figure 2.5: Example showing that correlation that is invariant to linear transformation. The figure plots time series of $\mathbf{y}_1$ and $\mathbf{y}_2$. Time series of $\mathbf{y}_2$ is generated through $\mathbf{y}_2 = \mathbf{y}_1/10+2$. The correlation between $\mathbf{y}_1$ and $\mathbf{y}_2$ is 1.0. However, the difference between $\mathbf{y}_1$ and $\mathbf{y}_2$ is big.

and thus R^2 is the same as the square of the correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$. Coefficient of determination R^2 measures the goodness of linear fit of variables, and has a value between 0 and 1.

2.5.3 A Discussion on Pearson Correlation and Coefficient of Determination

Section 2.5.1 and 2.5.2 introduces Pearson correlation r and coefficient of determination R^2. However, they can only detect the linear dependencies between $\mathbf{y}$ and $\hat{\mathbf{y}}$. Nonlinear preprocessing (e.g. squaring, taking the square, and the log) is needed when we want to detect nonlinear relations between the variable and the target.

In addition, we should notice that correlation is invariant under separate changes in location and scale in the two variables, i.e. it is invariant to linear transformation. Sometimes correlation can be misleading. In Figure 2.5, we plot the time series of $\mathbf{y}_1$ and $\mathbf{y}_2$, where time series of $\mathbf{y}_2$ is generated through $\mathbf{y}_2 = \mathbf{y}_1/10+2$. The correlation between $\mathbf{y}_1$ and $\mathbf{y}_2$ is 1.0. However, the difference between $\mathbf{y}_1$ and $\mathbf{y}_2$ is big, i.e. $\mathbf{y}_2$ is not a good estimation for $\mathbf{y}_1$. Therefore, when we evaluate the effectiveness of disease surveillance models, we have to combine correlation r and error metrics we define below. Correlation metrics are appropriate when the trends are on different scales.

2.5.4 Mean Absolute Error

Mean absolute error (MAE) is a measure of difference between two continuous variables. Given the the disease rates obtained from health agencies $\mathbf{y}$, and predicted disease rates by inference models $\hat{\mathbf{y}}$, mean absolute error is defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \,. \tag{2.20}$$

Mean absolute error is the average vertical distance between each point and the $y = \hat{y}$ line, which is also known as the One-to-One line. It has a value from 0 to infinity. A MAE of 0 means $\mathbf{y}$ and $\hat{\mathbf{y}}$ are identical, i.e. we make a perfect prediction. The bigger the MAE is, the worse the inference model is.

2.5.5 Mean Squared Error and Root Mean Square Error

Mean squared error (MSE) or mean squared deviation is similar to mean absolute error, but they are different. MSE measures the average of the squares of the errors, which is the average squared difference between the predicted disease rates $\hat{\mathbf{y}}$ and the observed disease rates $\mathbf{y}$. MSE is defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \,. \tag{2.21}$$

Root mean square error (RMSE) or root mean squared deviation, represents the square root of the differences between the predicted disease rates $\hat{\mathbf{y}}$ and the observed disease rates $\mathbf{y}$. RMSE is defined as

$$\text{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \,. \tag{2.22}$$

2.5.6 A Discussion on Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error

Both MAE and RMSE express average model prediction error in units of the variable of interest, while MSE is in squared units. All three metrics range from 0 to infinity, and they are negatively oriented scores, which means lower values are

better.

We have to notice that MSE and RMSE have squared the errors, which means MSE and RMSE give a relatively high weight to large errors. Therefore, MSE and RMSE have the benefit of penalizing large errors more so can be more appropriate in some cases. If a system makes small errors on average, but has some very large errors, then those large errors will affect MSE (or RMSE) more than MAE. This can be a useful property if we care about having no or few large errors, even if that makes other errors slightly worse. If we focus on the performance of inference models during outbreak periods, MSE or RMSE are more appropriate.

2.5.7 AIC and BIC

Another category of metrics measures model fit, i.e. how well the model explains or matches the data. Closely related to mean squared error is the log-likelihood of the true values under the regression model. Akaike information criterion (AIC) (Akaike, 1974) is a common metric that is based on log-likelihood, but adjusts the score to penalize models with large numbers of parameters, since more complex models models may not generalize well to future data. Another similar metric, Bayesian information criterion (BIC) (Schwarz, 1978) can also be used. They are defined by

$$\begin{aligned} \text{AIC} &= n \log\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}\right) + p_C \\ \text{BIC} &= n \log\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}\right) + p_C \log n \end{aligned} \tag{2.23}$$

where p_C is the number of free parameters in the inference model.

Chapter 3

Enhancing Feature Selection using Word Embeddings

A prevalent paradigm in disease surveillance is the formulation of a supervised learning task based on a textual representation of user-generated content (Polgreen et al., 2008; Lampos and Cristianini, 2010; Paul et al., 2014; Yang et al., 2015). This often involves a large number of features, but a moderate number of training samples, i.e. we have a $p \gg n$ problem, where p denotes the number of features, and n denotes the number of samples. When the problem appears, there are insufficient degrees of freedom to estimate the full model. To tackle the problem, it is common to apply statistical methods that are able to project the data to a lower dimensional space or maintain the most relevant features (Lampos et al., 2013; Park et al., 2015). This can be done using the statistical feature selection methods we discussed in Section 2.3 such as Lasso, and elastic net. A common criticism of such approaches is that some of the selected features may have little or no semantic link to the regression task.

In Figure 3.1, we plot the frequencies of query "undergraduate internships" (red line), "ski racing" (blue line), and "high school basketball teams" (green line) in the US against the ILI rates obtained from CDC over a period of September 2008 to 2016. All frequencies of queries as well as the ILI rates are z-scored such that they can be compared using the same scale.[1] The three queries are highly

[1] Z-score is defined as $z = \frac{x-\mu_x}{\sigma_x}$, where μ_x and σ_x are the population mean and standard deviation.

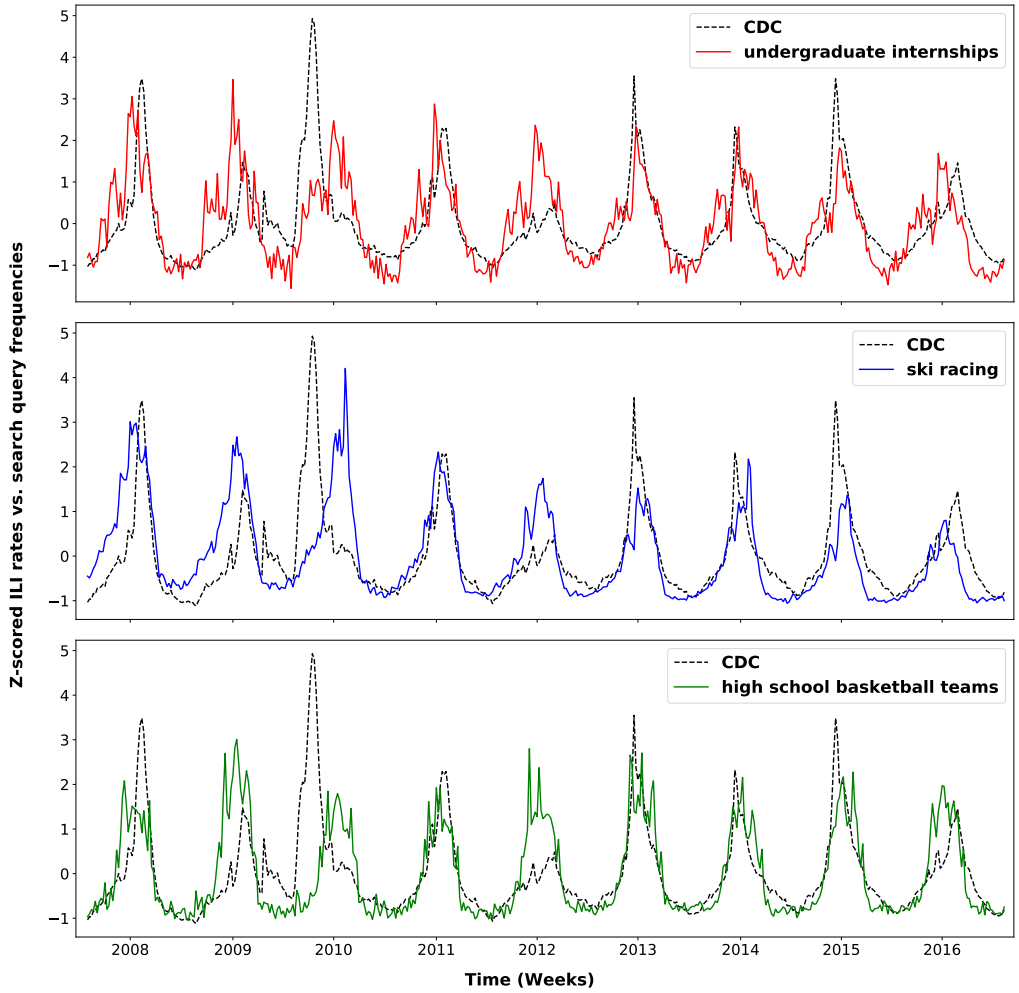


Figure 3.1: Example queries that are correlated with ILI rates but are irrelevant to the "flu" topic. The black (dotted) line represent z-scored ILI rates obtained from CDC from September 2008 to September 2016. The red, blue, and green lines represents z-scored frequencies of queries that are highly correlated with ILI rates "undergraduate internships" ($r = 0.606$, $p < 0.01$), "ski racing" ($r = 0.552$, $p < 0.01$), and "high school basketball teams" ($r = 0.581$, $p < 0.01$) in US for the same period, respectively.

correlated with the time series of ILI rates. Query "undergraduate internships" has a correlation of 0.606 with ILI rates ($p < 0.01$), query "ski racing" has a correlation of 0.552 ($p < 0.01$), and query "high school basketball teams" has a correlation of 0.581 ($p < 0.01$). These three events seasonally peak between December and January each year, and this highly correlates with the seasonality of flu activity in winter. However, these three queries are irrelevant to the "flu" topic, and using them as features can potentially lead to errors.

To alleviate this effect, methods in natural language processing have incorpo-

rated classification schemes, aiming to encourage a relatedness between the input information and the target concept (Aramaki et al., 2011; Bollen et al., 2011; Choudhury et al., 2013; Paul and Dredze, 2011). However, the classification models often require considerable human effort, especially in obtaining a sufficient number of labels, and are limited to a specific task.

In this chapter, we take advantage of current developments in statistical natural language processing and propose a method that is able to overcome the aforementioned deficiencies. We form general textual concepts by adopting word embeddings (Mikolov et al., 2013c), and then use them in conjunction with conventional feature selection methods to encourage a level of topicality. This approach can be regarded as an unsupervised classification layer that favors textual features that belong to a theme of interest. We evaluate our method on two large-scale, practical, text regression tasks. In the first task, we infer the ILI rates from time series of search query frequencies. Word embeddings are trained using microblogging text snippets from Twitter. Supervised learning is conducted and evaluated using official syndromic surveillance rates for ILI. Our empirical analysis shows that the proposed joint 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 topicality of the maintained textual features. In the second task, we estimate the number of Infectious Intestinal Disease (IID) cases reported by traditional health surveillance methods from Twitter. As a whole, 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.

The contributions of this chapter are listed as follows.

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

- The aforementioned approach is bound with conventional ways for feature selection, which significantly improves model reliability and, consequently, inference

performance under linear as well as nonlinear regression models.

- We conduct experiments on two large-scale, practical, text regression tasks, i.e. monitoring the ILI and gastrointestinal rates in a population, to verify the effectiveness of our proposed method. The significantly improved estimates are showcased on a live web service, the "Flu Detector."[2]

The rest of this chapter is structured as follows. We first review related work in Section 3.1. Then we give an overview of the linear and non linear models that we use for performing text regression in Section 3.2. We describe our approach in utilizing word embeddings to create concepts and refine feature selection in Section 3.3. To demonstrate the effectiveness of our proposed method, we conduct a case study on influenza-like illness surveillance in Section 3.4 and another case study on IID surveillance in 3.5. We finally conclude in Section 3.6.

3.1 Related Work

Regularization for feature selection has been routinely applied in supervised learning NLP tasks (Lampos et al., 2013; Owoputi et al., 2013; Yano et al., 2012). Word embeddings have also facilitated a number of text regression approaches, such as extending a financial lexicon for modeling risk (Tsai and Wang, 2014), or improving the inference of movie revenues based on textual reviews (Bitvai and Cohn, 2015). Notably, during initial experimentation we determined that using search query embeddings directly as features in a regression model introduced a level of compression that significantly reduced the inference performance.

Gaussian Processes models for text regression have provided solutions in NLP applications (Bitvai and Cohn, 2015; Lampos et al., 2014; Preoţiuc-Pietro et al., 2015). 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 (Lampos et al., 2015; Yang et al., 2015). Here, we use a straightforward GP kernel that is more suitable for directly assessing the predictive capacity of the selected features.

[2] **Flu Detector**, https://fludetector.cs.ucl.ac.uk/

Finally, many works have focused on disease text disambiguation by training various forms of classifiers (Collier et al., 2011; Doan et al., 2012; Paul et al., 2014), or developing laborious, task dependent NLP schemes (Lamb et al., 2013). In contrast, we have described an unsupervised, potentially task-independent approach for achieving this.

3.2 Regression methods

In regression, we learn a function f that maps an input space $\mathbf{X} \in \mathbb{R}^{n \times p}$ (where n and p respectively denote the number of samples and the dimensionality) to a target variable $\mathbf{y} \in \mathbb{R}^n$. Our input space $\mathbf{X}$ represents the frequency of p search queries (or Ngrams in tweets) during n (weekly) time intervals. In text regression, we usually operate with a high-dimensional textual feature space and a considerably smaller number of samples ($p \gg n$). To mitigate the effects of overfitting, a standard approach is to introduce a degree of regularization during the optimization of f (Hastie et al., 2009). We use a linear regression model elastic net, and a nonlinear regression Gaussian Processes. Elastic net was described in Section 2.4.3.4; we only present the Gaussian Processes approach here.

3.2.1 Nonlinear Regression using Gaussian Processes

Numerous applications have provided empirical proof for the predictive strength of Gaussian Processes in Machine Translation tasks, text and multi-modal regression problems (Beck et al., 2015; Cohn et al., 2014; Cohn and Specia, 2013; Lampos et al., 2015; Preoţiuc-Pietro et al., 2015). One caveat is that Gaussian Processes are not very efficient when operating in high dimensional spaces (Bull, 2011). Thus, while we perform modeling with a nonlinear regressor, we rely on a pre-selected subset of features.

As described in Section 2.4.4.1, Gaussian Processes are a family of statistical distributions, and are specified through a mean and a covariance (or kernel) function, i.e.

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \,. \tag{3.1}$$

By setting $\mu(\mathbf{x}) = 0$, a common practice in Gaussian Processes modeling, we focus

only on the kernel function. We use the Matérn covariance function (Matérn, 1986) to handle abrupt changes in the predictors given that the experiments are based on a sample of the original Google search (or Twitter) data. It is defined as

$$k_{\mathrm{M}}^{(\nu)}(\mathbf{x},\mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{\ell}r\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}}{\ell}r\right), \tag{3.2}$$

where K_ν is a modified Bessel function, ν is a positive constant,[3] ℓ is the length-scale parameter, and $r = \|\mathbf{x}-\mathbf{x}'\|_2$. We also use a squared exponential (SE) covariance function to capture more smooth trends in the data, defined by

$$k_{\mathrm{SE}}(\mathbf{x},\mathbf{x}') = \sigma^2 e^{-\frac{r^2}{2\ell^2}}, \tag{3.3}$$

where σ^2 is the signal variance.

We have chosen to combine these kernels through a summation. Note that the summation of Gaussian Processes kernels results in a new valid Gaussian Processes kernel (Rasmussen and Williams, 2006). An additive kernel allows modeling with a sum of independent functions, where each one can potentially account for a different type of structure in the data (Duvenaud, 2014). 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

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{2}\left(k_{\mathrm{M}}^{(\nu=3/2)}(\mathbf{x},\mathbf{x}';\sigma_i,\ell_i)\right) + \\ + k_{\mathrm{SE}}(\mathbf{x},\mathbf{x}';\sigma_3,\ell_3) + \sigma_4^2\delta(\mathbf{x},\mathbf{x}'), \end{aligned} \tag{3.4}$$

where δ is a Kronecker delta function, and σ_4^2 is the noise variance.

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.

Given a new observation $\mathbf{x}_*$, the joint distribution of the new observation and

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

the function points at the test location under the prior is given by

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathrm{K}(\mathbf{X},\mathbf{X}) & \mathrm{K}(\mathbf{X},\mathbf{x}_*) \\ \mathrm{K}(\mathbf{x}_*,\mathbf{X}) & \mathrm{K}(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix}\right), \tag{3.5}$$

where K represents the covariance matrix, which can be computed using Equation (3.4) element-wise. Deriving the conditional distribution we arrive at the key predictive equations for Gaussian Processes regression

$$\bar{y}_*|\mathbf{X},\mathbf{y},\mathbf{x}_* \sim \mathcal{N}(\mu_*, \sigma_*^2), \tag{3.6}$$

where the prediction can be calculated through

$$\mu_* = \mathbb{E}[y_*|\mathbf{y},\mathbf{X},\mathbf{x}_*] = \mathrm{K}(\mathbf{x}_*,\mathbf{X})^\mathsf{T}(\mathrm{K}(\mathbf{X},\mathbf{X}))^{-1}\mathbf{y}, \tag{3.7}$$

and the predictive uncertainty can be estimated by using the variance

$$\sigma_*^2 = \mathrm{K}(\mathbf{x}_*,\mathbf{x}_*) - \mathrm{K}(\mathbf{x}_*,\mathbf{X})^\mathsf{T}(\mathrm{K}(\mathbf{X},\mathbf{X}))^{-1}\mathrm{K}(\mathbf{x}_*,\mathbf{X}). \tag{3.8}$$

Observing Equation (3.4) we have 7 parameters in total. The parameters can be learnt by minimizing the negative log marginal likelihood

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\mathsf{T}(\mathrm{K}(\mathbf{X},\mathbf{X}) + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log|\mathrm{K}(\mathbf{X},\mathbf{X}) + \sigma_n^2\mathbf{I}| - \frac{n}{2}\log 2\pi. \tag{3.9}$$

3.3 Concept Formulation and Feature Selection

Word embeddings have been used as an input in various models and tasks in recent years (Goldberg, 2016). 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. In this section, we first introduce word embeddings and the word2vec model we use to train word embeddings, then we present our method of formulating concepts and selecting features.

3.3.1 Word Embeddings and Word2vec

Many natural language processing tasks use bag-of-words models and treat words as atomic units - there is no notion of similarity between words, as these are represented as indices in a vocabulary. Such simple techniques have been successfully applied in many natural language processing tasks (Brants et al., 2007). However, bag-of-words models ignore the context of words and do not respect the semantics of the word. For example, the words "car" and "automobile" are often used in the same context. However, the vectors corresponding to these words are orthogonal in bag-of-words models. The problem become more serious while modeling sentences. For examples, "buy used cars" and "purchase old automobiles" are represented by orthogonal vectors in bag-of-words models. But these two texts refer to almost the same thing.

Recent studies show that distributed representations of words (or word embeddings) in a vector space help learning algorithms to achieve better performance in natural language processing tasks by grouping similar words. They aim to quantify semantic similarities between linguistic terms based on their distributional properties in large samples of language data. Many different types of models were proposed for estimating continuous representations of words, including the well-known Latent Semantic Analysis (LSA) (Dumais, 2004) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). In this thesis, we focus on word embeddings learned by Recurrent Neural Networks (RNN), as it was previously shown that they perform significantly better for preserving linear regularities among words (Mikolov et al., 2013d; Zhila et al., 2013); LDA moreover becomes computationally very expensive on large data sets. In particular, we use the word2vec model proposed by Mikolov et al. (2013c,a).

The RNN based language model was proposed to overcome certain limitations of the feedforward neural network language model, such as the need to specify the context length, and because theoretically RNN can efficiently represent more complex patterns than shallow neural networks (Brants et al., 2007; Mikolov et al., 2010). The RNN model does not have a projection layer; only input, hidden and

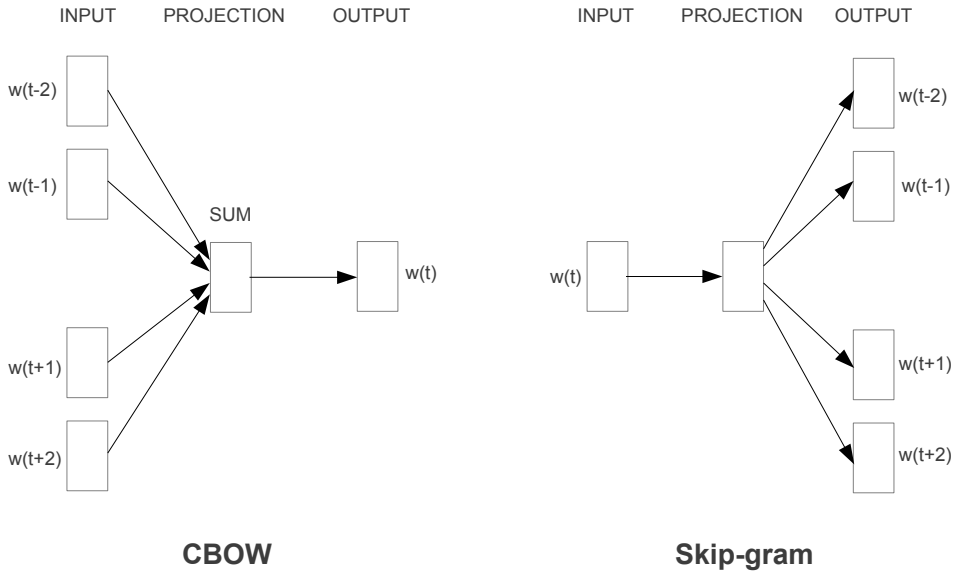


Figure 3.2: Architectures of CBOW and Skip-gram. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

output layer. What is special for this type of model is the recurrent matrix that connects hidden layer to itself, using time-delayed connections. This allows the recurrent model to form some kind of short term memory, as information from the past can be represented by the hidden layer state that gets updated based on the current input and the state of the hidden layer in the previous time step. The most time complexity of RNN is caused by nonlinear hidden layer in the model. The Continuous Bag of Words (CBOW) and Skip-Gram models are proposed to improve efficiency (Mikolov et al., 2013a,c).

Figure 3.2 displays two architectures that can be used to efficiently train word embeddings. CBOW is similar to the feedword neural networks language model, where the nonlinear hidden layer is removed and the projection layer is shared for all words; thus all words are projected into the same position. When predicting the current word, CBOW not only consider previous words, but also future words. Skip-gram model is similar to CBOW, but instead of predicting the current word based on the text, it tries to predict surrounding words given the current word.

Table 3.1: A set of concepts ($\mathcal{C}$) with their defining positive and negative context Ngrams, as well as the top-10 most similar search queries (using multiplicative cosine similarity on their embedding representations). Concepts $\mathcal{C}_1$ to $\mathcal{C}_5$ are based on Twitter content, whereas $\mathcal{C}_6$ 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
$\mathcal{C}_2$	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
$\mathcal{C}_3$	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
$\mathcal{C}_4$	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
$\mathcal{C}_5$	health	doctors, health, healthcare, nhs	cinema, football	vaccinations nhs, nhs dental, nhs sexual health, nhs nurses, nhs doctors, nhs appendicitis, nhs pneumonia, physiotherapy nhs
$\mathcal{C}_6$	gastro-intestinal 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
$\mathcal{C}_7$	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

3.3.2 Semantic Feature Selection

We consider a search query q to be a set of t textual tokens, $\{\varepsilon_1,\ldots,\varepsilon_t\}$, where standard English stop words are ignored.[4] The embedding of q, $\mathbf{e}_q$, is estimated by averaging across the embeddings of its tokens, that is

$$\mathbf{e}_q = \frac{1}{t}\sum_{i=1}^{t}\mathbf{e}_{\varepsilon_i}\,, \tag{3.10}$$

where $\mathbf{e}_{\varepsilon_i}$ denotes the word embedding of a search query token ε_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 of related or *positive* Ngrams, $\{P_1,\ldots,P_k\}$, and a set of non related or *negative* ones, $\{N_1,\ldots,N_z\}$. When the number of grams is bigger than 1, we retrieve the average embedding across the unigrams.

We then compute a similarity score, $S(q,\mathcal{C})$, between query embeddings and the formulated concept, using an extended version of the multiplicative cosine similarity (`3COSMUL`) introduced by Levy et al. (2014):

$$S(q,\mathcal{C}) = \frac{\prod_{i=1}^{k}\cos\left(\mathbf{e}_q,\mathbf{e}_{P_i}\right)}{\prod_{j=1}^{z}\cos\left(\mathbf{e}_q,\mathbf{e}_{N_j}\right)+\gamma}\,. \tag{3.11}$$

The numerator and denominator of Eq. (3.11) are products of cosine similarities between the embedding of the search query and each positive or negative concept term respectively. All cosine similarities (x) are transformed to the interval $[0,1]$ through $(x+1)/2$ 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. A multiplicative similarity is used as it is shown to be more balanced than an additive one, resulting in superior performance in various tasks (Levy et al., 2014, 2015). However, we note that the extension applied here (using more than 2 positive and 1 negative terms) has not received a dedicated evaluation in the literature, something that is hard given its unconstrained nature.

[4] We use a standard English language stop word list as defined in the NLTK software library (http://www.nltk.org).

Table 3.1 lists the concepts we formed and experimented with in our empirical analysis. 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 decided, perhaps with the assistance of an expert. Note that due to this reason, our approach is partially supervised. However, compared to building a classifier with a large number of manual labels (Paul and Dredze, 2014; Paul et al., 2014), the only manual effort is to define a few contextual keywords. We still consider this as an unsupervised learning approach. As described in the following sections, the optimal performance (in terms of MAE) is obtained when a broad version of this similarity based filter is combined with more traditional feature selection methods.

3.4 Case Study 1: Influenza-Like Illness Surveillance

To evaluate the effectiveness of the word embedding based semantic feature selection method, we apply it in the task of inferring ILI rates in England. We first assess the predictive capacity of the semantic feature selection method 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 correlation based feature selection to refer to this combination of bivariate linear correlation and elastic net regression. Finally, we propose a feature selection that combines the above two approaches, showcasing significant performance gains. The selected features from the various investigated feature selection approaches are also tested under the Gaussian Processes regressor described in Section 3.2.1.

3.4.1 Data Sets

Two data streams were used in our experiments: Google data and official health surveillance records obtained from PHE.

3.4.1.1 Google Data

The core input, user-generated data set for our supervised learning task was formed by time series of search query frequencies. It is a non standardized version of the publicly available Google Trends outputs and was retrieved through a Google Health Trends API, provided for academic research with a health-oriented focus. The query time series express the probability of a short search session for a specific geographical region and temporal resolution, drawn from a uniformly distributed 10%-15% sample of all corresponding sessions[5]. We have used a set of 35,572 non-preprocessed search queries (limited examples of which are presented in Table 3.1) and obtained their weekly frequency in England during an extensive period of 449 weeks, from January 1, 2007 to August 9, 2015.

To create word embeddings that capture more informal or direct ways of written expression, we used a Twitter data set. We collected tweets from users located in the UK[6]. The main incentive for this was to accommodate geographically constrained dialects and conversation themes. The total number of tweets was approx. 215 million, dated from February 1, 2014 to March 31, 2016. We applied the word2vec neural embedding algorithm (Mikolov et al., 2013a,c) as implemented in the gensim library[7]. 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 words with fewer than 500 occurrences, we obtained an embedding corpus of 137,421 unigrams. Note that we have not optimized word2vec's settings for our task, but the above parametrization falls within previously reported configurations (Amir et al., 2015). To capture more formal linguistic properties, we also used word embeddings trained on a Wikipedia corpus. The latter were based on the work of Levy and Goldberg (Levy and Goldberg, 2014) and have a dimensionality of 300.

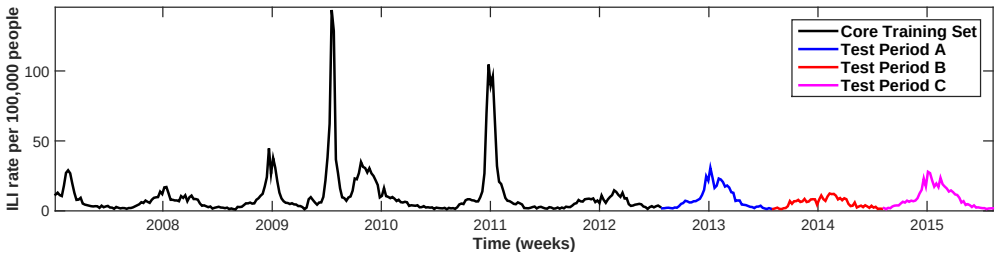


Figure 3.3: Weekly influenza-like illness (ILI) rates in England (per 100,000 people) from January 1, 2007 to August 9, 2015 obtained by RCGP and PHE. Training and test periods are denoted with different colorings.

3.4.1.2 Influenza-like Illness Surveillance Data

The inference target in our regression task consists of influenza-like illness (ILI) rates as reported by the RCGP and PHE. The 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 is displayed in Figure 3.3; different colorings denote training and testing periods.

3.4.2 Experiment Settings and Evaluation Metrics

We evaluate performance based on two metrics: Pearson correlation r, MAE between the inferred and target variables. We assess predictive performance on the last 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. 3.3). 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 $a = 0.5$, and decide the value of λ automatically by validating it on a held-out stratified subset ($\approx 7\%$) of the training set.

3.4.3 Semantic Feature Selection using Word Embeddings

The first row of Table 3.1 describes concept $\mathcal{C}_1$, which we refer to as "flu infection", that was chosen as the main concept for our experimental evaluation. The rationale 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,

[5] Note that the publicly available Google Trends represent a significantly smaller sample.
[6] The Twitter users are geographically distributed proportionally to regional UK population.
[7] **gensim library**, https://radimrehurek.com/gensims

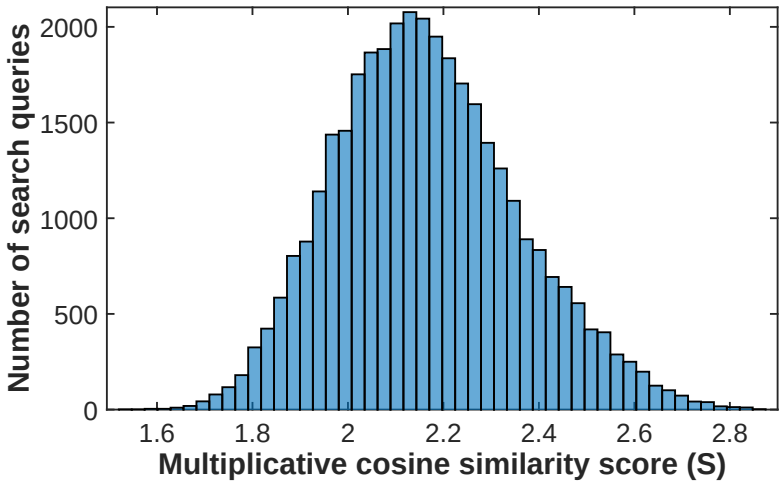


Figure 3.4: Histogram presenting the distribution of the search query multiplicative cosine similarity scores (S) with the flu infection concept $\mathcal{C}_1$.

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"[8] 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 3.4.5.

Figure 3.4 shows the distribution of the similarity scores (Eq. (3.11)) 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 3.2. 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[9] correlation of the data with the ground truth

[8] gp is an abbreviation for General Practitioner.
[9] Represents the mean frequency of all search queries.

Table 3.2: Linear regression (elastic net) performance estimates for the word embedding based feature selection. Column $S > \mu_S$ means we maintain queries with a similarity score that is greater than the mean similarity score. Columns $|\mathcal{Q}|$ and $|\mathcal{Q}_{\text{en}}|$ denote the average number of candidate and selected search queries in the three test periods, respectively. Columns r_{train} and r denotes the average correlation of the data with the ground truth in the training and test set, respectively. Column MAE is the mean absolute error between the inferred and target variables. NA (last row) denotes that no word embedding based feature selection has been applied.

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

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 illustrates 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 3.2) is equal to 32.33% (in terms of MAE). This indicates that selecting features via a semantically informed manner is better than solely relying on a naïve statistical approach. However, while the obtained performance is quite strong, the correlation based feature selection outperforms it, as we report in the next section.

3.4.4 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 (Guyon and Elisseeff, 2003). This is often applied prior to training a regression model, as

Table 3.3: Performance results for linear regression (elastic net) by applying a correlation based or a joint feature selection. Column corr > means we maintain queries with a similarity score that is greater than the values listed in the column. Columns $|\mathcal{Q}|$ and $|\mathcal{Q}_{\text{en}}|$ denote the average number of candidate and selected search queries in the three test periods, respectively. The subscript S means semantic filter is used. Column r denotes the average correlation of the data with the ground truth in the test set.

	Correlation only				**Correlation + word embeddings**			
corr>	$\lvert\mathcal{Q}\rvert$	$\lvert\mathcal{Q}_{\text{en}}\rvert$	**r**	**MAE**	$\lvert\mathcal{Q}^{\mathbf{S}}\rvert$	$\lvert\mathcal{Q}^{\mathbf{S}}_{\text{en}}\rvert$	**r**	**MAE**
.00	15,942	214	.560	5.864	2,275	168	.899	2.772
.10	3,238	128	.841	4.639	669	121	.918	2.206
.20	719	127	.811	3.861	256	53	.897	2.122
.30	279	121	.891	2.199	168	50	**.913**	**1.880**
.40	165	80	.876	2.137	118	43	.906	2.119
.50	104	65	.888	2.245	72	42	.905	2.347
.60	61	38	.850	2.577	40	18	.828	2.962
.70	26	9	.863	3.853	20	10	.863	3.855

a procedure that can reduce overfitting and offer performance gains (which we also report below). This form of feature selection has been applied in the task of ILI rate modeling from social media or search queries (Culotta, 2010; Ginsberg et al., 2009; Lampos et al., 2015). However, a correlation filter is not always successful in removing spurious features 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. 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 3.3 presents the performance outcomes under elastic net for correlation based and joint feature selection. The left part enumerates the results for a number of correlation thresholds (corr $> \rho$, $\rho \in [0,1)$), whereas on the right we report the

corresponding results using a combination of a correlation and similarity threshold (corr $> \rho \cap S > \mu_S + \sigma_S$, $\rho \in [0,1)$). Correlation based feature selection improves the performance estimates as opposed to using all the features (last row of Table 3.2), yielding its best performance, in terms of MAE, for corr $> .40$. This supports similar findings in the literature (Lampos et al., 2015). 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 3.3, where, based on the joint approach, queries that may be sufficiently correlated, but dissimilar to the specified concept are automatically omitted, we observe that the performance is enhanced significantly, reaching a relative improvement of 12.03% (from 2.137 to 1.880 in terms of MAE). As the correlation filter becomes more strict (corr$> .50$), 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.

Figure 3.5 compares the best-performing models, under elastic net, for the two approaches of performing feature selection (corr $> .40$ vs. corr $> .30 \cap S > \mu_S + \sigma_S$). It is evident that the correlation based approach makes some odd inferences at certain points in time, whereas the joint 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 joint feature selection model. As we reduce the correlation threshold, such problems are amplified and less relevant search queries are embedded into the model, expressing seasonality and other confounders.

To evaluate the proposed feature selection approach with the nonlinear Gaussian Processes regression model, we focus on the linear regression setups (correlation based or joint 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

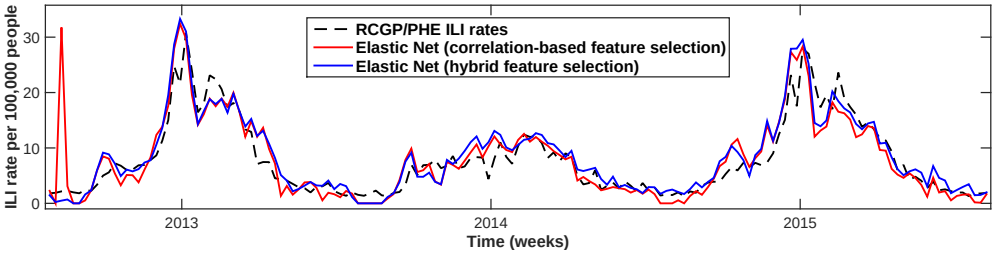


Figure 3.5: Comparative plot of the optimal models for the correlation based and joint feature selection under elastic net for the estimation of ILI rates in England.

Table 3.4: Nonlinear regression (Gaussian Processes) performance estimates. Column corr $>$ means we maintain queries with a similarity score that is greater than the values listed in the column. Column $S > \mu_S + \sigma_S$ means we maintain queries with a similarity score that is greater than one standard deviation from the mean similarity score. Check marks indicate the applied feature selection method(s). Columns *r* denotes the average correlation of the data with the ground truth in the test set. Their application sequence follows the left to right direction of the table columns.

corr $>$	$\mathbf{S > \mu_S + \sigma_S}$	**Elastic Net**	**r**	**MAE**
.10	-	✓	.568	5.344
	✓	✓	.912	2.057
.20	-	✓	.814	4.015
	✓	✓	.920	1.892
.30	-	-	.857	2.858
	-	✓	.891	2.686
	✓	-	**.942**	**1.567**
	✓	✓	.928	1.696
.40	-	-	.864	2.475
	-	✓	.895	2.347
	✓	-	.913	2.110
	✓	✓	.934	2.030
.50	-	-	.887	2.197
	-	✓	.921	2.308
	✓	-	.908	2.267
	✓	✓	.926	2.292
.60	-	-	.819	2.742
	-	✓	.851	2.598
	✓	-	.865	2.614
	✓	✓	.831	2.880

are enumerated in Table 3.4 and point again to the conclusion that the joint feature selection yields the best performance. In terms of MAE, this amounts to an improvement of 28.7% against the best nonlinear correlation based performance outcome, and a 16.6% against the best linear model. Interestingly, when word em-

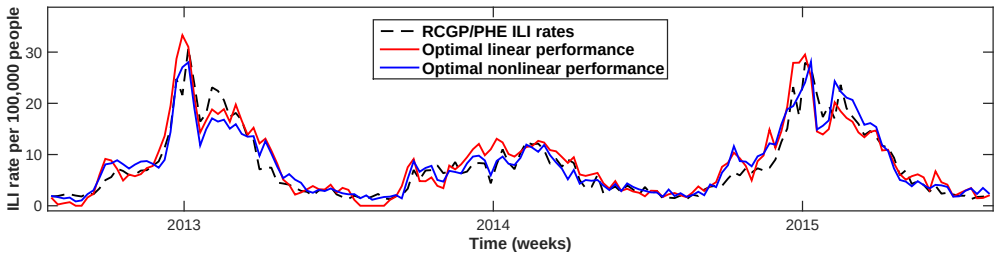


Figure 3.6: Comparative plot between the optimal nonlinear and linear models (both using joint feature selection) for the estimation of ILI rates in England.

bedding feature selection is not applied, the nonlinear model can seldom exceed the performance of the corresponding linear model, providing an indirect indication of the inappropriateness of the selected features.

Figure 3.6 draws a comparison between the inferences of the best nonlinear and linear models, both of which happen to use the same feature basis ($\text{corr} > .30 \cap S > \mu_S + \sigma_S$). The Gaussian Processes model provides more smooth estimates and an overall better balance between stronger and milder flu seasons.

3.4.5 How are Inferences Affected by the Choice of a Different Concept

The main human intervention[10] 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 experimented with a number of different concepts (see Table 3.1). $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ are variations of the flu infection topic, $\mathcal{C}_4$ is on infectious diseases in general, $\mathcal{C}_5$ is about health, 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 3.5 enumerates the best obtained performance (under elastic net) for all investigated concepts for variants of the joint 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, and when the focus is drawn on a different disease (gastroin-

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

Table 3.5: Optimal performance estimates after applying the feature selection method based on correlation and word embeddings for varying concepts under elastic net. Column $\mathbf{S} > \mu_{\mathbf{S}} + \sigma_{\mathbf{S}}$ means we maintain queries with a similarity score that is greater than one standard deviation from the mean similarity score. Check marks indicate the applied feature selection method(s). Column $\cap\, \mathcal{C}_1$ measures the average percentage of common features with the ones formed by using $\mathcal{C}_1$. Column corr $>$ means we maintain queries with a similarity score that is greater than the values listed in the column. Columns *r* denotes the average correlation of the data with the ground truth in the test set.

ID	$\mathbf{S} > \mu_{\mathbf{S}} + \sigma_{\mathbf{S}}$	corr $>$	$\cap\, \mathcal{C}_1(\%)$	**r**	**MAE**
$\mathcal{C}_1$	✓	.30	100%	.913	1.880
$\mathcal{C}_2$	✓	.30	98.6%	.914	1.864
$\mathcal{C}_3$	✓	.30	98.4%	**.913**	**1.788**
$\mathcal{C}_4$	✓	.30	87.5%	.920	2.084
$\mathcal{C}_5$	✓	.30	43.1%	.891	2.237
$\mathcal{C}_6$	✓	.20	8.3%	.616	5.217
$\mathcal{C}_7$	✓	.30	94.2%	.909	2.116

testinal; $\mathcal{C}_6$), 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 ($\sim 98\%$) 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 was indicated by Table 3.1), which end up providing inferior performance to the ones trained on Twitter.

3.5 Case Study 2: Infectious Intestinal Diseases Surveillance

A case study on ILI inference has shown the effectiveness of our proposed semantic feature selection method based on word embeddings. In this section, we utilize this method on another task, inferring IID rates in England. To the best of our knowledge, this is the first work that models IIDs from user-generated content. IIDs have a number of characteristics that are distinct from diseases that have been previously investigated using user-generated content, such as influenza we studied in the previous case study (Culotta, 2010; Lampos and Cristianini, 2010; Ginsberg et al., 2009; Lampos et al., 2015). 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 (Bernardo et al., 2013; Tam, 2012; Wheeler, 1999).
- Self-diagnosis in user-generated content (e.g. as in "I am down with the flu") is less frequent, resulting in sparser textual feature representations; for example, a feature as informative as the keyword 'flu' does not exist.
- IIDs generally exhibit a less stronger seasonality than other infectious diseases.

In this study, we directly apply word embedding based semantic feature selection to choose features that are semantically relevant to the IIDs topic. Similar to Section 3.4, together with semantic feature selection, we also apply a correlation based feature selection method. Then, we apply a regularized linear (elastic net), as well as a nonlinear (Gaussian Processes), regression function for inference.

3.5.1 Datesets

Two data streams are used in our experiments: Twitter data and official health surveillance records obtained from PHE.

3.5.1.1 Twitter data

Tweets were retrieved using the Twitter API. Approximately 585 million tweets geolocated in England over a 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 approximately 410 million tweets.

3.5.1.2 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 (1) campylobacter and (2)

norovirus (the most recurrent organisms related to IIDs according to PHE reports). We also consider (3) 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).

3.5.1.3 Extracting Features from Tweets

To create vector space representations of the Twitter corpus, we first extract all n-grams ($1 \leq n \leq 3$) from the Twitter dataset; to form an n-gram, we filter out a list of common English stop words,[11] 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 semantic feature selection method, we form a vocabulary $\mathcal{S}_{\text{IID}}$ of 597 1-grams that have the highest multiplicative cosine similarity with the predefined IID topic. The IID topic includes positive keywords ("vomit", "indigestion", "heartburn", "nausea", "reflux", "diarrhea", "hiccups") and negative words ("flu", "cold"). We use a tighter – and more semantically coherent – set of the top 212 1-grams, to perform keyword matching with 2- and 3-grams.[12] 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 normalized using the total number of tweets published in a week.

3.5.2 Experiment Settings and Evaluation Metrics

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 modeling food poisoning cases, given the smaller time span of the data. When applying elastic net, we use the same values for the regularization parameters (λ_1, λ_2) in all folds, and in each fold's training set we pre-filter features by applying a Pearson correlation threshold

[11] The applied list of English stop words was a concatenation of various lists available online.
[12] Both cutoff thresholds (597 and 212 top terms) have been decided through manual inspection.

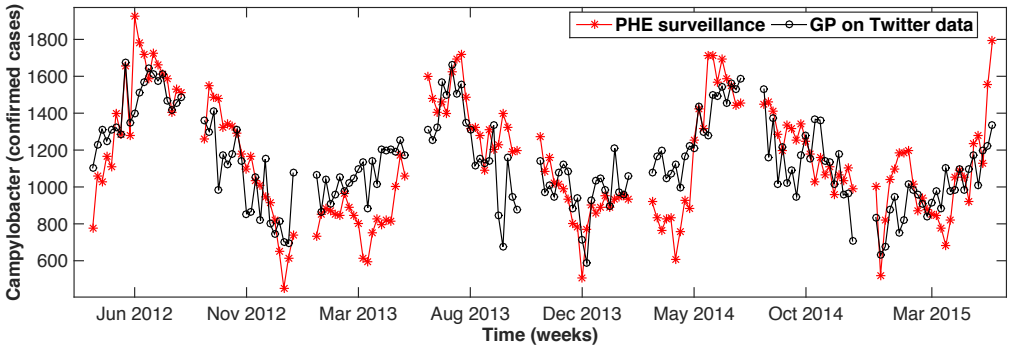


Figure 3.7: Comparative plot between laboratory confirmed campylobacter cases in England (reported by PHE) and the indication inferred from Twitter content based on the Gaussian Processes model. The gaps separate the folds in the 8-fold cross validation process.

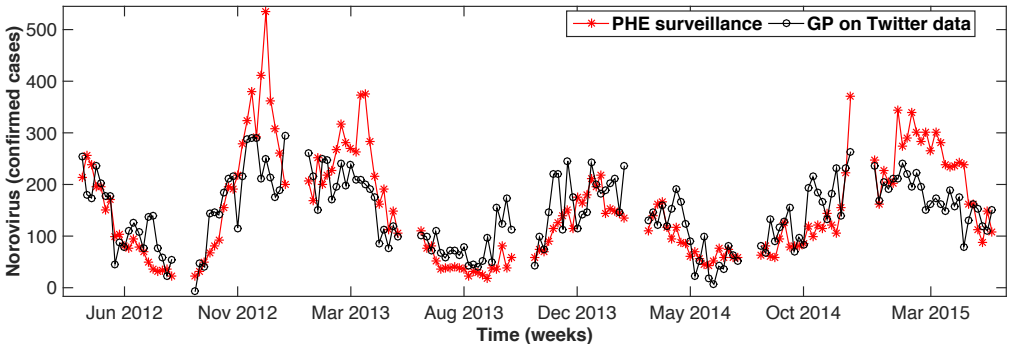


Figure 3.8: Comparative plot between laboratory confirmed norovirus cases in England (reported by PHE) and the indication inferred from Twitter content based on the Gaussian Processes model. The gaps separate the folds in the 8-fold cross validation process.

with the corresponding ground truth. The Gaussian Processes model is applied on the positively weighted features selected by the elastic net (per fold). We use MAE and Pearson correlation (r) to measure the performance of the models; note that $\mathbf{y}$ has been standardized (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 bigger than its mean value, to assess the performance of the models during periods of increased incidence of an IID.

3.5.3 Results

Table 3.6 enumerates the results for the two methods. The Gaussian Processes method outperforms Elastic Net; the difference in their mean performance (using MAE) is statistically significant according to a Kolmogorov-Smirnov test (Massey

Table 3.6: Performance indicators for the IID indicator inference task from Twitter content in England.

	Elastic Net		Gaussian Processes	
IID target	**r**	**MAE**	**r**	**MAE**
Campylobacter	.625	.572	**.633**	**.545**
Norovirus	.596	.554	**.607**	**.513**
Food poisoning	.702	.700	**.711**	**.624**

and Frank (1951); $p < .05$).[13] For campylobacter, norovirus and food poisoning, the average MAE between inferences and standardized target values is .545, .513 and .624, whereas their linear correlation is .633, .607 and .711, according to the better-performing Gaussian Processes model. Figures 3.7, 3.8, and 3.9 present the Gaussian Processes inferences in all the folds for the three case studies.

[13]Given the small sample for food poisoning, we could not assess its statistical significance.

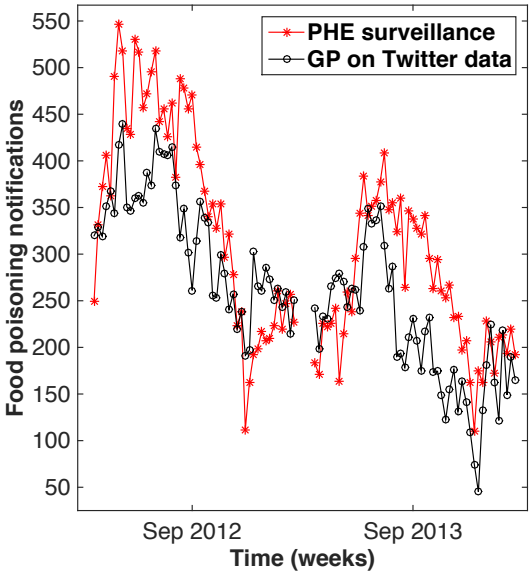


Figure 3.9: Comparative plot between food poisoning cases in England (reported by PHE) and the indication inferred from Twitter content based on the Gaussian Processes model. The gaps separate the folds in the 2-fold cross validation process.

Table 3.7: Comparison of the inference performance (average MAE), when the IID activity is above its mean value.

IID target	**Elastic Net**	**Gaussian Processes**
Campylobacter	.623	**.562**
Norovirus	.790	**.732**
Food poisoning	.927	**.802**

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 Gaussian Processes model. Looking at the average peak-MAE performance figures (Table 3.7), we see that the performance gap between elastic net and Gaussian Processes models increases, emphasizing the value of a nonlinear approach when the IID signal gains a significant presence.

3.6 Summary

In this chapter, we have presented a feature selection method for text regression that employs neural word embeddings to improve the topicality of the selected features. Our approach can be seen as an unsupervised filter for a target thematic concept that can be easily applied in conjunction with current feature selection techniques. Following the feature selection, a regularized linear (elastic net) and a nonlinear (Gaussian Processes) regression model are deployed as inference functions. We conducted two experiments to demonstrate the effectiveness of our proposed method, inferring ILI and IID rates in England from Web data. We have shown that the proposed feature selection method can significantly outperform competitive approaches. Future work will focus on further generalizations of the reported outcomes, including different application domains and more detailed qualitative interpretations.

Chapter 4

Multi-Task Learning for Disease Surveillance

Existing algorithms for disease surveillance from online user-generated content are predominantly based on supervised learning paradigms (Ginsberg et al., 2009; Culotta, 2010; Lampos and Cristianini, 2010; Paul et al., 2014). These frameworks propose single task learning solutions that do not consider the correlations of data across different geographies. They are also not considering situations, where significantly fewer health reports are available for training a model.

In this chapter, we investigate the utility of multi-task learning for disease surveillance using Web search data. Multi-task learning can train a number of disease models jointly. Compared to single task learning, it has the potential to improve the generalization of a model by exploiting shared structures in the data. Previous work has shown that this may result in significant performance gains (Caruana, 1997; Baxter, 2000; Bakker and Heskes, 2003; Ben-David and Schuller, 2003; Evgeniou and Pontil, 2004; Argyriou et al., 2006). In the context of disease surveillance, we investigate whether multi-task learning can provide an improved estimate of disease rates when (1) training data is available for multiple geographic locations, specifically geographic regions of the US, and (2) when ground truth training data (health reports) is sporadic. In addition, we investigate the utility of multi-task learning to estimate disease rates in a different country by exploiting a denser health reporting scheme of a reference country. We explore both linear and nonlinear re-

gression models, namely multi-task elastic net (Lee et al., 2010) and Multi-task Gaussian Processes (Bonilla et al., 2007), comparing them to their respective single task formulations.

We use ILI as a case study and conduct experiments on the US and England. Our experiments show that multi-task learning models improve regional as well as national ILI rates estimates from Google search data for the US. The percentage improvement increases as the historical training data is reduced, up to 14.8%, indicating that multi-task learning can facilitate the derivation of accurate models using significantly less training data. We also simulate situations, where partial ground truth data are available, perhaps due to unexpected reasons (natural disasters, a spreading epidemic, technical problems) or due to limitations of a public health system. Our experimental results indicate that multi-task learning models can mitigate such effects. Finally, we expand this concept to cross-country settings, where complete data for a country could improve the models of another country with insufficient health reports. In that case, multi-task learning is shown to improve ILI estimates for England (up to 40% of error decrease) under the assumption that increasingly limited historical data exist, when training models jointly with data from the US.

This chapter's main contributions are the following:

- This is the first work to assess the utility of multi-task learning in infectious disease surveillance from Web search data.
- We show that multi-task learning models improve:
  - regional as well as national ILI models for the US,
  - regional US models for ILI, under the assumption of increasingly limited historical health reports (simulated by using three different sampling methods), and
  - country-level ILI models for England, when training is performed jointly with data from a different, but not culturally distant, country (the US).

The rest of this chapter is structured as follows. We first review related work in Section 4.1. Then we provide a description for the disease surveillance task, under both single and multi-task learning settings in Section 4.2. We present the linear and nonlinear techniques for performing single and multi-task regression in Section 4.3 and 4.4, respectively. A case study on influenza-like illness surveillance is presented in Section 4.5. Finally, we make a summary in Section 4.6.

4.1 Related Work

The fundamentals of multi-task learning have been thoroughly presented in (Caruana, 1997). Compared to single task learning that attempts training on isolated tasks, multi-task learning performs this jointly using a shared representation. The tasks can be used as valuable sources of inductive bias for each other, leading to a more accurate model (Caruana, 1997). This may also allow more difficult problems, such as target variables with partial observations, to be modeled successfully (Caruana, 1997; Bakker and Heskes, 2003; Ben-David and Schuller, 2003). The majority of multi-task regression models were developed by extending their single-task formulations. Some examples for linear regression are the multi-task ℓ_1-norm regularization (Argyriou et al., 2008) and the $\ell_{2,1}$-norm regularization (Liu et al., 2009). Nonlinear multi-task regression models have also been explored, extending Support Vector Machines (Evgeniou and Pontil, 2004), Gaussian Processes (Bonilla et al., 2007), Convolutional or Recurrent Neural Networks (Abdulnabi et al., 2015; Liu et al., 2016a).

In this work, we study the utility of multi-task learning in disease surveillance from Web search data. Existing approaches have routinely used single task models such as regularized regression (Polgreen et al., 2008; Ginsberg et al., 2009; Culotta, 2010; Lampos et al., 2015), Gaussian Processes (Lampos et al., 2015, 2017), and autoregressive frameworks (Shaman and Karspeck, 2012; Paul et al., 2014; Lampos et al., 2015). Here, we have chosen to apply multi-task elastic net (Lee et al., 2010) and multi-task Gaussian Processes (Bonilla et al., 2007) for the following reasons: (a) elastic net and Gaussian Processes have been applied in many text re-

gression (Lampos et al., 2014; Preoţiuc-Pietro et al., 2015) and disease modeling approaches (Lampos et al., 2015; Zou et al., 2016; Lampos et al., 2017), and (b) the sample sizes we are operating on are limited and no performance gain would have been achieved by deploying neural network structures (Collobert and Weston, 2008; Zhang et al., 2014b, 2015).

Multi-task learning has been applied in the context of user-generated data modeling (Lampos et al., 2013; Lukasik et al., 2015) and computational health (Zhou et al., 2012; Benton et al., 2017; Bickel et al., 2008; Emrani et al., 2017; Zhao et al., 2015). Given various tasks and objectives, multi-task learning frameworks can be different. Zhou et al. (2012) and Emrani et al. (2017) formulated a fused sparse group lasso and a graph regularization approach, respectively, aiming to model disease progression. Both models focused on the temporal relation between the various tasks and utilized image data from patients. However, our work focuses on textual user-generated content and the spatial relation among tasks. Benton et al. (2017) used online multimodal user-generated content to train a multi-task feed-forward neural network for classifying the mental health condition of online users. This model tries to capture shared structures of user attributes in relation to mental conditions. Our work, however, focuses on a collective regression task, aiming to exploit relationships at a higher level, determined by geography, rather than specific user characteristics. Finally, Zhao et al. (2015) proposed a linear regularized multi-task regression model to detect civil unrest events in various locations using Twitter data. In our work, apart from a different thematic focus, we also deploy nonlinear multi-task learning frameworks.

4.2 Problem Formulation

Our aim is to infer disease rates as reported by an established health surveillance system using the frequencies of Web search queries. We formulate this as a regression task, where we learn a function f: $\mathbf{X} \rightarrow \mathbf{y}$ that maps the input space $\mathbf{X} \in \mathbb{R}^{n \times p}$ to the target variable $\mathbf{y} \in \mathbb{R}^n$; n denotes the number of samples and p is the size of our feature space, i.e. the number of unique search queries we consider. $\mathbf{X}$ contains

time series of normalized frequencies of search queries and $\mathbf{y}$ represents the disease rates at the same time points as reported by the health agency. A normalized query frequency is defined as the count of a query divided by the total number of searches during a fixed time interval, e.g. one week.

In multi-task disease rate inference, we are modeling disease rates simultaneously for a number of different geographical locations (tasks). A tensor $\mathbf{Q} \in \mathbb{R}^{n \times p \times m}$ is used to represent our input data for the m tasks.[1] $\mathbf{Q}$ can simply be interpreted as m versions of $\mathbf{X}$; in the remainder of the script, we denote them using $\mathbf{Q}_j$, where j refers to the j^{th} task or geographical location. An element of $\mathbf{Q}$, $\mathbf{Q}_{tij}$, represents the normalized frequency of a query i for the location j during the time interval t. The corresponding target variables, i.e. the disease rates for the m locations are denoted by $\mathbf{Y} \in \mathbb{R}^{n \times m}$. Similarly, we use $\mathbf{Y}_j$ to refer to the disease rates at the location j. Based on the aforementioned formulations, our task now becomes to learn a function f, such that f: $\mathbf{Q} \rightarrow \mathbf{Y}$.

4.3 Multi-Task Elastic Net (MTEN)

Linear regression models have been successfully applied for conducting disease surveillance from web search and social media data (Ginsberg et al., 2009; Culotta, 2010; Lampos and Cristianini, 2010; Paul et al., 2014; Zou et al., 2016). We use multi-task elastic net (Zou and Hastie, 2005) to train several linear regression models jointly.

Multi-task elastic net is an extension to elastic net described in Section 2.4.3.4 (Zhao et al., 2015). It is specified by the following optimization task

$$\underset{\mathbf{W},\beta}{\operatorname{argmin}} \left(\|\mathbf{Y} - \mathbf{QW} - \beta\|_F^2 + \lambda_1 \|\mathbf{W}\|_{2,1} + \lambda_2 \|\mathbf{W}\|_F^2 \right), \tag{4.1}$$

where $\mathbf{W} \in \mathbb{R}^{p \times m}$, $\beta \in \mathbb{R}^m$ are the weight matrix and intercept vector for all the m

[1] Note that the number of samples n may be different for different locations (tasks).

tasks, and the norms $\ell_{2,1}$ and Frobenius (F) are given by

$$\begin{aligned} \|\mathbf{W}\|_{2,1} &= \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{m} W_{ij}^2} \\ \|\mathbf{W}\|_{F} &= \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{m} W_{ij}^2}. \end{aligned} \tag{4.2}$$

The $\ell_{2,1}$ norm encourages all tasks to select a common set of features, while the Frobenius norm enhances the robustness of the model (Zhao et al., 2015).

4.4 Multi-Task Gaussian Processes (MTGP)

We also deploy nonlinear regression models using Gaussian Processes as previous works have shown that the relationship between query frequencies and disease rates is significantly better captured by a nonlinear function (Lampos et al., 2017; Wagner et al., 2017).

Gaussian Processes models were extended to a multi-task version (MTGP) by Bonilla et al. (2007) and have been used in various tasks, including natural language processing applications (Cohn and Specia, 2013; Beck et al., 2014). The MTGP model incorporates all m tasks into a single GP that is defined by

$$f(\mathbf{Q}) \sim \mathcal{GP}\left(\mu_{\mathrm{M}}(\mathbf{x}), k_{\mathrm{M}}(\mathbf{x}, \mathbf{x}')\right), \tag{4.3}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are inputs from tasks j and j', respectively. As with the single-task Gaussian Processes, we assume $\mu_{\mathrm{M}}(\mathbf{x}) = 0$. MTGP's covariance function, $k_{\mathrm{M}}(\mathbf{x}, \mathbf{x}')$, is formed by placing a Gaussian Processes prior over the kernel function in Eq. (3.4), so that we directly induce correlations between the tasks (Bonilla et al., 2007). It is given by

$$k_{\mathrm{M}}(\mathbf{x}, \mathbf{x}') = k^{\mathrm{c}}(j, j') \times k^{\mathrm{x}}(\mathbf{x}, \mathbf{x}'), \tag{4.4}$$

where k^{c} is a correlation kernel that explains the relation between tasks j and j', and k^{x} is the covariance that explains the relation of inputs $\mathbf{x}$ and $\mathbf{x}'$. This approach is also known as the intrinsic correlation model (Wackernagel, 2014).

Let $\mathbf{K}_{\mathrm{M}}$ be the covariance matrix of $\mathbf{Q}$, $\mathbf{K}^{\mathrm{c}}$ the task correlation matrix, and $\mathbf{K}^{\mathrm{x}}$ the covariance matrix of inputs. We define $\mathbf{K}_{\mathrm{M}}$ as

$$\mathbf{K}_{\mathrm{M}} = \mathbf{K}^{\mathrm{c}} \otimes \mathbf{K}^{\mathrm{x}}, \tag{4.5}$$

where $\otimes$ denotes a Kronecker product. $\mathbf{K}^{\mathrm{c}}$ is assumed to be a valid covariance matrix (satisfying Mercer's theorem). Its diagonal elements describe the correlation of the tasks with themselves and the non-diagonal elements correspond to the correlation between tasks. It can be constructed using the Cholesky decomposition and is parameterized by the elements of the lower triangular matrix of

$$\mathbf{K}^{\mathrm{c}}(j,j') = \mathbf{J}\mathbf{J}^{\top},\ \mathbf{J} = \begin{pmatrix} \theta^{\mathrm{c}}_{1} & 0 & \dots & 0 \\ \theta^{\mathrm{c}}_{2} & \theta^{\mathrm{c}}_{3} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta^{\mathrm{c}}_{\zeta-m+1} & \theta^{\mathrm{c}}_{\zeta-m+2} & \dots & \theta^{\mathrm{c}}_{\zeta} \end{pmatrix}, \tag{4.6}$$

where $\theta^{\mathrm{c}} = \{\theta^{\mathrm{c}}_{u}\}$, $u \in \{1,2,\dots,\zeta\}$ is the set of $\mathbf{K}^{\mathrm{c}}$'s hyperparameters, with $\zeta = m(m+1)/2$.

Inference and hyperparameter learning in MTGPs is conducted similarly to the single task Gaussian Processes (Bonilla et al., 2007; Durichen et al., 2014). Given a new data point $\mathbf{x}_*$, for task j, the predictions (y_*) can be made by using the conditional distribution $p(\mathbf{y}_*|\mathbf{x}_*,\mathbf{Q},\mathbf{Y}) \sim \mathcal{N}(\mu_{j*},\sigma^2_{j*})$, where

$$\mu_{j*} = \left(\mathbf{k}^{\mathrm{c}}_{j} \otimes \mathbf{k}^{\mathrm{x}}_{*}\right)^{\top} \mathbf{K}^{-1}_{\mathrm{M}}\mathbf{Y}, \text{ and} \tag{4.7}$$

$$\sigma^2_{j*} = \mathbf{K}_{\mathrm{M}} + \mathbf{D} \otimes \mathbf{I}. \tag{4.8}$$

In the above equations, $\mathbf{k}^{c}_{j}$ is the j^{th} column of $\mathbf{K}^{\mathrm{c}}$, $\mathbf{k}^{x}_{*}$ is the vector of covariances between $\mathbf{x}_*$ and the training points, and $\mathbf{D}$ is an $m \times m$ matrix in which the $(j,j)^{\text{th}}$ element is the noise variance (σ^2_j) for the j^{th} task.

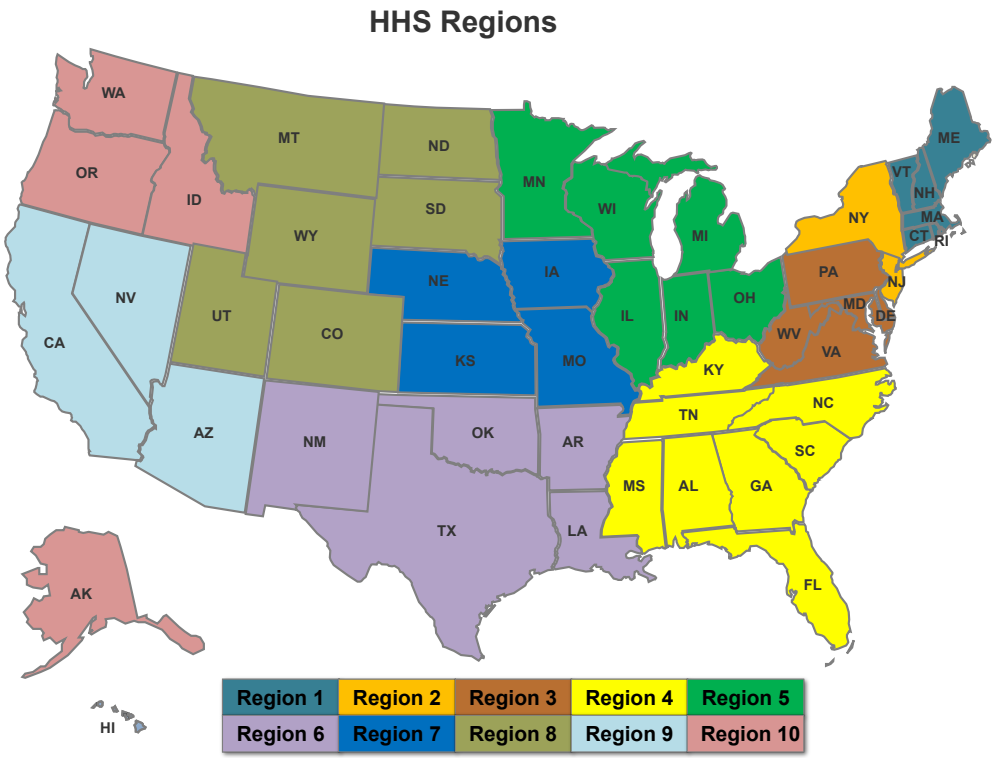


Figure 4.1: The 10 US regions as specified by the Department of Health & Human Services (HHS).

4.5 Case Study on Influenza-like Illness Surveillance

Our experiments assess a number of different disease modeling scenarios, where we expect that multi-task learning will have a positive impact. We focus on the estimation of ILI rates, which is a well-studied task (Ginsberg et al., 2009; Polgreen et al., 2008; Lampos et al., 2015; Yang et al., 2015). The locations of interest are the US at the national level, US regions as defined by HHS, and England.

4.5.1 Data Sets and Experiment Settings

4.5.1.1 ILI Rates from Health Agencies

For the US, we use weekly ILI rates from CDC. These rates represent the average percentage of all outpatient visits to health care providers normalized by the respective regional population figures and are recorded by ILINet. The 10 HHS US regions considered by the CDC are shown in Fig. 4.1. Our data spans from September 1, 2007 to August 31, 2016 (both inclusive), which includes 9 consecutive influenza seasons as defined by the CDC. Each (expanded) flu season begins

on September 1 and ends on August 31 of the next year. To provide further insight, we have plotted the ILI rates of US regions 1, 2, and the US as a whole in Fig. 4.2. As expected, we see that the time series are strongly correlated, but each signal may be peaking at different moments throughout a flu season. For England, we obtain weekly ILI rates from PHE through the syndromic surveillance network developed by the RCGP. We focus on the same time period as for the US.

4.5.1.2 Search Query Frequencies

We iteratively used Google Correlate starting with flu-related query seeds (such as the word 'flu') to obtain a set of 1,641 candidate search queries. However, due to the existing seasonal confounders, many of the candidate queries we ended up with, such as 'college basketball' or 'spring break', were not related to flu. To remove these unrelated queries in a principled fashion, we applied a topic filter specified using word embeddings. The filtering process was similar to the one we proposed in (Lampos et al., 2017), but without the notion of a negative context. Embeddings were trained using word2vec on Google news (Mikolov et al., 2013c,a).[2] We consider a query q to be a set of z textual tokens, $\{\varepsilon_1, \ldots, \varepsilon_z\}$. The embedding of q, $\mathbf{e}_q$, is computed by averaging across the embeddings of its tokens,

$$\mathbf{e}_q = \frac{1}{z} \sum_{i=1}^{z} \mathbf{e}_{\varepsilon_i} \,. \tag{4.9}$$

We define a topic about flu, $\mathcal{T}$, as a set of two flu-related terms, specifically the name of the disease and one of its main symptoms, $\mathcal{T} = \{\text{'flu'}, \text{'fever'}\}$. For each of the queries, we calculate a similarity score defined as the product of the cosine similarities between the embeddings of the terms in $\mathcal{T}$ and $\mathbf{e}_q$, i.e.

$$S(\mathbf{q}, \mathcal{T}) = \prod_{i=1}^{2} \cos(\mathbf{e_q}, \mathbf{e}_{\mathbf{T}_i}) \,, \tag{4.10}$$

[2]The embeddings were downloaded from code.google.com/archive/p/word2vec. The specific training settings are detailed in (Mikolov et al., 2013c,a).

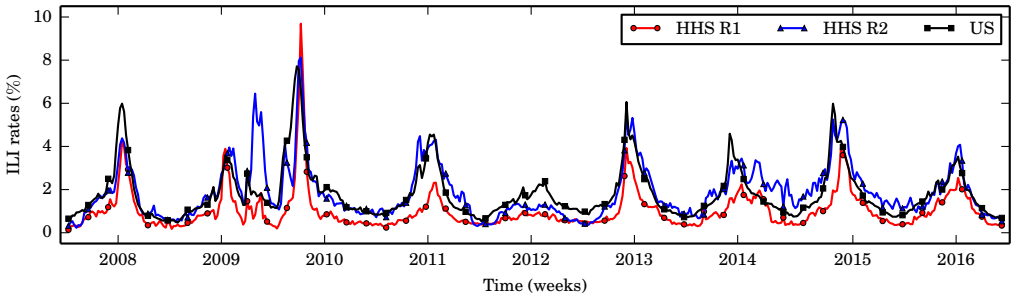


Figure 4.2: Weekly ILI rates (from CDC) for the US (national level) as well as the US Regions 1 and 2.

where each cosine similarity component is mapped to $[0,1]$ via $(\cos(\cdot,\cdot)+1)$ / 2.[3] Queries with $S \leq 0.5$ are filtered out and are not considered in our experiments. The 0.5 threshold guarantees that even in the extreme case, where a candidate query has a perfect cosine similarity (equal to 1) with one of the two concept queries, it also needs to have a non-negative cosine similarity (prior to the $[0,1]$ mapping) with the other concept query. The semantic filter succeeds in eliminating some confounding features, i.e. queries that may be highly correlated with ILI rates, but are referring to different topics.[4]

We retain 128 search queries after applying the word embedding filter described above.[5] The frequencies of these queries are retrieved through a private Google Health Trends API, provided for academic research with a health-oriented focus. The query frequency expresses the probability of a short search session[6] conducted within a geographic region and during a specified time period. The probability is estimated based on a 10-15% sample of all Google searches. We obtained daily frequencies at the state-level (for the US) and the national-level (for the US and England) from September 1, 2007 to August 31, 2016 (both inclusive). Weekly frequencies were estimated by averaging the daily frequencies. Similarly, regional US frequencies were computed by averaging the state-level frequencies.

[3] This resolves misleading similarity scores based on different sign combinations.

[4] All candidate queries together with their similarity scores are listed at https://github.com/binzou-ucl/google-flu-mtl.

[5] For the experiments on England, two queries referring to medication available in the US are replaced by England-based equivalent medication.

[6] 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.

4.5.1.3 Baselines, Evaluation and Parameter Learning

To demonstrate the effectiveness of multi-task learning models, we compare MTEN and MTGP with their single-task formulations, EN and GP, respectively. We use Pearson correlation (r) and the MAE between inferred and target ILI rates as our evaluation metrics. For reporting the performance of multi-task learning models, we use the average MAE and correlation of the different test periods across all tasks (locations). The statistical significance of a performance improvement is tested via a paired-sample t-test by using the mean MAEs across all locations for the applied test periods (for the two methods under comparison). In our results, we use an asterisk (*) to indicate that a difference in performance is not statistically significant at the .05 level ($p\text{-}value \geq .05$). For learning the regularization parameters of the linear models, we perform grid search on 20% of the training data; all models are trained on the remaining 80% subset of the training data. We begin by training a model on data from the first ϕ flu seasons, and test the model in the following season ($\phi+1$). Then, we increase our training data by including one more flu season ($\phi+1$) and test in the following season ($\phi+2$); we repeat this process until we have tested on the last flu season in our data set. Before training a model, we only retain search queries that have a Pearson correlation higher than .3 with the respective disease rates (per location). This correlation threshold choice was motivated by the extensive experiments we conducted in (Lampos et al., 2017) (see Table 3 in that paper). Note that the correlation filter is applied to each training data set separately and it may result in retaining different features for each task. Whenever this is the case, we maintain the intersection of features among the tasks. In addition, the GP and MTGP models are trained on the features that received a nonzero weight by the respective elastic net model, similarly to the methodology proposed in (Lampos et al., 2015).

4.5.2 Regional and National ILI Surveillance Tasks

First, we investigate whether multi-task learning can improve the accuracy of regional US models for the estimation of ILI rates. We test this hypothesis under a decreasing number of training samples, where L varies from 5 to 1 year(s) of his-

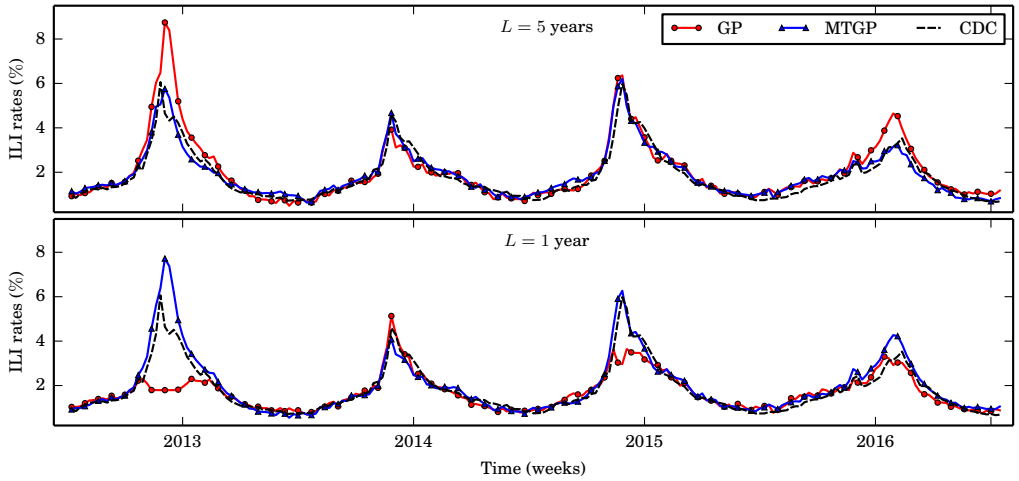


Figure 4.3: Comparing GP (red) and MTGP (blue) ILI estimates for the US using $L = 5$ years and $L = 1$ year of training data.

torical data. By doing this we can additionally assess whether multi-task learning models can have a positive impact when the historical training data are limited. The multi-task learning models are trained on data from the 10 US HSS regions jointly and their performance is compared to the performance obtained by learning these models separately.

Table 4.1 enumerates the performance for the aforementioned comparison.[7] We observe that, in general, multi-task learning models perform better than their single-task alternatives both in terms of MAE and correlation. In addition, the non-linear models tend to outperform the linear ones. However, performance gains from multi-task learning (in MAE) only become statistically significant when $L \leq 2$ years of historical training data are used. The greatest improvement occurs for $L = 1$; for this case MTEN reduces EN's MAE by 7.5%, whereas the MTGP reduces GP's MAE by 12.7%.

We next expand our observations by adding data for the US at a national level. Hence, we are now considering 11 tasks (US plus the 10 US regions). The aim is to test whether we can obtain a better model at the national level by training it together with regional data in a multi-task learning fashion. The results enumerated

[7]Numbers in the table represent the average performance across the 10 US regions and the 4 test periods. For additional clarity, all individual performance estimates (for $L = 1$ and $L = 5$) are enumerated at https://github.com/binzou-ucl/google-flu-mtl.

Table 4.1: Performance of single and multi-task learning models for estimating ILI rates on US HHS regions. Numbers in the table represent the average performance across the 10 US regions and the 4 test periods. L denotes the length of the training period in years. The asterisk (*) indicates that a multi-task learning model does **not** yield a statistically significant improvement over its single-task formulation.

	EN		**MTEN**		**GP**		**MTGP**	
L	r	MAE	r	MAE	r	MAE	r	MAE
5	.928	.347	.935	.344*	.936	.335	.944	.330*
4	.919	.379	.927	.371*	.926	.355	.938	.346*
3	.912	.398	.921	.385*	.916	.382	.929	.369*
2	.901	.438	.913	.414	.906	.424	.924	.398
1	.845	.531	.858	.491	.844	.535	.867	.467

in Table 4.2 confirm that this is the case. The impact of multi-task learning is greater and statistically significant (in terms of MAE), when $L \leq 3$ years. The greatest improvement happens for $L = 1$; for this case MTEN reduces EN's MAE by 12.6%, whereas the MTGP reduces GP's MAE by 14.8%. In Fig. 4.3, we compare the estimates from the GP and MTGP models for the ILI rates in the US during the test periods from 2012 to 2016 (4 flu seasons) under two different training data lengths (5 vs. 1 year of historical data) and against the rates reported by CDC. Even under the 5-year training period, where the difference in average performance between the models is small, we see that the GP makes a significant over-prediction of the peak during the 2012/13 flu season, something that the MTGP does not. The bottom sub-figure, where $L = 1$ year, showcases more clearly the level of improvement

Table 4.2: Performance of single and multi-task learning (including regional data) models for estimating US ILI rates. Numbers in the table represent the performance at the national level across 4 test periods; notational conventions as in Table 4.1. The asterisk (*) indicates that a multi-task learning model does **not** yield a statistically significant improvement over its single-task formulation.

	EN		**MTEN**		**GP**		**MTGP**	
L	r	MAE	r	MAE	r	MAE	r	MAE
5	.960	.353	.962*	.351*	.965	.253	.966*	.245*
4	.951	.356	.954*	.353*	.947	.265	.949*	.251*
3	.939	.398	.945	.374	.942	.286	.947*	.268
2	.930	.408	.936	.362	.933	.351	.941	.323
1	.854	.531	.868	.464	.854	.513	.875	.437

obtained by applying a multi-task learning scheme; MTGP delivers a quite accurate model despite being trained on a few samples. This is an important characteristic as it suggests that we can develop accurate disease prevalence models with much less historical data than previously considered (Ginsberg et al., 2009; Lampos and Cristianini, 2012; Lampos et al., 2015).

4.5.3 Mitigating the Effect of Sporadic ILI Health Reports

In many real-world scenarios, health surveillance reports are or can become temporally and/or geographically sporadic. For instance, syndromic surveillance networks, especially in developing countries, may focus on a few regions rather than an entire country due to infrastructure and economic constraints. Furthermore, established health surveillance schemes may be exposed to data loss due to unprecedented events, such as technical faults, natural disasters or a spreading epidemic during which doctor visits are discouraged. In the following experiments, we assess whether multi-task learning can help us establish more accurate disease models under various scenarios of sporadic health reporting. To assess this, we have performed several forms of down-sampling on the training data of several US HHS regions. All experiments were conducted by setting $L = 1$, i.e. based on 1-year long training periods, and results represent the average performance after 50 sampling trials.

We have applied the following sampling techniques: (**A**) *random weekly sampling*, (**B**) *random monthly sampling*, and (**C**) *random burst-error sampling*. In (A), we simply take random samples from our data, thereby simulating scenarios where reports for a specific week may be missing. In (B), we first partition our data into non-overlapping monthly periods and then randomly sample over these periods, thereby simulating situations where health systems may be affected for longer time periods. Finally, in (C) we randomly discard a block of temporally contiguous data points, and use the remaining points only. We apply a sampling rate $\gamma = \{0.1, 0.2, \ldots, 1\}$, where $\gamma = 1$ means that all data are used (no sampling), and $\gamma = 0.1$ that 10% of the weekly data (for A) or monthly periods (for B) are maintained. In C, γ determines the size of the error block B, $B = (1 - \gamma)\tau$, where

Table 4.3: Performance of single and multi-task learning models for estimating ILI rates on US HHS regions belonging to $\mathcal{R}$-odd under three sampling methods (A, B and C). Training data in $\mathcal{R}$-odd regions is down-sampled using a sampling rate (γ). The asterisk (*) indicates that a multi-task learning model does **not** yield a statistically significant improvement over its single-task formulation.

		EN		**MTEN**		**GP**		**MTGP**	
	γ	r	MAE	r	MAE	r	MAE	r	MAE
	1.0	.825	.492	.843	.488*	.828	.502	.856	.460
A	0.9	.823	.504	.840	.494*	.825	.503	.852	.465
	0.8	.806	.512	.839	.498*	.817	.505	.850	.465
	0.7	.805	.523	.834	.499*	.811	.506	.849	.467
	0.6	.800	.528	.824	.501*	.804	.512	.835	.468
	0.5	.798	.541	.823	.502*	.804	.513	.835	.469
	0.4	.789	.550	.822	.508	.801	.534	.829	.469
	0.3	.768	.555	.817	.511	.801	.545	.825	.474
	0.2	.758	.567	.803	.520	.789	.564	.824	.476
	0.1	.698	.694	.793	.554	.700	.686	.824	.482
B	0.9	.813	.516	.835	.495*	.814	.519	.851	.463
	0.8	.806	.531	.827	.505*	.805	.528	.843	.468
	0.7	.793	.549	.823	.511*	.792	.540	.834	.475
	0.6	.775	.555	.821	.516	.776	.565	.825	.476
	0.5	.752	.574	.820	.523	.756	.570	.823	.478
	0.4	.702	.598	.818	.534	.751	.594	.819	.485
	0.3	.621	.751	.815	.544	.650	.748	.817	.491
	0.2	.510	.781	.814	.547	.516	.776	.814	.497
	0.1	.425	.942	.806	.583	.433	.930	.809	.503
C	0.9	.817	.524	.836	.497*	.818	.525	.848	.466
	0.8	.805	.539	.829	.506*	.810	.532	.839	.470
	0.7	.796	.554	.817	.513	.801	.552	.832	.471
	0.6	.784	.576	.814	.528	.788	.569	.825	.473
	0.5	.756	.606	.807	.535	.766	.588	.819	.477
	0.4	.689	.637	.799	.543	.713	.626	.818	.480
	0.3	.621	.739	.794	.557	.632	.711	.804	.492
	0.2	.483	.792	.781	.561	.506	.791	.800	.498
	0.1	.414	.934	.780	.571	.424	.906	.796	.505

τ is equal to the size of the training data. In all experiments, we are sampling per location, meaning that the time points in the training data can vary across locations.[8]

We begin by assessing the added value of multi-task learning in situations,

[8]We have also conducted experiments where sampling is temporally synchronized across regions, but we did not observe a significant difference in the performance outcomes. Due to space constraints, we only report the non-synchronized results.

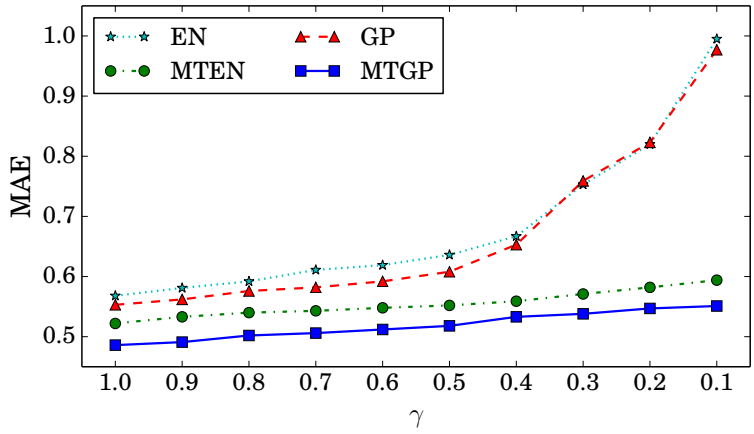


Figure 4.4: Comparing the performance of EN (dotted), GP (dashed), MTEN (dash dot) and MTGP (solid) on estimating the ILI rates for US HHS Regions (except Regions 4 and 9) for varying burst error sampling (type C) rates (γ).

where progressively less health reports are obtained for half of the regions of a country. To simulate this, we partition the 10 US HSS regions into two sub-groups, $\mathcal{R}$-odd and $\mathcal{R}$-even consisting of the odd and even regions respectively (following the numbering of Fig. 4.1). For the regions in $\mathcal{R}$-odd, we have increasingly down-sampled their training data; regions in $\mathcal{R}$-even were not subject to down-sampling.

Table 4.3 enumerates the results of this experiment. The numbers in the table represent the average MAE of all test periods over the $\mathcal{R}$-odd regions. Generally, the performance of the multi-task learning models degrades less as down-sampling increases, i.e. there are less training data. MTGP always offers a statistically significant improvement over GP, whereas MTEN, in the worst case (for sampling type A), requires a $\gamma \leq 0.4$ to achieve this. Type A sampling, which can be seen as having missing weekly reports in various regions at random time points, affects single task learning models much more than multi-task learning models. For example, for the EN model, the MAE increased from .492 for $\gamma = 1$ (no down-sampling), to .694 for $\gamma = 0.1$, a degradation of 41.1%. In contrast, the MTEN model degrades by 13.5%. The effect is more pronounced for the nonlinear models, with GP degrading by 36.7% while MTGP degrades by only 4.8%. Note that MTGP's MAE is equal to .482 when the fewest data points are used (10% for $\gamma = 0.1$), which is smaller than EN's or GP's MAEs, when no sampling is taking place (.492 and .502 respectively).

All models degrade worse for B and C sampling methods, which drop blocks

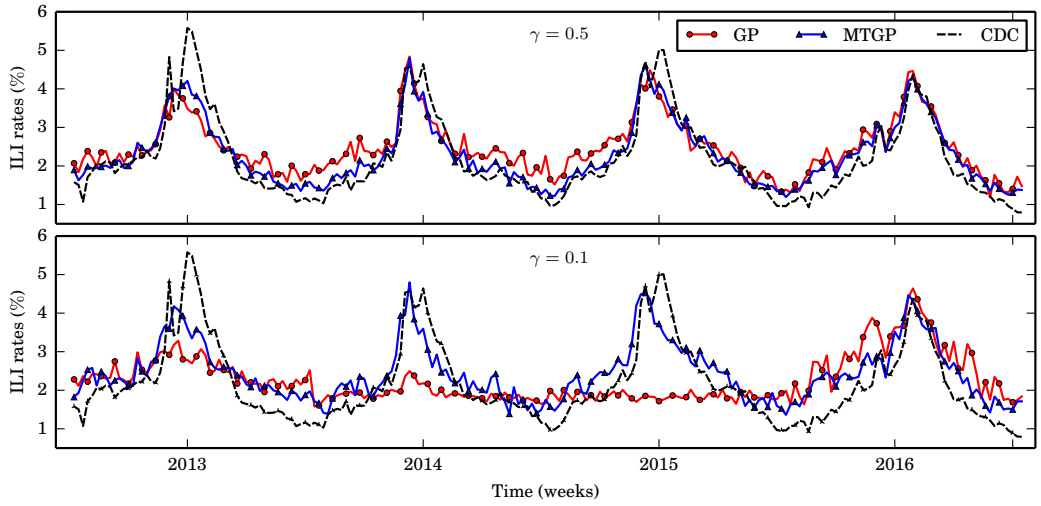


Figure 4.5: Comparing GP (red) and MTGP (blue) ILI estimates for US Region 9 for two burst error sampling (type C) rates (γ).

of data points from the training set. However, the degradation in performance of the multi-task learning models is much less than for the comparative EN or GP models. For example, when $\gamma = 0.1$, MTGP improves GP's MAE by 45.9% and 44.3% for B and C sampling types, respectively. Fig. 4.5 illustrates this performance difference by comparing the ILI estimates from the GP and MTGP models for US region 9 under burst error sampling, for $\gamma = 0.5$ (top) and $\gamma = 0.1$ (bottom).[9] Clearly, for low sampling rates ($\gamma = 0.1$) the MTGP model is still able to provide acceptable performance.

In a subsequent experiment, we performed burst-error sampling on all but two US regions with the highest population figures (Regions 4 and 9). The rationale behind this setting is that in many occasions health reports are available for central locations in a country (i.e. two big cities), but are limited anywhere else. Fig. 4.4 compares the performance of all regression models under this scenario. It confirms that the pattern observed in the previous experiment still holds, i.e. that the multi-task models are much less affected by down-sampling. We can also see that MAE in single task learning models increases at an exponential rate as γ decreases.

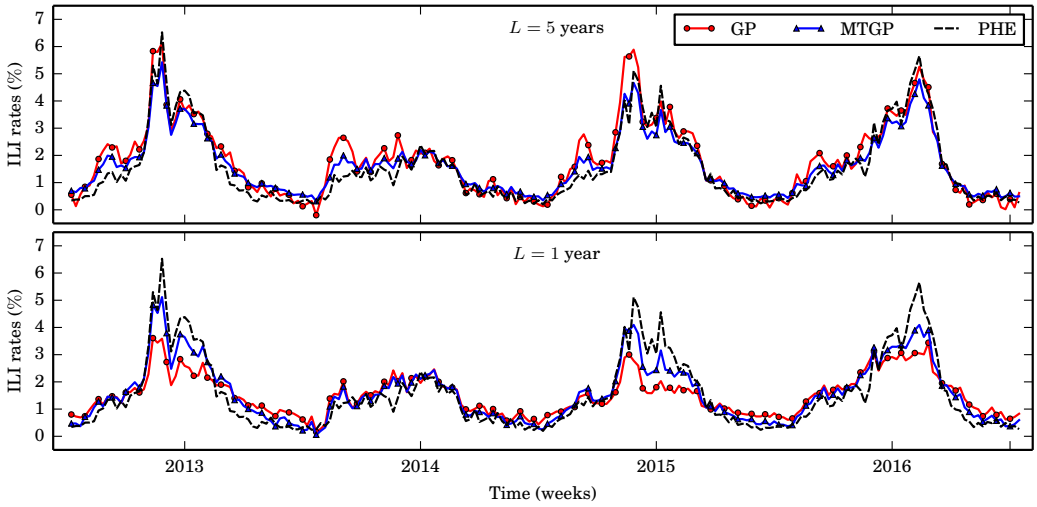


Figure 4.6: Comparing GP (red) and MTGP (blue) ILI estimates for England under varying training data sizes.

4.5.4 ILI Surveillance Tasks across Countries

We expand on the previous results to test whether a stable data stream for a country could be used to enhance a disease model for a different, but culturally similar, country. The underlying assumption here is that countries that share a common language and have cultural similarities may also share common patterns of user search behavior.

For this purpose, we use data from the US and England and assume that there are increasingly fewer historical health reports for England only, in a similar fashion to the experiments described in Section 4.5.2 (L from 5 to 1 year). For the US data, we always assume that the training window is based on the past $L = 5$ years. The search queries used in both countries are the same, with the following exception. Two of the US search queries about medication were changed to their British equivalent because their search frequencies in England are low; we changed "tussin" to "robitussin" and "z pak" to "azithromycin".

Table 4.4 shows a similar pattern of results to the previous experiments. All multi-task learning models register statistically significant improvements compared to the single task learning ones. As the length of the training period is reduced,

[9]Region 9 includes the states of California, Nevada and Arizona and one of the largest in terms of population ($\approx$ 49.1 million).

the improvements are greater; MTGP reduces MAE by 20.9% and 40.0% for $L = 5$ and $L = 1$ year, respectively. Fig. 4.6 presents the estimates for the GP and MTGP models for these extreme cases. Whereas both models seem to be inferring the trends of the time series correctly, the multi-task estimates are closer to the actual values of the signal's peaks.

The results confirm our original hypothesis that data from one country could improve a disease model for another country with similar characteristics. This motivates the development of more advanced transfer learning schemes (Pan and Yang, 2010), capable of operating between countries with different languages by overcoming language barrier problems, using variants of machine translation.

4.6 Summary

In this chapter, we have investigated the utility of multi-task learning to disease surveillance from Web search data. Disease surveillance models for various geographies — inside a country and across different countries – were trained jointly such that knowledge between different tasks could be shared. We explored both linear and nonlinear models (MTEN and MTGP) and used ILI surveillance as a case study. Experiments were conducted on the US and England. Our empirical results indicate that multi-task learning improves regional as well as national models for the US. The percentage of improvement increases as we reduce the historical training data. For a 1-year training period, the MTGP model improved MAE by 14.8% at the

Table 4.4: Performance of single and multi-task learning models for estimating ILI rates in England; notational conventions as in Table 4.1. The asterisk (*) indicates that a multi-task learning model does **not** yield a statistically significant improvement over its single-task formulation.

	EN		**MTEN**		**GP**		**MTGP**	
L	r	MAE	r	MAE	r	MAE	r	MAE
5	.885	.696	.896	.491	.891	.599	.903	.474
4	.873	.734	.887	.504	.880	.664	.894	.491
3	.860	.788	.876	.530	.868	.742	.883	.517
2	.854	.842	.871	.554	.859	.815	.875	.528
1	.836	.999	.857	.603	.846	.977	.860	.586

regional level. Furthermore, in simulated scenarios, where health reports (training data) are limited, we showed that multi-task learning helps to maintain stable inference performance across all the affected locations. Experiments, where data for England were modeled in conjunction with US data, indicated that more accurate estimates were obtained for England, maxed at 40% MAE reduction when using 1-year long training periods. This suggests that multi-task learning can benefit models across different countries as well.

Chapter 5

Transfer Learning for Disease Surveillance

Recent research efforts have shown that traditional disease surveillance can be complemented by alternative methods trained on data from online user activity, e.g. social media or online search behavior (Milinovich et al., 2014; Gomide et al., 2011; Choudhury et al., 2013; Lampos and Cristianini, 2010; Culotta, 2010; Paul and Dredze, 2011; Polgreen et al., 2008; Ginsberg et al., 2009; Lampos et al., 2015; Yang et al., 2015; Biggerstaff et al., 2018; Wagner et al., 2018). The main advantages of these complementary methods are timeliness, and sampling from a larger (and perhaps different) segment of the population, including people who may not visit a doctor while being ill. It is also often cited that such approaches may be very useful in regions where health infrastructure is poor or absent. However, in practice, this is often impossible as the proposed machine learning solutions rely on training data which, apart from the user-generated inputs, need to contain confirmed disease rates at the target location, broadly referred to as "ground truth". This data is typically provided by existing syndromic surveillance systems. Hence, for locations where ground truth is not available, user-data driven approaches are not realistically applicable.

In this chapter, we propose a statistical framework to circumvent problems associated with a lack of training data in some geographic regions. Our approach is based on the broad notion of transfer learning, where we aim to transfer parts

of the knowledge gained while solving a certain task to better solve a different, but related one (Pan and Yang, 2010). In particular, our goal is to transfer a well-performing disease rate inference model from a source location, where supervised learning is possible, to a target location, where supervision is not possible, given the lack of ground truth. We focus our experiments on influenza (flu) and utilize Google search query statistics as our descriptive variable for aggregate, population-level, online user activity. For example, CDC monitor and report ILI rates on a weekly basis, providing sufficient ground truth to learn a function that maps online search query frequencies to these rates. In our experiments we show that we can adapt this function to derive estimates of ILI rates at different locations (outside the US). Language may or may not differ between the source and target locations. Online search statistics can be obtained for these target locations, but we assume that there is no ground truth data.

The proposed approach is composed of 3 steps. After learning a source regression model (step 1), we seek ways to map the selected source search queries to sets of queries in the target location. To derive this mapping we deploy a hybrid metric, which combines semantic similarity with a time series correlation component (step 2). Semantic similarities are estimated using cross-lingual or monolingual word embeddings and correlations are computed using query frequencies. Finally, query weights from the source model are transferred to the identified target queries (step 3). This framework is evaluated on three transfer learning tasks, where the source model is always based in the US, and the target countries are France, Spain and Australia. While ground truth is available for all the target countries, we only use it to evaluate the performance of the transferred models. Transferred models, assessed on four flu seasons (2012 to 2016), can accurately estimate the peak of each flu season, achieving, on average, Pearson correlations greater than .92 and root mean squared errors comparable to the ones obtained by the corresponding fully supervised models ($\leq 21.6\%$ increase in errors). Therefore, they can be considered as practical solutions for locations that lack historical ground truth data.

The chapter's main contributions are the following:

- We propose a novel, end-to-end transfer learning framework for mapping a disease model trained on online search data from a location, where ground truth is available, to a location, where ground truth is not available.

- We investigate variations of this model, exploring different query mapping functions using semantic or temporal similarities or combinations of the two.

- We empirically show that our approach works in three case studies, two of which require a transfer to a different language (English to French or Spanish), and one that maintains the same language (English), but demands a model transfer to a different hemisphere (US to Australia).

5.1 Problem Formulation

The estimation of disease rates from web search data is commonly formulated as a regression task (Ginsberg et al., 2009; Lampos et al., 2015). The aim is to learn a function $f\colon \mathbf{X} \rightarrow \mathbf{y}$ that maps the input space of search query frequencies, $\mathbf{X} \in \mathbb{R}^{n \times s}$, to the target variable, $\mathbf{y} \in \mathbb{R}^{n}$, representing disease rates; n denotes the number of samples and s is the size of the feature space, i.e. the number of unique search queries we are considering. More specifically, $\mathbf{X}$ contains the time series of search query frequencies, and $\mathbf{y}$ represents the number of disease diagnoses per 100,000 people (as reported by a health agency) at corresponding times. The time interval for computing the frequency of queries is often set to one week to match the temporal frequency of syndromic surveillance reports.

Regression approaches require observations of the target variable $\mathbf{y}$ (ground truth) for training a machine learning model. This restricts the application of such techniques to areas where historical disease rates are available. This chapter attempts to address this limitation by proposing a transfer learning methodology, that maps an existing disease model from an area, where disease rates are available, to another location, where disease rates cannot be obtained. In this scenario, we seek to map an established ILI rate estimation function $f\colon \mathbf{X} \rightarrow \mathbf{y}$ from a source country, where ground truth exists, to a target country, where ground truth is unavailable.

We define the source domain as $\mathcal{D}_S = \{(\mathbf{x}_i, y_i)\}$, $i \in \{1,\ldots,n\}$, where $\mathbf{x}_i$ is an s-dimensional vector holding the frequencies of the s queries for the time interval i, y_i is the corresponding disease rate, and n is the number of observations. The target domain is denoted by $\mathcal{D}_T = \{\mathbf{x}'_i\}$, $i \in \{1,\ldots,m\}$, where $\mathbf{x}'_i$ is a t-dimensional vector of the frequencies of the t queries in the target domain that are associated with the s queries in the source domain. No ground truth is available for the target domain, although in practice we have access to ground truth which we later on use only to evaluate a transferred model. Note that t need not equal s, thus allowing one-to-many query mappings, which are discussed shortly. In theory, the m time intervals may precede or overlap the n time intervals in the source region. In our experiments, results are reported only for the case where the m target intervals are after the n source intervals.

5.2 Related Work

The fundamental properties of transfer learning have been thoroughly discussed in relevant literature (Ben-David et al., 2007; Mansour et al., 2009; Pan and Yang, 2010; Torrey and Shavlik, 2009; Ben-David et al., 2010; Weiss et al., 2016). In contrast to traditional machine learning methods, which assume that the training and test data belong to the same domain, i.e. they are drawn from the same feature space and distribution, transfer learning aims to improve the learning function in a target domain by transferring knowledge from a related, source domain. This concept has been successfully applied to various tasks, including text classification (Dai et al., 2007; Pan et al., 2009; Glorot et al., 2011; Chen et al., 2011), part of speech tagging (Blitzer et al., 2006; Jiang and Zhai, 2007), machine translation (Koehn and Schroeder, 2007; Foster et al., 2010), and image classification (Kulis et al., 2011; Zhu et al., 2011; Duan et al., 2012).

In this work, we present a statistical framework for transferring a disease surveillance model from a source country, where supervised learning is applicable, to a target country, where no ground truth is available. We formulate it as a cross-lingual transductive regression task (Pan and Yang, 2010), which poses the

following challenges: (a) ground truth is not available in the target domain, and (b) features (queries) may not belong in the same feature space due to linguistic or cultural differences. Due to (a), multi-task learning models, such as this solution for ILI (Zou et al., 2018), cannot be used because they still require partial ground truth to capture the relationship between the different tasks (Caruana, 1997). To solve (b), a few studies have attempted to learn a mapping of both source and target languages to the same space (Wan, 2009; Prettenhofer and Stein, 2010; Huang et al., 2013; Smith et al., 2016). For example, Prettenhofer and Stein (2010) used unlabeled documents along with a word translation oracle to automatically induce task-specific, cross-lingual correspondences for cross-lingual text classification. In this chapter, we used cross-lingual word embeddings proposed in (Smith et al., 2016) to align different languages. Methods have also been proposed for reducing the distance between the source and target features (Pan et al., 2009; Zhou et al., 2014). For example, Pan et al. (2009) proposed Transfer Component Analysis (TCA) to learn transfer components across source and target domains in a reproducing kernel Hilbert space using maximum mean discrepancy. Zhou et al. (2014) constructed a sparse feature transformation matrix based on compressive sensing theory to map the weight vector of classifiers learned from the source domain to the target domain. However, their tasks are very different from the regression task studied in this chapter. These models were not able to capture efficiently the time series structure in our data.

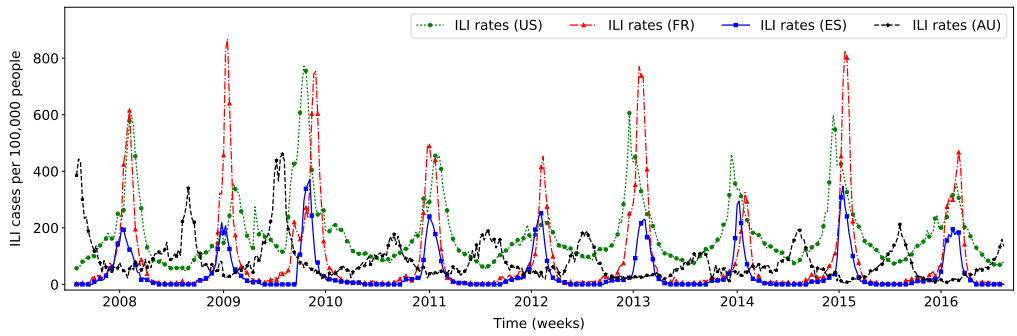
5.3 Data Sets

Our experiments rely on two sources of data, namely Google search query frequency statistics and ILI rates from established health organizations.

5.3.1 Google Search Query Frequency Statistics

Time series of weekly search query frequencies were retrieved through Google Correlate. A frequency represents the weekly search activity of a query (number of times issued) within a geographical region. It is normalized by dividing by the total number of search queries issued during that week. This normalization controls for

Figure 5.1: ILI rates for the United States (US), France (FR), Spain (ES) and Australia (AU).



variations in the number of searches issued each week. Variations in the weekly number of searches can be due to a variety of causes, including (1) summer vacations, (2) responses to news events, and (3) the general trend of increased Web usage (Mohebbi et al., 2011). Normalized query frequencies are subsequently standardized, such that their time series have a zero mean and a standard deviation of one. This results in expressing query frequencies under the same units, indicating how many standard deviations a query frequency value is away from its mean value, for different geographical regions with potentially varying population size and search usage patterns.

Overall, we obtained weekly frequencies of search queries from September 1, 2007 to August 31, 2016 inclusive (470 weeks) for US, France, Spain, and Australia. Given that an exhaustive list of user search queries was not available to us, we extracted search queries by using a set of 12 flu-related queries per country as a seed to Google Correlate and iterating through this process (using correlated queries as new seeds). This process extracted 34,121, 29,996, 15,673 and 8,764 queries for US, France, Spain and Australia, respectively. Queries were not limited to the topic of flu, given that various other spurious queries may also correlate (different illnesses, activities or products based on seasonality).

5.3.2 Influenza-Like Illness Rates

We obtained weekly ILI rates for US, France, Spain and Australia from their established syndromic surveillance systems, namely CDC, SN, SISSS, and ASPREN, respectively. ILI rates represent the fraction of the population that has been di-

agnosed with influenza-like symptoms. The data spans from September 1, 2007 to August 31, 2016 inclusive, which covers approximately 9 consecutive influenza seasons. Note that for Spain, we only have ILI rates from Week 40 in a year to Week 20 in the following year. The prevalence of influenza outside this period is typically very low. We denote ILI rates from each syndromic surveillance system using the corresponding country code (US, FR, ES, and AU). All ILI rates in this chapter represent the number of ILI cases per 100,000 people in a population.

In our experiments, described in the following sections, we are transferring a flu model trained on US data (search query frequencies, ILI rates) to one of the other three countries. To provide some insight about the difficulty of the task, we have plotted the historical ILI rates for all countries in Fig. 5.1. ILI rates may correlate between countries, e.g. the Pearson correlation between the US and FR rates is equal to .6 ($p < 0.01$), but peaks and troughs are occurring at different times and with very different intensities. The US and AU ILI rates are negatively correlated ($-.4$, $p < 0.01$), as expected, since these countries are situated in different hemispheres and influenza is strongly seasonal.

5.4 Transfer Learning Framework

In this section, we first describe our fundamental assumption in Section 5.4.1. Then we provide an overview of the transfer learning framework in Section 5.4.2. Each step of the transfer learning framework is presented in Section 5.4.3, 5.4.4, and 5.4.5.

5.4.1 A Fundamental Assumption about Online Search Behavior in Different Countries

As the transfer learning function is explained in more detail in the following paragraphs, it will become apparent that the proposed statistical framework is grounded on a fundamental assumption, which is that online search behavior will be similar in the source and the target countries. Narrowing this assumption down for our specific task, this implies that the conditional probability of issuing a query q under a certain health status h (with or without experiencing influenza-like symptoms),

Table 5.1: Mean ratio of query frequency over ILI rate in United States, France, Spain, and Australia.

Search queries	US	FR	ES	AU
flu (US/AU), grippe (FR), gripe (ES)	.036	.033	.032	.031
symptoms of flu (US/AU), symptômes de la grippe (FR), síntomas de gripe (ES)	.030	.031	.029	.027
flu in children (US/AU), grippe chez le bébé (FR), gripe en el bebé (ES)	.017	.020	.019	.022

$P(q|h)$, will be similar for the populations of the source and the target countries. Relevant literature offers some evidence of this with regards to user search behavior for various health-related themes (Andreassen et al., 2007; Ybarra and Suman, 2008; Barry et al., 2011; Alicino et al., 2015). In addition, we also offer some empirical evidence using our data. Table 5.1 shows the average query frequency over the corresponding ILI rate ratio for three basic queries in the US and AU. It also shows these ratios for translations of these queries in FR and ES (e.g. flu $\rightarrow$ grippe (FR) $\rightarrow$ gripe (ES)). The main observation is that these ratios do not vary much over the time span of our data, which is almost a decade. Although, this is a limiting observation, in that it does not involve many different search queries, it serves as a strong indication that user search behavior, at least for this specific area of interest, has similarities among different countries. The transfer learning framework, described in the following paragraphs, tries to exploit these similarities.

5.4.2 Overview of the Transfer Learning Framework

Here we provide an overview of the proposed transfer learning framework. It consists of three basic steps:

- **Step 1** learns a regression function based on data from the source domain. That is, given the frequencies of source queries together with estimates of the disease rate, we learn a linear regression function comprised of a set of s non-negative weights (one weight per query).

- **Step 2** maps the s source queries to the t target queries. As previously mentioned, this need not be a unique mapping, i.e. a source query, q_s may map to 0, 1 or more queries in the target domain, and two or more queries in the source domain may share queries in the target domain.

- **Step 3** transfers the regression weight associated with each source query to the corresponding queries in the target domain.

5.4.3 Step 1 — Learning a Regression Function in the Source Domain

Regularized regression has been successfully applied to various text regression tasks, including the estimation of disease rates from social media or online search data (Lampos et al., 2015; Zou et al., 2016). In this chapter, we use elastic net (Zou and Hastie, 2005) as our regression function (see Section 2.4 for detail). Given $\mathbf{X} \in \mathbb{R}^{n \times s}$ and $\mathbf{y} \in \mathbb{R}^n$ from the source domain $\mathcal{D}_\mathrm{S}$, we apply a constrained version of elastic net which solves the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{w}}{\operatorname{argmin}} \left(\|\mathbf{y} - \mathbf{X}\mathbf{w} - \beta\|_2^2 + \lambda_1 \|\mathbf{w}\|_2^2 + \lambda_2 \|\mathbf{w}\|_1 \right), \\ &\text{subject to } \mathbf{w} \geq 0, \end{aligned} \tag{5.1}$$

where $\lambda_1 > 0$, $\lambda_2 > 0$ are, respectively, the ℓ_1-norm and ℓ_2-norm regularization parameters, and β denotes the intercept term. The non-negativity constraint for $\mathbf{w}$ may result in a worse performing model for the source country, but, at the same time, makes the weight transfer from a source to a target country more comprehensive (positive weights are easier to interpret) and eventually more accurate in terms of performance (see Section 5.5.4).

Due to the seasonal nature of influenza, our dataset of candidate queries contains a significant number of confounders, i.e. queries that are correlated with flu, but have no link to flu, such as "college basketball" and "spring break". To remove these unrelated queries we applied a semantic filter based on word embedding representations, similar to the one proposed in (Lampos et al., 2017; Zou et al., 2018). English word embeddings were trained on the English Wikipedia corpus using the `fastText` method (Bojanowski et al., 2017). A topic about flu, $\mathcal{T}$, was defined as a simple set of two flu-related terms, $\mathcal{T} = \{\text{'flu'}, \text{'fever'}\}$. For each of the source queries, we calculate a similarity score defined as the product of the cosine similar-

ities between the embeddings of the terms in $\mathcal{T}$ and $\mathbf{e}_q$, i.e.

$$g(\mathbf{q},\mathcal{T}) = \prod_{i=1}^{2} \cos(\mathbf{e_q}, \mathbf{e}_{\mathbf{T}_i}), \tag{5.2}$$

where each cosine similarity component is mapped to $[0,1]$ via $(\cos(\cdot,\cdot)+1)/2$.[1] Queries with $g \leq .5$ are filtered out and are not considered in our experiments.

5.4.4 Step 2 — Mapping Source to Target Queries

The identified and weighted set of search queries in the source domain ($\mathcal{Q}_\mathrm{S}$) should be mapped to a set of queries in the target domain from a potential pool of target query candidates ($\mathcal{P}_\mathrm{T}$). Queries about the same topic may vary in their textual formulation, especially when they are issued by users located in different countries. Even in cases, where countries share the same language, cultural and socioeconomic differences may result in different querying preferences. Thus, simple approaches where search queries from the source country are translated or directly mapped to queries in the target country are not efficient. In our approach, we utilize word embeddings (mono- or cross-lingual) to map source to target queries based on their broad semantic relationship. We consider both one-to-one and one-to-many query mappings from the source to the target domain.

In addition, the weights associated with each source query reflect on how correlated the query is with the modeled disease rate. Therefore, a desired property is to map source queries to target ones based on their pairwise temporal correlation as this may enhance the statistical relevance of the mapping. Consequently, there is a tension between mapping based on semantic similarity and mapping based on the similarity in temporal correlation. To capture both, we define a combined similarity metric, Θ, that is the weighted sum of a semantic similarity Θ_s and a correlation similarity, Θ_c, i.e.

$$\Theta = \gamma\Theta_\mathrm{s} + (1-\gamma)\Theta_\mathrm{c}, \tag{5.3}$$

where $\gamma \in [0,1]$ controls the relative weighting of each. When $\gamma = 1$ the mapping

[1] This resolves misleading similarity scores based on different sign combinations.

is based only on semantic similarity. Conversely, when $\gamma = 0$ the mapping is based only on the correlation similarity.

5.4.4.1 Semantic similarity (Θ_s)

If the source and target domains have different languages, a translation module is required. For this purpose, we deploy cross-lingual word embeddings. Cross-lingual embeddings are trained using corpora from multiple languages, and can be used to compute word similarities in different languages (Vulić and Moens, 2015a,b; Smith et al., 2016). Empirical evidence indicates that they can also facilitate better knowledge transfer between languages (Mogadala and Rettinger, 2016; Ammar et al., 2016; Mrkšić et al., 2017). The majority of cross-lingual word embedding models are trained by exploiting sources of monolingual text alongside a smaller cross-lingual corpus of aligned text (Ruder, 2017). The alignment can be made at word (Mikolov et al., 2013b; Dinu et al., 2014; Vulić and Moens, 2015a; Ammar et al., 2016; Smith et al., 2016; Artetxe et al., 2018), sentence (Zou et al., 2013; Levy et al., 2017), and document level (Vulić and Moens, 2016; Mogadala and Rettinger, 2016). In this chapter, we utilize a method for learning bilingual word embeddings proposed by Smith et al. (2016).

First, for each of the source and target languages, we respectively learn a word embedding space based on monolingual text. For all languages considered in our experiments (English, French and Spanish) we obtained word embeddings by applying `fastText` on corresponding Wikipedia corpora (Bojanowski et al., 2017).[2] The dimensionality of the word embeddings was set to $d = 300$. Then, we used a core selection of exact translation pairs ($\sigma \rightarrow \tau$) from the source to the target domain language to generate bilingual embeddings. Given the embedding matrices of this alignment dictionary, $\mathbf{E}_\sigma$ and $\mathbf{E}_\tau$ both $\in \mathbb{R}^{m \times d}$, where m, d denote the number of translation pairs and the dimensionality of the word embedding respectively, we learn a transformation matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ such that $\mathbf{E}_\tau \approx \mathbf{E}_\sigma \mathbf{W}$. $\mathbf{W}$ is an orthogonal matrix learned by minimizing the squared Euclidean distance between $\mathbf{E}_\sigma$ and $\mathbf{E}_\tau$, i.e.

[2]The embeddings were obtained from https://github.com/facebookresearch/fastText

$$\begin{aligned} &\underset{\mathbf{W}}{\operatorname{argmin}} \|\mathbf{E}_\sigma \mathbf{W} - \mathbf{E}_\tau\|_2^2 , \\ &\text{subject to } \mathbf{W}^\top \mathbf{W} = \mathbf{I}. \end{aligned} \tag{5.4}$$

The orthogonality constraint ensures that the transformation works both ways, that is $\mathbf{E}_\tau \approx \mathbf{E}_\sigma \mathbf{W}$, $\mathbf{E}_\sigma \approx \mathbf{E}_\tau \mathbf{W}^\top$, and $\mathbf{E}_\tau \approx \mathbf{E}_\tau \mathbf{W}^\top \mathbf{W}$ (Smith et al., 2016). In addition, Artetxe et al. (2016) have empirically shown that it also improves the performance of machine translation. The exact solution of Eq. 5.4 is given by $\mathbf{W} = \mathbf{V}\mathbf{U}^\top$, where $\mathbf{E}_\tau^\top \mathbf{E}_\sigma = \mathbf{U}\Sigma\mathbf{V}^\top$ is the singular value decomposition of $\mathbf{E}_\tau^\top \mathbf{E}_\sigma$ (Golub and Reinsch, 1970).[3]

A query's embedding is defined as the average of the embeddings of its tokens, an effective practice for short texts (Mikolov et al., 2013c; Xu et al., 2015; Benton et al., 2016; Zou et al., 2018). We denote with $\mathbf{v}_{\mathrm{S}_i}$, $\mathbf{v}_{\mathrm{T}_j}$ both $\in \mathbb{R}^{1\times d}$, the embeddings of a source query (from $\mathcal{Q}_\mathrm{S}$) and of a target query from $\mathcal{P}_\mathrm{T}$, respectively. Then, an element ω_{ij} from the cosine similarity matrix $\Omega \in \mathbb{R}^{s\times|\mathcal{P}_\mathrm{T}|}$ between the embeddings of source and valid target queries is given by

$$\omega_{ij} = \frac{\mathbf{v}_{\mathrm{S}_i} \mathbf{W} \mathbf{v}_{\mathrm{T}_j}^\top}{\|\mathbf{v}_{\mathrm{S}_i} \mathbf{W}\|_2 \left\|\mathbf{v}_{\mathrm{T}_j}\right\|_2}. \tag{5.5}$$

Note that the cosine similarities are computed after projecting the embeddings of the source domain to the target domain using the transformation matrix $\mathbf{W}$.

In theory, we can directly use Eq. 5.5 to determine the k most similar target queries to the source query, thus providing a one-to-many mapping. However, in practice when conducting translations based on cross-lingual word embeddings, this may result in the presence of "hubs", i.e. target words or queries that are similar to unrealistically many different source words, a development that reduces the performance of translation (Dinu et al., 2014; Smith et al., 2016). Smith et al. (2016) mitigate this effect by using an inverted softmax ranking, described next.

Given q_i in the source language, its translation is determined by finding candi-

[3]The proof of the solution can be found in (Artetxe et al., 2016).

date target queries q'_j that maximize the probability defined by

$$P_{j\to i} = \frac{\exp\left(\eta\, \omega_{ij}\right)}{\alpha_j \sum_{z=1}^{s} \exp\left(\eta\, \omega_{iz}\right)}, \tag{5.6}$$

where α_j is a normalization factor that ensures $P_{j\to i}$ is a probability, and s is the number of source queries in the vocabulary. The inverted softmax estimates the probability $P_{j\to i}$ that a candidate target query translates back to the source query, rather than the other way around, $P_{i\to j}$ (Dinu et al., 2014; Smith et al., 2016). If a target query is a hub, then the denominator in Eq. 5.6 will be large, preventing this target query from being selected. The parameter η is learned by maximizing the log probability over the alignment dictionary ($\sigma \to \tau$), i.e.,

$$\underset{\eta}{\operatorname{argmax}} \sum_{\text{pairs } ij} \ln\left(P_{j\to i}\right). \tag{5.7}$$

The top-k queries from $\mathcal{P}_\mathrm{T}$ with the highest pairing probability ($P_{j\to i}$) are then selected as possible translations of the source query q_i. Then, we compute the semantic (cosine) similarity score Θ_s between the source query q_i and the target query q_j using

$$\Theta_\mathrm{s}(q_i, q_j) = \frac{\mathbf{e}_{q_i}\mathbf{W}\,\mathbf{e}_{q_j}^{\top}}{\left\|\mathbf{e}_{q_i}\mathbf{W}\right\|_2 \left\|\mathbf{e}_{q_j}\right\|_2}, \tag{5.8}$$

where $\mathbf{e}_{q_i}$, $\mathbf{e}_{q_j}$ are the embeddings of q_i, q_j, respectively. Our experiments report results for a variety of values of k.

If the language in the source and the target domain is the same, the previously described approach is not applicable. Given potential differences in querying preferences, some of the source queries, $\mathcal{Q}_\mathrm{S}$, may not be present in the pool of candidate target queries, $\mathcal{P}_\mathrm{T}$. Therefore, we use cosine similarity to map each source query to the k most similar target ones based on Eq. 5.8 and using the common word embedding space for the shared language.

5.4.4.2 Temporal Correlation Similarity (Θ_c)

We compute the Pearson correlation between the frequency time series of the source and target queries over a fixed period (set to 5 years in our experiments). Since the flu season may be offset in the target domain with respect to the source domain, we computed the maximum correlation between these two frequency time series using a shifting window of $\pm\xi$ weeks. The range of possible values for ξ is determined based on the seasonal offset between the source and target countries (see Section **??**).

Given a source query, q_i, and a target query, q_j which is a member of a mapping set $\mathcal{T}_i$ (consisting of $k \geq 1$ queries from $\mathcal{P}_T$), and their associated daily search frequencies, $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$, respectively, the temporal correlation similarity, Θ_c, is given by

$$\Theta_c(q_i, q_j) = \rho\left(\mathbf{x}_i(t), \mathbf{x}_j(t + l_{ij})\right), \tag{5.9}$$

where $\rho(x_i(t), x_j(t + l_{ij}))$ denotes the optimal Pearson correlation coefficient between $\mathbf{x}_i$, $\mathbf{x}_j$ within the shifting window. Note that the optimal window is independently computed for each target query in $\mathcal{T}_i$, and thus optimal shifts may vary.

5.4.5 Step 3 — Weighting target queries

In the previous steps, we have established that a source query q_i, which has received a regression weight w_i, is mapped to a set, $\mathcal{T}_i$, of $k \geq 1$ queries in the target domain. If $k = 1$, then we can directly assign w_i to the single target query. If $k > 1$, then the source query's weight, w_i, should be distributed across these k mapping target queries. To perform this, we have considered three alternatives:

1. **Uniform.** We divide the source query weight, w_i, by the number of queries q'_j in $\mathcal{T}_i$, and assign each query in $\mathcal{T}_i$ a weight equal to

$$w'_j = w_i / k. \tag{5.10}$$

2. **Non-uniform.** The k target query weights are determined based on each target query's similarity score Θ_{ij}, $j \in \{2, k\}$ (see Eq. 5.3) with the source query. More

specifically, a target weight w'_j is defined as

$$w'_j = w_i \Theta_{ij'} / \sum_{q'_j \in \mathcal{T}_i} \Theta_{ij'} \,. \tag{5.11}$$

To obtain a baseline performance estimate, we randomly shuffle the established query mappings in Step 2, and then transfer the source weights to the k target queries using the uniform approach. We repeat this process multiple times and report the mean performance of these randomized transfer learning models.

5.5 Experiments

We deploy the proposed transfer learning framework to estimate ILI rates in 3 target countries without using any ground truth from these countries to supervise modeling. The US is always the source country, while the target countries are FR, ES and AU. We assess the performance of the proposed model, comparing it to various baselines, and also provide a qualitative analysis, aiming to interpret the inner workings of our approach.

5.5.1 Experiment Settings

After applying the semantic filter (Eq. 5.2) to the pool of 34,121 US queries, 1,403 queries were retained. The evaluation protocol was as follows. We trained a source model (US) using the first 5 flu seasons (2007-2012). A flu season is conventionally defined as the 1-year long period from the first week in September to the last week of August in the next year.[4] Prior to applying elastic net, we maintained search queries that had a Pearson correlation $\geq .3$ with the US ILI rates (these queries may vary per training fold). We then transfer the model to FR, ES, and AU and subsequently test the model in the following season (2012-2013). Then, we move our training data window to include the 2012-13 flu season, removing the first flu season (2007-2008) and test in the following season (2013-2014), so that we still have 5 flu seasons to train. We repeat this process until we have tested on the last

[4]Note that for AU this may result in including the end of a flu season and the beginning of the next in training and test folds.

flu season in our data set (2015-2016), testing 4 times in total. The window size (ξ) used for identifying optimal correlations between the frequency time series of the source and target queries (see Section 5.4.4.2) is set to ± 6 weeks for FR and ES. The window is the same for AU, although prior to applying it, the time series are shifted by 6 months to account for the seasonal difference in the northern and southern hemispheres. For a 1-to-k mapping from a source to a set of target queries, we explore sizes up to $k = 5$ (values > 5 did not yield any different insights or performance improvements). We evaluate the performance of transferred models by comparing our estimates with their national public health estimates, using Pearson correlation (r), MAE, and RMSE, with the two last (MAE, RMSE) being the most reliable metrics.

5.5.2 Baseline Models

To demonstrate the effectiveness of our transferring learning framework, we compare it with four baseline models:

- **Random**. After determining the one-to-k mapping between source and target queries, the mappings are randomly permuted. The source query weight is uniformly distributed across the k target queries. We repeat this process 2,000 times and report the average inference performance. This random assignment of query weights provides a possibly worst case baseline.

- **Transfer component analysis (TCA)**. TCA is a transfer learning approach that aims to learn transfer components across source and target domains in a reproducing kernel Hilbert space using maximum mean discrepancy (Pan et al., 2009). After we map source to target queries, TCA is applied to source and target query frequencies.

- **Unsupervised**. We apply a semantic filter (described in Eq. 5.2) to filter queries that are irrelevant to the flu topic. The term pairs {'grippe', 'fièvre'}, {'gripe', 'fiebre'} and {'flu', 'fever'} are used to define this semantic filter in FR, ES and AU, respectively. Queries with $g \leq .5$ are filtered out and are not considered in our experiments. The mean frequency of the retained queries is regarded as a

proxy of the estimated ILI rates. These estimates are in different scale with true ILI rates, thus we only report correlation (r).

- **Supervised**. We first apply a semantic filter (see point above) to the queries of each target country. We then train an elastic net, using an additional correlation filter in each fold ($r \geq .3$ with the target values in the training data). This is inline with previously proposed, state-of-the-art supervised models for the task (Lampos et al., 2017) and is considered to be the top performance we could obtain, if we had access to ground truth in the target countries.

5.5.3 Quantitative Analysis

Performance estimates are enumerated in Tables 5.2, 5.3, and 5.4 for each transfer learning task (US→FR, US→ES, US→AU). We first explored the extreme cases of $\gamma = 0$ and $\gamma = 1$ (Eq. 5.3) that result in using only temporal correlation or semantic similarity, respectively.

For $\gamma = 0$, spurious queries are very likely to be included in the target domain's mappings. This is a result of the way the pool of target queries, $\mathcal{P}_{\mathrm{T}}$, was originally formed (see Section 5.3.1). Seasonal search queries, correlating with the occurrence of flu incidents in a population, are very likely to be selected as mappings, e.g. "symptoms flu" was mapped to "ski serre chevalier" in the US→FR task. Seasonal activities or expressions may change in time, and thus such queries are very unstable predictors. In fact, the best average performance we can obtain for $\gamma = 0$ is considerably worse (MAEs of 61.532, 25.977 and 42.348 for FR, ES, and AU) than for alternative values. In general, the uniform weight allocation for $k = 1$ seems to be a good choice, although performance does not seem to be affected much by different choices of weighting (uniform vs. non-uniform) or different numbers of queries in a mapping (k).

For $\gamma = 1$, we obtain on average more accurate estimates than for $\gamma = 0$. As a precursor to the joint similarity, we also introduce a correlation-based weighting scheme (denoted by "C"), which uses the optimal correlation between source and target queries (after deploying a shifting window) to determine the proportion of

Table 5.2: Performance estimates for the US→FR transfer learning task. Different values of γ determine how queries are mapped from the source to the target domain (γ=1: semantic similarity only, γ=0: temporal correlation only, $\gamma \in (0,1)$: joint similarity score). The best performance among all transfer learning models is denoted in bold. The best performance among models under a different γ is underlined. Only the best random mapping performance (R) is enumerated per choice of γ. The last two rows show the performance of the baseline models.

Mapping	k	**w**	09/2012 – 09/2013			09/2013 – 09/2014			09/2014 – 09/2015			09/2015 – 09/2016			**Average**		
			r	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**
$\gamma = 0$	1	—	.797	78.905	136.098	.789	59.584	93.752	.900	56.107	92.324	.855	51.533	78.073	.835	61.532	100.062
	2	U	.803	80.044	137.247	.794	59.961	94.853	.890	58.372	96.282	.843	55.532	84.438	.833	63.477	103.205
	3	U	.802	79.010	135.905	.796	59.750	95.350	.896	57.241	94.451	.844	57.306	86.588	.834	63.327	103.073
	4	U	.798	79.077	135.892	.795	59.529	95.295	.895	58.380	95.852	.834	59.729	90.180	.830	64.179	104.305
	5	U	.799	78.881	135.743	.794	58.508	95.036	.893	58.439	96.988	.829	60.075	91.182	.829	63.976	104.737
	2	NU	.803	80.012	137.180	.794	59.971	94.869	.891	58.360	96.268	.843	55.502	84.399	.833	63.461	103.179
	3	NU	.802	78.999	135.881	.796	59.763	95.360	.896	57.244	94.453	.844	57.271	86.538	.834	63.319	103.058
	4	NU	.799	79.068	135.875	.795	59.519	95.278	.895	58.367	95.834	.834	59.676	90.106	.830	64.157	104.273
	5	NU	.799	78.868	135.725	.794	58.499	95.015	.893	58.434	96.972	.829	60.029	91.110	.829	63.957	104.706
	1	R	.771	125.422	152.275	.731	93.122	105.769	.807	138.579	158.000	.825	102.972	113.607	.783	115.024	132.413
$\gamma = 1$	1	—	.964	51.885	77.728	.928	24.373	35.801	.974	51.623	69.254	.917	75.416	92.946	.946	50.824	68.932
	2	U	.967	41.298	68.164	.939	22.993	33.287	.973	62.869	81.119	.924	84.469	102.422	.951	52.907	71.248
	3	U	.967	39.789	67.336	.947	21.219	30.446	.972	58.654	79.471	.933	76.235	93.338	.955	48.974	67.648
	4	U	.965	40.120	65.882	.947	24.037	33.095	.970	63.290	85.390	.939	77.601	93.301	.955	51.262	69.417
	5	U	.965	37.632	61.217	.952	26.136	35.651	.972	66.825	90.248	.943	78.479	93.855	.958	52.268	70.243
	2	NU	.968	41.272	68.016	.939	22.925	33.213	.973	61.971	80.280	.924	83.058	101.160	.951	52.306	70.667
	3	NU	.967	39.665	66.933	.948	21.189	30.378	.973	58.568	79.476	.933	75.661	92.917	.955	48.770	67.426
	4	NU	.966	39.754	65.480	.948	23.794	32.767	.971	62.957	85.275	.939	76.868	92.866	.956	50.843	69.097
	5	NU	.966	37.295	60.749	.952	25.925	35.383	.972	66.890	90.583	.943	77.969	93.647	.958	52.020	70.091
	3	R	.891	83.535	113.537	.890	79.396	86.904	.949	116.532	124.478	.922	109.746	119.219	.913	97.302	111.034
	2	C	.968	39.972	65.695	.941	21.639	31.190	.974	59.103	77.964	.926	78.798	97.444	.952	49.878	68.073
	3	C	.967	38.062	64.349	.949	20.408	29.002	.973	56.188	77.822	.933	72.492	90.289	.956	46.788	65.365
	4	C	.965	38.225	63.063	.949	22.869	31.161	.971	60.623	83.764	.938	73.644	90.367	.956	48.840	67.089
	5	C	.966	35.827	58.820	.953	24.940	33.619	.973	63.562	87.764	.942	74.547	90.793	.958	49.719	67.749
$\gamma_{opt} = .5$	1	—	.968	33.475	53.775	.951	22.615	34.416	.973	34.793	58.007	.944	45.324	62.417	**.959**	**34.052**	**52.153**
	2	U	.959	37.461	60.529	.939	24.885	38.056	.967	43.197	69.883	.930	54.504	74.766	.949	40.012	60.809
	3	U	.954	38.786	63.909	.939	26.390	39.771	.968	44.241	71.312	.931	61.182	81.592	.948	42.650	64.146
	4	U	.948	41.150	69.125	.934	29.553	43.996	.966	47.021	74.662	.932	62.330	82.811	.945	45.014	67.649
	5	U	.945	41.936	71.322	.925	30.387	46.164	.963	46.108	75.703	.931	61.750	82.670	.941	45.045	68.965
	2	NU	.959	37.414	60.456	.939	24.881	38.036	.967	43.118	69.763	.930	54.329	74.599	.949	39.936	60.714
	3	NU	.954	38.675	63.792	.940	26.423	39.789	.968	44.452	71.495	.931	61.147	81.601	.948	42.674	64.169
	4	NU	.948	40.867	68.727	.935	29.381	43.748	.966	47.093	74.691	.932	62.323	82.804	.945	44.916	67.492
	5	NU	.946	41.610	70.892	.926	30.201	45.863	.963	46.192	75.685	.931	61.788	82.685	.942	44.948	68.781
	1	R	.913	86.752	110.096	.846	72.130	83.158	.943	94.681	109.176	.942	97.352	104.952	.911	87.729	101.845
Unsupervised	—	—	.936	—	—	.870	—	—	.947	—	—	.910	—	—	.916	—	—
Supervised	—	—	.977	27.331	50.643	.979	23.665	33.994	.992	34.345	62.803	.987	15.011	21.956	.984	25.088	42.349

k: number of target queries (1-to-k mapping), **w**: weighting approach, **U**: uniform, **NU**: non-uniform, **C**: correlation, **R**: random

the source weight that will be allocated to the k mapped queries. In countries that deploy a translation module based on bilingual word embeddings, the "C" scheme outperforms the other two (uniform, non-uniform). For the US→AU task, where high semantic similarity often means that very similar queries are being mapped to each other (given the common language), the optimal model is obtained for $k = 1$, and thus, no further distribution of the weights is required. With or without the "C" weighting scheme, better performance is achieved compared to setting $\gamma = 0$ (MAEs of 46.788/48.77, 33.224/34.834 and 34.509/30.275 for FR, ES, and AU).

The joint similarity scheme attempts to combine the positive attributes of se-

Table 5.3: Performance estimates for US→ES transfer learning task. Different values of γ determine how queries are mapped from the source to the target domain (γ=1: semantic similarity only, γ=0: temporal correlation only, $\gamma \in (0,1)$: joint similarity score). The best performance among all transfer learning models is denoted in bold. The best performance among models under a different γ is underlined. Only the best random mapping performance (R) is enumerated per choice of γ. The last two rows show the performance of the baseline models.

Mapping	k	**w**	09/2012 – 09/2013			09/2013 – 09/2014			09/2014 – 09/2015			09/2015 – 09/2016			**Average**		
			r	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**
$\gamma = 0$	1	—	.808	25.068	41.104	.807	25.789	42.137	.843	29.221	47.360	.791	25.134	38.497	.812	26.303	42.275
	2	U	.799	25.589	42.631	.843	23.850	39.092	.844	30.069	48.120	.821	24.470	36.902	.827	25.994	41.686
	3	U	.795	25.756	42.883	.840	23.669	38.934	.843	29.509	48.189	.813	24.989	37.713	.823	25.981	41.930
	4	U	.783	26.504	43.662	.835	23.671	39.207	.844	29.850	48.335	.809	25.715	38.745	.818	26.435	42.487
	5	U	.783	26.579	43.391	.840	23.605	38.861	.842	30.336	48.800	.806	26.586	39.843	.818	26.776	42.724
	2	NU	.799	25.579	42.610	.843	23.852	39.095	.844	30.060	48.111	.821	24.472	36.907	.827	25.991	41.681
	3	NU	.795	25.748	42.867	.840	23.670	38.936	.843	29.503	48.176	.813	24.989	37.712	.823	25.977	41.922
	4	NU	.784	26.491	43.643	.835	23.671	39.209	.844	29.842	48.325	.809	25.708	38.734	.818	26.428	42.478
	5	NU	.783	26.567	43.380	.840	23.605	38.866	.842	30.324	48.785	.806	26.575	39.826	.818	26.768	42.714
	3	R	.830	40.548	46.584	.903	36.241	40.718	.846	53.929	61.098	.813	45.762	49.637	.848	44.120	49.509
$\gamma = 1$	1	—	.954	28.614	34.944	.976	27.777	30.129	.919	44.638	50.082	.899	43.761	46.590	.937	36.197	40.436
	2	U	.955	27.342	33.979	.976	27.118	29.294	.923	44.723	50.213	.925	44.518	49.547	**.945**	35.926	40.758
	3	U	.958	25.523	31.885	.971	28.293	32.055	.916	47.603	53.909	.917	48.513	54.053	.941	37.483	42.975
	4	U	.960	25.316	31.623	.973	27.998	31.797	.918	46.862	53.458	.918	47.443	52.823	.942	36.905	42.425
	5	U	.957	24.445	30.821	.975	27.169	30.959	.917	45.775	52.620	.914	45.854	51.505	.941	35.811	41.476
	2	NU	.955	26.336	32.978	.977	26.069	28.232	.923	43.543	49.056	.925	43.389	48.409	**.945**	34.834	39.669
	3	NU	.958	25.532	31.879	.971	28.327	32.076	.917	47.471	53.737	.917	48.356	53.908	.941	37.422	42.900
	4	NU	.960	25.324	31.587	.973	28.020	31.814	.919	46.770	53.334	.917	47.391	52.769	.942	36.876	42.376
	5	NU	.958	24.432	30.759	.975	27.197	30.990	.917	45.778	52.576	.915	45.951	51.574	.941	35.839	41.475
	2	R	.731	47.277	53.345	.804	44.924	52.394	.795	60.370	70.934	.719	48.986	56.506	.762	50.389	58.295
	2	C	.954	25.520	33.516	.976	24.408	26.693	.923	41.827	47.782	.924	41.142	46.278	.944	33.224	38.567
	3	C	.957	23.642	31.398	.970	25.353	29.090	.916	44.358	50.846	.916	45.405	51.174	.940	34.690	40.627
	4	C	.960	23.339	30.912	.973	24.900	28.709	.919	43.431	50.236	.918	44.297	49.873	.942	33.992	39.933
	5	C	.957	24.137	30.513	.974	26.555	30.466	.917	44.598	51.464	.915	45.662	51.359	.941	35.238	40.950
$\gamma_{\text{opt}} = .2$	1	—	.931	21.419	30.004	.948	15.403	23.900	.907	27.050	39.864	.888	26.762	35.420	.918	**22.658**	**32.297**
	2	U	.926	21.433	29.944	.941	17.334	25.525	.899	30.166	43.243	.877	30.662	40.272	.911	24.899	34.746
	3	U	.936	21.249	28.841	.961	18.189	24.028	.908	31.608	42.568	.900	35.661	43.995	.926	26.677	34.858
	4	U	.945	21.016	28.161	.965	18.720	23.647	.917	32.235	41.483	.910	37.141	44.448	.934	27.278	34.435
	5	U	.946	20.977	28.041	.967	18.727	23.321	.910	33.330	43.018	.903	36.846	44.547	.932	27.470	34.732
	2	NU	.926	21.427	29.932	.941	17.321	25.510	.899	30.135	43.214	.877	30.626	40.233	.911	24.877	34.723
	3	NU	.936	21.254	28.845	.961	18.186	24.037	.908	31.583	42.554	.900	35.629	43.969	.926	26.663	34.851
	4	NU	.945	21.023	28.158	.965	18.739	23.682	.917	32.241	41.509	.910	37.128	44.438	.934	27.283	34.447
	5	NU	.946	20.983	28.033	.967	18.747	23.364	.910	33.337	43.028	.903	36.872	44.568	.932	27.485	34.748
	1	R	.865	32.859	40.942	.931	33.323	38.687	.878	49.262	57.762	.814	45.799	51.424	.872	40.311	47.204
$\gamma = .5$	1	—	.945	20.016	28.161	.965	17.720	23.647	.917	31.235	41.483	.910	36.141	44.448	.934	26.278	34.435
Unsupervised	—	—	.936	—	—	.976	—	—	.910	—	—	.878	—	—	.925	—	—
Supervised	—	—	.968	19.788	24.487	.993	30.642	41.059	.972	24.779	35.861	.954	13.271	20.992	.971	22.120	30.600

k: number of target queries (1-to-k mapping), **w**: weighting approach, **U**: uniform, **NU**: non-uniform, **C**: correlation, **R**: random

mantic and correlation based similarities. To assess its potential contribution, we performed a grid search using 9 values of γ (from .1 to .9), and presented the results for the best performing one (γ_{opt}). For completeness, we also show results for the default choices of $\gamma = .5$ and $k = 1$. Firstly, mappings and weightings based the joint similarity provide significant performance improvements in all tasks (MAEs of 34.052, 22.658 and 22.043 for FR, ES, and AU). Secondly, the best performing model consistently occurs for $k = 1$, i.e. for 1-to-1 query mappings, where no weight redistribution is required. Finally, although results do not deviate much from

Table 5.4: Performance estimates for the US→AU transfer learning task. Different values of γ determine how queries are mapped from the source to the target domain (γ=1: semantic similarity only, γ=0: temporal correlation only, $\gamma \in (0,1)$: joint similarity score). The best performance among all transfer learning models is denoted in bold. The best performance among models under a different γ is underlined. Only the best random mapping performance (R) is enumerated per choice of γ. The last two rows show the performance of the baseline models.

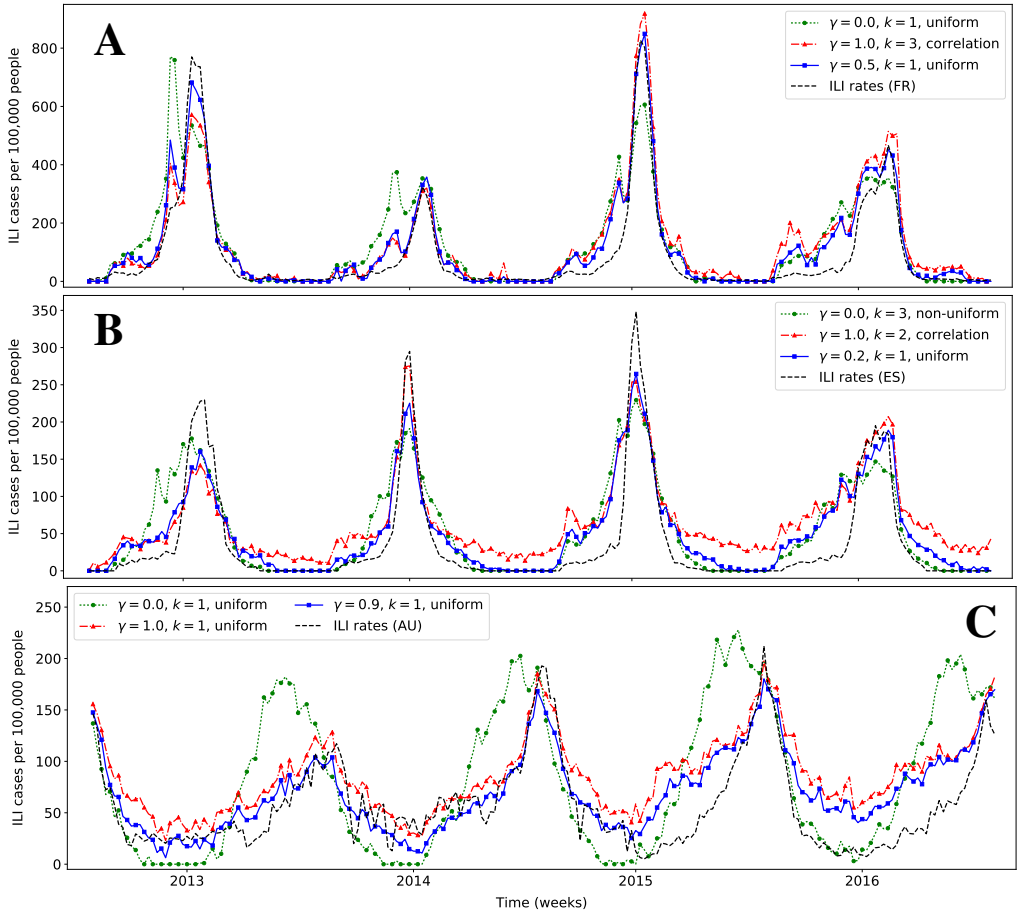
Mapping	k	**w**	09/2012 – 09/2013			09/2013 – 09/2014			09/2014 – 09/2015			09/2015 – 09/2016			**Average**		
			r	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**	**r**	**MAE**	**RMSE**
$\gamma = 0$	1	—	.704	38.804	50.140	.677	39.151	48.508	.630	51.412	65.215	.787	40.025	57.421	.700	42.348	55.321
	2	U	.622	41.824	55.943	.663	41.708	50.752	.633	52.017	66.448	.763	40.557	59.312	.670	44.027	58.114
	3	U	.621	42.263	56.819	.669	42.900	51.487	.631	53.041	67.754	.769	41.330	59.468	.672	44.883	58.882
	4	U	.607	42.040	56.755	.669	42.501	51.008	.634	51.868	66.404	.759	40.287	58.660	.667	44.174	58.207
	5	U	.600	41.900	56.618	.671	41.950	49.692	.647	50.958	64.744	.761	40.899	58.979	.670	43.927	57.508
	2	NU	.623	41.886	55.947	.663	41.642	50.818	.633	52.068	66.590	.763	40.617	59.384	.670	44.053	58.185
	3	NU	.620	42.263	56.812	.668	42.857	51.533	.631	53.062	67.852	.769	41.373	59.540	.672	44.889	58.934
	4	NU	.607	42.031	56.745	.669	42.466	51.039	.634	51.909	66.504	.759	40.343	58.732	.667	44.187	58.255
	5	NU	.600	41.885	56.601	.671	41.928	49.723	.647	51.011	64.844	.761	40.935	59.032	.670	43.940	57.550
	1	R	.653	60.835	71.392	.710	52.090	62.045	.628	67.895	78.856	.738	69.695	75.320	.683	62.629	71.903
$\gamma = 1$	1	—	.916	23.447	26.436	.871	13.994	18.129	.902	35.315	42.126	.971	48.344	50.617	.915	30.275	34.327
	2	U	.900	28.828	33.029	.880	18.583	22.656	.925	39.274	45.149	.989	59.174	60.026	.923	36.465	40.215
	3	U	.896	30.804	35.148	.881	19.492	23.743	.938	36.748	42.294	.990	57.829	58.516	.926	36.218	39.925
	4	U	.889	30.876	35.549	.872	21.475	26.089	.935	37.484	42.966	.994	57.871	58.397	.922	36.926	40.750
	5	U	.882	31.248	35.738	.868	21.320	25.883	.936	37.615	43.059	.992	58.773	59.318	.919	37.239	41.000
	2	NU	.902	28.789	32.947	.880	18.497	22.565	.925	39.278	45.150	.989	59.007	59.861	.924	36.393	40.131
	3	NU	.897	30.805	35.137	.882	19.510	23.775	.938	36.973	42.482	.990	57.779	58.462	.927	36.267	39.964
	4	NU	.890	30.839	35.484	.873	21.367	25.986	.936	37.554	42.999	.994	57.825	58.354	.923	36.896	40.706
	5	NU	.884	31.217	35.678	.870	21.261	25.830	.936	37.609	43.019	.992	58.770	59.309	.920	37.214	40.959
	1	R	.825	58.539	60.310	.793	42.200	46.818	.890	55.940	61.462	.963	65.023	66.924	.868	55.426	58.878
	2	C	.905	27.444	31.356	.881	17.547	21.520	.925	37.373	43.387	.989	58.318	59.229	.925	35.171	38.873
	3	C	.900	28.802	32.701	.882	18.039	22.091	.939	34.534	40.310	.990	56.660	57.516	**.928**	34.509	38.154
	4	C	.894	28.643	32.867	.874	19.505	23.747	.938	34.613	40.360	.994	56.309	57.011	.925	34.768	38.496
	5	C	.888	29.149	33.118	.870	19.259	23.507	.939	34.622	40.252	.992	57.220	57.962	.922	35.063	38.710
$\gamma_{opt} = .9$	1	—	.922	11.997	14.986	.879	15.084	18.011	.898	24.898	31.110	.985	36.191	38.271	.921	**22.043**	**25.594**
	2	U	.892	16.642	19.922	.881	15.719	19.009	.923	23.858	30.280	.988	39.919	41.175	.921	24.034	27.596
	3	U	.890	18.641	22.543	.876	18.391	21.453	.930	23.965	29.934	.989	41.232	42.249	.921	25.557	29.045
	4	U	.883	19.078	23.494	.866	19.766	22.757	.928	23.691	29.686	.991	40.159	41.138	.917	25.673	29.269
	5	U	.875	20.091	24.960	.862	18.791	21.614	.933	23.474	29.474	.991	41.433	42.483	.915	25.947	29.633
	2	NU	.894	16.565	19.826	.882	15.679	18.961	.923	23.830	30.226	.988	39.809	41.071	.922	23.971	27.521
	3	NU	.892	18.588	22.457	.877	18.312	21.353	.930	23.995	29.967	.989	41.230	42.245	.922	25.531	29.005
	4	NU	.885	19.043	23.410	.867	19.639	22.621	.929	23.690	29.673	.991	40.229	41.204	.918	25.650	29.227
	5	NU	.877	19.983	24.795	.864	18.716	21.530	.933	23.414	29.390	.991	41.416	42.462	.916	25.882	29.544
	1	R	.844	47.859	50.120	.817	37.727	40.926	.900	54.008	59.263	.940	55.980	59.071	.875	48.893	52.345
$\gamma = .5$	1	—	.871	18.642	23.367	.848	17.735	20.735	.873	27.140	32.733	.930	39.651	43.484	.880	25.792	30.080
Unsupervised	—	—	.815	—	—	.810	—	—	.881	—	—	.942	—	—	.862	—	—
Supervised	—	—	.891	19.353	25.297	.865	22.048	25.200	.939	18.658	22.473	.971	11.255	14.159	.916	17.829	21.782

k: number of target queries (1-to-k mapping), **w**: weighting approach, **U**: uniform, **NU**: non-uniform, **C**: correlation, **R**: random

the default settings of $\gamma = .5$ and $k = 1$, there are discrepancies between the optimal γ value for each task ($\gamma = .5$, .2 and .9 for FR, ES, and AU). We believe that this is an artifact of the intrinsic characteristics (size, semantic/temporal similarities) of the pool of candidate target queries used for each each task (see Section 5.5.4).

Better performance is always obtained (in terms of MAE and RMSE) compared to the random mapping allocation baseline ("R"), the best performance estimates of which per γ value are provided. The same holds for TCA, which performs

Figure 5.2: Comparison of transfer learning models for estimating ILI rates in France (A), Spain (B) and Australia (C) with the corresponding actual ILI rates obtained by health agencies in these countries.



even worse than random (results are omitted). One explanation for this is that TCA fails to capture the time series structure of this particular data set, an essential property for producing a meaningful solution. Furthermore, the optimal models (joint similarity) outperform the unsupervised baseline in terms of correlation, the only metric which is relevant in this case. Finally, compared to the fully supervised elastic net, the transfer learning unsupervised approach reaches to a comparable performance, which is worse by 23.15%, 5.55%, and 17.5% (in terms of RMSE), for FR, ES, and AU, respectively.

Fig. 5.2 plots the time series of a selection of these estimates, including the ones of the best performing models (blue solid line), in comparison to the ground truth (black dashed line), for each target country. We can see how estimates become

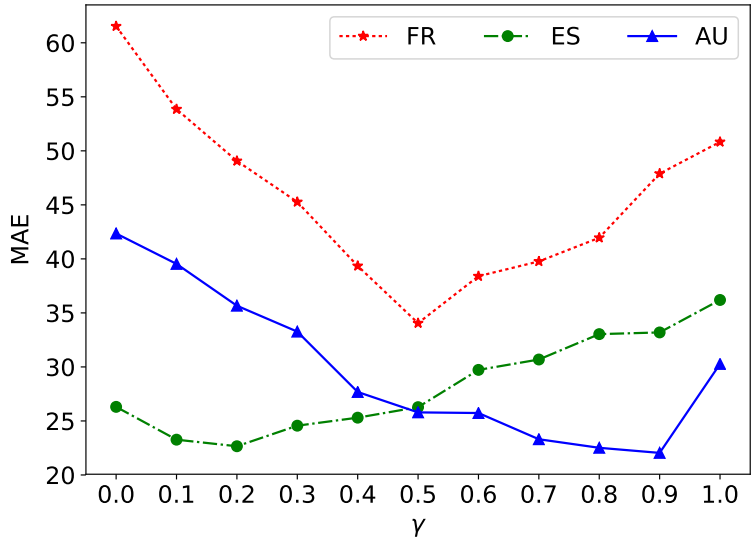


Figure 5.3: MAE under different γ values for the transfer learning models for FR, ES, and AU ($k = 1$).

significantly better when the joint similarity is used vs. its extremes. The transferred models can very often estimate the peak of the flu season accurately. This includes the time of occurrence as well as its intensity. Notably, ILI rates in these countries differ in terms of scale, but the proposed models are capable of capturing this effortlessly, providing further evidence about the search behavior similarities in different countries (Section 5.4.1). At the same time, most models show some inaccuracies, especially during the time periods without a significant amount of flu circulation (e.g. summer).

5.5.4 Qualitative Analysis

One fair criticism for the proposed framework is that in a practical scenario the optimal values for γ and k cannot be validated. However, we have already demonstrated that the default settings of $\gamma = .5$ and $k = 1$ provide very satisfactory performance in all our case studies. Fig. 5.3 looks further into this, depicting performance estimates (MAE) for different values of γ. As discussed previously, optimal γs differ per target country. Interestingly, all error trends are monotonically decreasing (as γ increases) until they reach a minimum, and then start to monotonically increase. We argue that the optimal γ reflects the actual pool of candidate target queries ($\mathcal{P}_T$), although we have too small a sample size to be able to empirically prove this. In

our data, the average correlation over the average semantic similarity ratio between all source-target query pairs is equal to 1.143, .982 and 2.261, for the FR, ES, and AU tasks respectively. These ratios depend on characteristics of the target queries which we are not controlling for in our approach. They do correlate with the respective optimal γ values (.5, .2, and .9), an insight that can be used to make a more informed choice of γ in future applications of the proposed framework.

Table 5.5 lists the top-5 query mappings that were the most impactful in the ILI estimates on average during the 10 weeks with the lowest and greatest MAEs (for the optimal transfer models). Impact is determined by the percentage of an ILI rate that is contributed by a query (frequency $\times$ weight / ILI rate). The identified pairs during the weeks with the lowest errors are topically coherent (about flu) and on many occasions, are accurate translations from the source to the target language. On the other hand, pairs responsible for the largest errors include inaccurate translations that sometimes lead to an off-topic target query selection. For example, "24 hour flu" is mapped to "grippe intestinale" (impact: 13.2%),[5] "child fever" to "sinusitis" (7.7%), and "child temperature" to "warmer" (9.8%). Nevertheless, it is encouraging that some of these mappings may have been avoided by carefully preprocessing the target query candidates to avoid spurious queries.

The optimal joint similarity transfer models do not improve by increasing the number of target queries ($k > 1$). An interpretation for that might be drawn by the fact that for $k = 1$ at most 77.9% of the selected target queries are unique (at least 22.1% are repetitive). As k increases, these mappings do not seem to improve and the error increases monotonically. This might be due to the existence of various spurious queries in the feature space as well as the indication that the mapping has already converged to a subset of the target queries.

Finally, the choice of adding a non-negativity constraint to the regularized regression function for the source domain (Eq. 5.1), was also empirically justified as MAE increases on average by 20.6%, 21.6%, and 20.5% for FR, ES, and AU respectively.

[5] Grippe intestinale translates to stomach flu (viral gastroenteritis).

Table 5.5: Top-5 target queries (with source mappings) in terms of mean ILI estimate impact (%) in the 10 weeks with the lowest and greatest MAE (all test periods), for all target countries (TC), based on their respective optimal transfer learning models.

TC	Mappings during accurate estimates	Impact (%)	Mappings during inaccurate estimates	Impact (%)
FR	flu incubation period → grippe durée	10.90	24 hour flu → grippe intestinale	13.24
	cough fever → la toux	6.30	influenza a treatment → grippe traitement	8.07
	how to treat flu → comment soigner une grippe	6.00	remedies for colds → rhume de cerveau	6.75
	fever flu → fièvre de la grippe	5.47	child temperature → température du corps	6.37
	flu treatment → traitement de la grippe	4.95	child fever → fièvre adulte	6.04
ES	symptoms of flu → symptômes grippe	9.04	mucinez for kids → tratmiento de la grippe	20.76
	fever flu → con gripe	7.49	child fever → sinusitis	7.76
	cough fever → la tos	6.34	influenza a treatment → con gripe	7.02
	flu incubation period → cuanto dura una gripe	5.19	symptoms pneumonia → bronquitis	6.04
	how to treat a fever → para bajar la fiebre	5.03	child temperature → temperatura corporal	5.62
AU	treatment for the flu → flu treatment	9.85	24 hour flu → flu duration	11.51
	cough fever → cough and fever	8.05	child temperature → warmer	9.77
	flu type → influenza type	5.37	how to treat a fever → have a fever	6.94
	symptoms of flu → symptoms of flu	5.11	tamiflu and breastfeeding → flu while pregnant	6.81
	flu incubation period → flu incubation period	5.03	robitussin cf → colds	5.18

5.6 Summary

Prior work on estimating disease rates from online user-generated content relies heavily on supervised learning models. Such models require ground truth data which is usually provided by public health organizations. Ground truth, however, is either sparse or absent from locations with a poor healthcare infrastructure. This is somewhat ironic as it is often stated that web-based approaches hold considerable promise for regions that lack syndromic surveillance systems. This chapter proposes a transfer learning framework as a potential solution to this problem. We leverage semantic and temporal relationships to map a supervised model from a source to a target location. We show that we can obtain satisfactory performance ($r > .92$ on average) that does not deviate much from a fully supervised model ($\leq 21.6\%$ increase in RMSE), without using any ground truth from the target domain.

There are a number of avenues for future work. It is highly desirable to perform a study where the target country is from a low or middle income region. However, such a study is complicated, since the lack of ground truth data does not allow the performance to be quantified. Nevertheless, a qualitative study that demonstrated ILI estimates that followed an expected seasonal pattern would be of value. Our experiments on regions with ground truth data allowed us to investigate parameters k and γ, i.e. the choice for the one-to-k mapping and the relative weight assigned

to the semantic and temporal similarities. Our experiments indicated that a one-to-one ($k = 1$) mapping performed best on average, and that the optimal γ differed per target country. In our analysis, we attempted to justify both outcomes, but further experiments on other regions are needed to understand the effect of these parameters better.

Chapter 6

Conclusions and Future Work

In this chapter, we summarize our contributions in Section 6.1. Then, we make a discussion on our work for real-world problems in Section 6.2. Finally, we provide some future directions of this work in Section 6.3.

6.1 A Summary of Contributions

The framework consists of five steps: (1) acquire data from the Web, (2) extract features from the Web data, (3) select features from extracted features, (4) train supervised learning models, and (5) estimate disease rates and provide early warning before disease outbreak happens. We have made three contributions to complement and improve the framework.

First, we have proposed a joint feature selection method, which consists of a time series similarity filter and a semantic filter. The former filter is based on Pearson correlation, and ensures the features remained are potentially good predictors. The later semantic filter is based on word embeddings, and succeeds in eliminating confounding features, i.e. queries that may be highly correlated with disease rates, but are not referring to the target disease. Using the proposed feature selection method, we have made two case studies, estimating ILI rates from Web search data and IID rates from Twitter data. For both linear (elastic net) and nonlinear regression models (Gaussian Processes), the experimental results demonstrate significant improvement over strong baseline models.

Second, we have investigated the utility of multi-task learning techniques to

disease surveillance from Web search data. A number of related disease surveillance models from different geographies are jointly trained using linear (multi-task elastic net) and nonlinear (multi-task Gaussian Processes) models. The data structures are shared during joint training, which exploits the relatedness between tasks and improves the generalization of the model. We used ILI estimation as a case study, and have shown that the multi-task learning can provide an improved estimate of disease rates when (1) training data is available for multiple geographic locations, and (2) when ground truth training data is sporadic. In addition, we have shown that multi-task learning can improve the estimates of a different country by exploiting a denser health reporting scheme of a reference country.

Third, we have also proposed a transfer learning framework for delivering considerably accurate disease rate models without the existence of ground truth information for a target location. Our framework consists of three steps: (1) learn a regularized regression model for a source country, (2) map the source queries to target ones using semantic and temporal similarity metrics, and (3) re-adjust the weights of the target queries. Our solution is evaluated on the task of estimating ILI rates. In the experiment, we learn a source model for the US, and subsequently transfer it to three other countries, namely France, Spain and Australia. Overall, the transferred models achieve strong performance in terms of Pearson correlation with the ground truth, and their mean absolute error does not deviate greatly from a fully supervised baseline.

6.2 Discussions

Disease surveillance, or more broadly speaking, public health surveillance, aims to monitor and assess the health of a population, and to craft health policies to address the identified health problems. According to Paul and Dredze (2017), there are mainly three components in a public health cycle: (1) assessment, (2) policy development, and (3) assurance.

The aim of the first stage assessment is to monitor the health of a population, and to identify health issues. In the second stage, health policies need to be devel-

oped to address the results of assessment. In the last stage, results of the policies are evaluated. Our work in this thesis focus on the first stage assessment. We estimate disease rates in (nearly) real-time, and this complements traditional disease surveillance in monitoring the health of a population. In addition, our estimations are 2 to 3 weeks ahead of the official numbers published by established health agencies. This can provide an early warning before epidemics happen and affect the second stage policy development. Our work can also be used to evaluate the effectiveness of the policies. For instance, Wagner et al. (2017) evaluated the population impact of a new pediatric influenza vaccination program in England using social media content.

Web-based disease surveillance systems can be used for both developed countries where well-established health systems exist and ground truth is sufficient, and low and middle income countries where such well-established health infrastructure is missing and ground truth partially and does not exist.

For developed countries, although well-established health systems exist, there is usually several weeks delay on reporting. Web-based disease surveillance can complement traditional disease surveillance systems by providing accurate estimates of disease rates in (nearly) real-time, as described in the literature and Chapter 3. This is especially useful during epidemics period, when disease rate is needed in very short time.

Low and middle income countries can benefit more from Web-based disease surveillance. When ground truth only partially or does not exist, supervised learning models are hardly to be used. Multi-task and transfer-learning techniques described in Chapter 4 and 5 are presented for this purpose. Experiments were conducted in developed countries for evaluation purpose. However, given user-generated content and partial (or no) ground truth in a new country, our models can be easily deployed. However, we have to admit that our models tend to work well when the culture between two countries are not distant. If the search behavior between countries are quite different, for example people in the target country are more concerned about diseases and therefore search more, our models will overestimate disease rates.

6.3 Future Work

During the development of the thesis, a lot of interesting ideas and issues have emerged. However, due to time and resource constraint, only some of them have been discussed. Here I summarize the possible directions of the future work.

First, we can make a case study that learns a disease surveillance model from a country with sufficient ground truth, and then transfers the model to a low or middle income country, where only a poor (or no) healthcare infrastructure is established. This is significant as web-based disease surveillance approaches will be particularly useful for regions that lack syndromic surveillance systems.

Second, we can take the advantage of current developments in deep learning and develop a disease surveillance model using deep neural networks. Deep neural networks have demonstrated their advantages in many tasks, such as computer vision, natural language processing, and information retrieval (Goodfellow et al., 2016; Goldberg, 2016). Compared to tradition machine learning techniques, deep learning is good at automatically identifying high-level features from data and solving complex problems. This may improve the current Web-based disease surveillance systems.

Third, we can exploit the methods for disease rate forecasting. The majority of the community work on estimating disease rates in real-time, i.e. they are nowcasting disease rates. Forecasting models, such as hidden Markov model and RNN, can predict the diseases rates days to weeks before the outbreak happen. This can provide early warning to public health agencies.

Appendix A

Full list of Publications

Bin Zou has published the following papers during his PhD

1. Bin Zou, Vasileios Lampos, and Ingemar J. Cox. Transfer Learning for Disease Surveillance Models from Web Search Data. In *International World Wide Web Conference*, 2019. (**Chapter 5**)

2. Bin Zou, Vasileios Lampos, and Ingemar J. Cox. Multi-Task Learning Improves Disease Models from Web Search. In *International World Wide Web Conference*, 2018. (**Chapter 4**)

3. Vasileios Lampos, Bin Zou, and Ingemar J. Cox. Enhancing Feature Selection using Word Embeddings: The Case of Flu Surveillance. In *International World Wide Web Conference*, 2017. (**Chapter 3**)

4. Bin Zou, Vasileios Lampos, Shansong Liang, Zhaochun Ren, Emine Yilmaz, and Ingemar J. Cox. A Concept Language Model for Ad-Hoc Retrieval. In *International World Wide Web Conference*, 2017.

5. Bin Zou, Vasileios Lampos, Russell Gorton, and Ingemar J. Cox. On Infectious Intestinal Disease Surveillance using Social Media Content. In *International Conference on Digital Health*, 2016. (**Chapter 3**)

6. Inferring the Socioeconomic Status of Social Media Users Based on Bahaviour And Language. In *European Conference on Information Retrieval*, 2016.

7. Bin Zou, Akshay Pai, Lauge Sørensen, and Mads Nilsen. Bias Correlation in Images. US patent application number US20170243336A1, GB patent number GB201416416D0, 2017.

Appendix B

Glossary

***n*-grams** A contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. 31

Akaike information criterion (AIC) An estimator of the relative quality of statistical models for a given set of data; the model with lowest AIC is preferred. 55

Autoregressive (AR) A type of random process, which is used to describe certain time-varying processes. It specifies that the output variable depends linearly on its own previous values and on a stochastic term. 50

Autoregressive integrated moving average (ARIMA) A generalization of autoregressive moving average model used in some cases where data show evidence of non-stationarity, where an initial differencing step can be applied one or more times to eliminate the non-stationarity. 51

Autoregressive moving average (ARMA) A parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials, one for the autoregression and the second for the moving average. 50

Autoregressive moving average model with exogenous inputs (ARMAX) Autoregressive moving average model that use exogenous inputs (X) for forecasting. 51

Bayesian information criterion (BIC) A criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. 55

Coefficient of determination Denoted as R^2, which measures the proportion of the variance in the dependent variable that is predictable from the independent variables. 52

Continuous Bag of Words (CBOW) A model loops on the words of each sentence and uses each of these contexts to predict the current word. 65

Correlation In this thesis, it refers to Pearson correlation, which is a measure of the linear correlation between two variables. It has a value between 1 and -1. 26

Disease surveillance The continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice. 21

Elastic net A regularized regression method that linearly combines the ℓ^1 and ℓ^2 penalties of the Lasso and ridge methods. 48

Feature extraction A dimensionality reduction process, where an initial set of raw variables is reduced to more manageable groups (features) for processing, while still accurately and completely describing the original data set. 31

Feature selection Also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. 26

Flu News Europe Joint ECDC and WHO/Europe weekly influenza update. 24

Gaussian Processes (GP) A stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed. 49

Google Flu Trends A web service operated by Google, which provided estimates of influenza activity for more than 25 countries. By aggregating Google Search queries, it attempted to make accurate predictions about flu activity. 36

Google Trends A website by Google that analyzes the popularity of top search queries in Google Search across various regions and languages. 34

Ground truth A term used in various fields to refer to information provided by direct observation (i.e. empirical evidence) as opposed to information provided by inference. 27

ILI rates Number of patients with influenza-like illness symptoms within a population. 24

Influenza-Like Illness (ILI) Also known as acute respiratory infection and flu-like symptoms, is a medical diagnosis of possible influenza or other illness causing a set of common symptoms. 23

Kernel methods A class of algorithms that use kernel functions, which enable them to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. 49

Latent Dirichlet Allocation (LDA) A generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. 64

Latent Semantic Analysis (LSA) A technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. 64

Least absolute shrinkage and selection operator (Lasso) A regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It is a combination of ordinary least squares and ℓ^1-norm regularization. 48

Linear regression A linear approach to modeling the relationship between a dependent variable and one or more explanatory variables (or independent variables). 46

Machine Learning (ML) A field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) from data, without being explicitly programmed. 26

Matérn covariance function A covariance function that handles abrupt changes in the predictors. 62

Mean absolute error (MAE) A measure of the averaged absolute difference between the estimated and observed variables. 54

Mean squared error (MSE) A measure of the averaged squared difference between the estimated and observed variables. 54

Moving average (MA) A common approach for modeling time series, which specifies that the output variable depends linearly on the current and various past values of a stochastic term. 50

Multi-task elastic net A multi-task version of elastic net, which explores the linear relations between different tasks. 84

Multi-task Gaussian Processes A multi-task version of Gaussian processes, which explores the nonlinear relations between different tasks. 84

Multi-task learning A subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks. 27

Natural Language Processing (NLP) A subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. 26

Neural networks An information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurones) working in unison to solve specific problems. 50

Nonlinear regression A form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. 49

Ordinary least squares A type of linear least squares method for estimating the unknown parameters in a linear regression model. The parameters are chosen by minimizing the sum of the squares of the differences between the observed dependent variable and those predicted by the linear function. 46

Recurrent Neural Networks (RNN) A class of neural network where connections between nodes form a directed graph along a sequence. This allows it to exhibit temporal dynamic behavior for a time sequence. 64

Regression Statistical processes for estimating the relationships among variables. More specifically, regression helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied. 26

Ridge regression Also known as Tikhonov regularization, a commonly used method of regularization of ill-posed problems. It is a combination of ordinary least squares and ℓ^2-norm regularization. 48

Root mean square error (RMSE) Also known as root mean square deviation, the square root of mean squared error. 54

Skip-Gram A model loops on the words of each sentence and tries to use the current word of to predict its neighbors. 65

Squared exponential (SE) A stationary covariance function with smooth sample paths. 62

Syndromic surveillance A type of disease surveillance. It refers to the surveillance of a specific syndrome (a set of related symptoms). 22

Topic modeling A type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. 41

Transfer learning A subfield of machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. 28

User-generated content Also known as user-created content, is any form of content, such as images, videos, text and audio, that have been posted by users of online platforms such as search queries log, blogs, wikis, discussion forums, posts, chats, tweets, and other forms of media. 25

Web-based disease surveillance Disease surveillance that utilizes online user-generated content as a data source. 25

Word embeddings Language modeling and feature learning techniques in natural language processing where words or phrases from the vocabulary are mapped

to vectors of real numbers. Conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with a much lower dimension. 26

Word2vec A model developed by Google to produce word embeddings. The models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. 64

Appendix C

Acronyms

API Application Programming Interface. 25

ASPREN Australian Sentinel Practices Research Network. 45

BRFSS Behavioral Risk Factor Surveillance Systems. 23

CDC Centers for Disease Control and Prevention. 23

ECDC European Centre for Disease Prevention and Control. 24

EISN European Influenza Surveillance Network. 24

HHS Department of Health and Human Services. 44

IID Infectious Intestinal Disease. 59

ILINet Influenza-Like Illness Surveillance Network. 23

NSDUH National Survey on Drug Use and Health. 23

PHE Public Health England. 44

RCGP Royal College of General Practitioners. 44

SISSS Spanish Influenza Sentinel Surveillance System. 45

SN French GPs Sentinelles Network. 45

Bibliography

S. Abbar, Y. Mejova, and I. Weber. You Tweet What you Eat: Studying Food Consumption through Twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206, 2015.

A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia. Multi-Task CNN Model for Attribute Prediction. *IEEE Transactions on Multimedia*, 17(11):1949–1959, 2015.

H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Twitter Improves Seasonal Influenza Prediction. In *Proceedings of the 2012 International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 61–70, 2012.

H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

C. Alicino, N. L. Bragazzi, V. Faccio, D. Amicizia, D. Panatto, R. Gasparini, G. Icardi, and A. Orsi. Assessing Ebola-related Web Search Behaviour: Insights and Implications from an Analytical Study of Google Trends-based Query Volumes. *Infectious Diseases of Poverty*, 4(54), 2015.

S. Amir, R. Astudillo, and W. Ling. INESC-ID: A Regression Model for Large Scale Twitter Sentiment Lexicon Induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 613–618, 2015.

W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith. Massively Multilingual Word Embeddings. *arXiv preprint arXiv:1602.01925*, 2016.

H. K. Andreassen, M. M. Bujnowska-Fedak, C. E. Chronaki, R. C. Dumitru, I. Pudule, S. Santana, H. Voss, and R. Wynn. European Citizens' Use of E-health Services: A Study of Seven Countries. *BMC Public Health*, 7(53), 2007.

E. Aramaki, S. Maskawa, and M. Morita. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In *Proceeding of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, 2011.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-Task Feature Learning. In *Proceedings of Advances in Neural Information Processing Systems 19*, 2006.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex Multi-Task Feature Learning. *Machine Learning*, 73(3):243–272, 2008.

M. Artetxe, G. Labaka, and E. Agirre. Learning Principled Bilingual Mappings of Word Embeddings while Preserving Monolingual Invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, 2016.

M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised Neural Machine Translation. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

B. Bakker and T. Heskes. Task Clustering and Gating for Bayesian Multitask Learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

M. M. Barry, C. Domegan, O. Higgins, and J. Sixsmith. A Literature Review on Health Information Seeking Behaviour on the Web: A Health Consumer and Health Professional Perspective. 2011.

J. Baxter. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12(1):149–198, 2000.

D. Beck, T. Cohn, and L. Specia. Joint Emotion Analysis via Multi-Task Gaussian Processes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1798–1803, 2014.

D. Beck, T. Cohn, C. Hardmeier, and L. Specia. Learning Structural Kernels for Natural Language Processing. *Transactions of the Association for Computational Linguistics*, 3:461–473, 2015.

S. Ben-David and R. Schuller. Exploiting Task Relatedness for Multiple Task Learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of Representations for Domain Adaptation. In *Proceedings of Advances in Neural Information Processing Systems 19*, pages 137–144, 2007.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A Theory of Learning from Different Domains. *Machine Learning*, 79(1-2):151–175, 2010.

A. Benton, R. Arora, and M. Dredze. Learning Multiview Embeddings of Twitter Users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 14–19, 2016.

A. Benton, M. Mitchell, and D. Hovy. Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–162, 2017.

T. M. Bernardo, A. Rajic, I. Young, K. Robiadek, M. T. Pham, and J. A. Funk. Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation. *Journal of Medical Internet Research*, 15(7), 2013.

S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-Task Learning for HIV Therapy Screening. In *Proceedings of the 25th International Conference on Machine Learning*, pages 56–63, 2008.

M. Biggerstaff, M. Johansson, D. Alper, L. C. Brooks, P. Chakraborty, D. C. Farrow, S. Hyun, S. Kandula, C. McGowan, N. Ramakrishnan, et al. Results from the

Second Year of A Collaborative Effort to Forecast Influenza Seasons in the United States. *Epidemics*, 24:26–33, 2018.

Z. Bitvai and T. Cohn. Predicting Peer-to-Peer Loan Rates using Bayesian Non-Linear Regression. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2203–2209, 2015.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

J. Blitzer, R. McDonald, and F. Pereira. Domain Adaptation with Structural Correspondence Learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, 2006.

S. J. Blumberg and J. V. Luke. Coverage Bias in Traditional Telephone Surveys of Low-income and Young Adults. *Public Opinion Quarterly*, 71(5):734–749, 2007.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146, 2017.

J. Bollen, H. Mao, and X. Zeng. Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2(1):1–8, 2011.

E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams. Multi-Task Gaussian Process Prediction. In *Proceedings of Advances in Neural Information Processing Systems 20*, pages 153–160, 2007.

J. C. Bosley, N. W. Zhao, S. Hill, F. S. Shofer, D. A. Asch, L. B. Becker, and R. M. Merchant. Decoding Twitter: Surveillance and Trends for Cardiac Arrest and Resuscitation Communication. *Resuscitation*, 84(2):206–212, 2013.

G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.

T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.

S. Brody and N. Elhadad. Detecting Salient Aspects in Online Reviews of Health Providers. In *AMIA Annual Symposium Proceedings*, volume 2010, page 202, 2010.

D. A. Broniatowski, M. J. Paul, and M. Dredze. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLOS One*, 8(12):e83672, 2013.

D. A. Broniatowski, M. Dredze, M. J. Paul, and A. Dugas. Using Social Media to Perform Local Influenza Surveillance in an Inner-city Hospital: A Retrospective Observational Study. *Journal of Medical Internet Research*, 1(1), 2015.

A. D. Bull. Convergence Rates of Efficient Global Optimization Algorithms. *Journal of Machine Learning Research*, 12:2879–2904, 2011.

D. Carmel, A. Mejer, Y. Pinter, and I. Szpektor. Improving Term Weighting for Community Question Answering Search using Syntactic Analysis. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 351–360, 2014.

R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.

G. Chandrashekar and F. Sahin. A Survey on Feature Selection Methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

L. Chen, K. S. M. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash. Syndromic Surveillance of Flu on Twitter using Weakly Supervised Temporal Topic Models. *Data Mining and Knowledge Discovery*, 30(3):681–710, 2016.

M. Chen, K. Q. Weinberger, and J. Blitzer. Co-Training for Domain Adaptation. In *Proceedings of Advances in Neural Information Processing Systems 24*, pages 2456–2464, 2011.

C. K. Y. Cheng, H. Channarith, and B. J. Cowling. Potential Use of School Absenteeism Record for Disease Surveillance in Developing Countries, Case Study in Rural Cambodia. *PLOS One*, 8(10):e76859, 2013.

C. Chew and G. Eysenbach. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLOS One*, 5(11):e14118, 2010.

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

M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting Depression via Social Media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pages 128–137, 2013.

R. J. W. Cline and K. M. Haynes. Consumer Health Information Seeking on the Internet: The State of the Art. *Health Education Research*, 16(6):671–692, 2001.

N. K. Cobb, A. L. Graham, M. J. Byron, and D. B. Abrams. Online Social Networks and Smoking Cessation: A Scientific Research Agenda. *Journal of Medical Internet Research*, 13(4):e119, 2011.

T. Cohn and L. Specia. Modelling Annotator Bias with Multi-Task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 32–42, 2013.

T. Cohn, D. Preotiuc-Pietro, and N. D. Lawrence. Gaussian Processes for Natural Language Processing. In *Proceedings of 52th Annual Meeting of the Association for Computational Linguistics*, pages 1–3, 2014.

N. Collier, N. T. Son, and N. M. Nguyen. OMG U Got Flu? Analysis of Shared Health Messages for Bio-Surveillance. *Journal of Biomedical Semantics*, 2(5): S9, 2011.

R. Collobert and J. Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, 2008.

C. Cook, F. Heath, and R. L. Thompson. A Meta-Analysis of Response Rates in Web-or Internet-Based Surveys. *Educational and Psychological Measurement*, 60(6):821–836, 2000.

S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi. Assessing Google Flu Trends Performance in the United States During the 2009 Influenza Virus A (H1N1) Pandemic. *PLOS One*, 6(8):e23610, 2011.

P. Copeland, R. Romano, T. Zhang, G. Hecht, D. Zigmond, and C. Stefansen. Google Disease Trends: An Update. *Nature*, 457:1012–1014, 2013.

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

A. Culotta. Lightweight Methods to Estimate Influenza Rates and Alcohol Sales Volume from Twitter Messages. *Language Resources and Evaluation*, 47(1): 217–238, 2013.

H. Dai, B. R. Lee, and J. Hao. Predicting Asthma Prevalence by Linking Social Media Data and Traditional Surveys. *The ANNALS of the American Academy of Political and Social Science*, 669(1):75–92, 2017.

W. Dai, G. Xue, Q. Yang, and Y. Yu. Transferring Naive Bayes Classifiers for Text Classification. In *Proceedings of the 22nd International Conference on Artificial Intelligence*, pages 540–545, 2007.

M. De Choudhury, S. Sharma, and E. Kiciman. Characterizing Dietary Choices, Nutrition, and Language in Food Deserts via Social Media. In *Proceedings of the 19th ACM Conference on Computer-supported Cooperative Work & Social Computing*, pages 1157–1170, 2016.

E. De Quincey and P. Kostkova. Early Warning and Outbreak Detection using Social Networking Websites: The Potential of Twitter. In *Proceedings of 2009 International Conference on Electronic Healthcare*, pages 21–24, 2009.

E. Diaz-Aviles and A. Stewart. Tracking Twitter for Epidemic Intelligence: Case study: Ehec/hus Outbreak in Germany, 2011. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 82–85, 2012.

G. Dinu, A. Lazaridou, and M. Baroni. Improving Zero-shot Learning by Mitigating the Hubness Problem. *arXiv preprint arXiv:1412.6568*, 2014.

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

L. Duan, D. Xu, and I. W. Tsang. Learning with Augmented Features for Heterogeneous Domain Adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 667–674, 2012.

A. F. Dugas, Y.-H. Hsieh, S. R. Levin, J. M. Pines, D. P. Mareiniss, A. Mohareb, C. A. Gaydos, T. M. Perl, and R. E. Rothman. Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics. *Clinical Infectious Diseases*, 54(4):463–469, 2012.

A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, and R. E. Rothman. Influenza forecasting with google flu trends. *PLOS One*, 8(2):e56176, 2013.

S. T. Dumais. Latent Semantic Analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

R. Durichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton. Multitask Gaussian Process Models for Biomedical Applications. In *Proceedings of the 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics*, pages 492–495, 2014.

D. Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

V. L. Edge, F. Pollari, G. Lim, J. Aramini, P. Sockett, S. W. Martin, J. Wilson, and A. Ellis. Syndromic Surveillance of Gastrointestinal Illness using Pharmacy Over-the-counter Sales: A Retrospective Study of Waterborne Outbreaks in Saskatchewan and Ontario. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, pages 446–450, 2004.

S. Emrani, A. McGuirk, and W. Xiao. Prognosis and Diagnosis of Parkinson's Disease Using Multi-Task Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1457–1466, 2017.

J. Eschler, Z. Dehlawi, and W. Pratt. Self-Characterized Illness Phase and Information Needs of Participants in an Online Cancer Forum. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, 2015.

T. Evgeniou and M. Pontil. Regularized Multi-Task Learning. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004.

G. Eysenbach and J. Wyatt. Using the Internet for Surveys and Health Research. *Journal of Medical Internet Research*, 4(2), 2002.

S. Feng and L. Hossain. Risk-informed decisions for epidemics. *Journal of Decision Systems*, 25(sup1):240–247, 2016.

G. Foster, C. Goutte, and R. Kuhn. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 2010 Con-*

ference on Empirical Methods in Natural Language Processing, pages 451–459, 2010.

S. Fox and M. Duggan. Health Online 2013. *Pew Internet & American Life Project*, 1, 2013.

I. C.-H. Fung, Z. T. H. Tse, C.-N. Cheung, A. S. Miu, and K.-W. Fu. Ebola and the Social Media. *The Lancet*, 384(9961):2207, 2014.

F. Galton. Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

V. R. K. Garimella, A. Alfayad, and I. Weber. Social Media Image Analysis for Public Health. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5543–5547, 2016.

N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky. Global Disease Monitoring and Forecasting with Wikipedia. *PLOS Computational Biology*, 10(11):e1003892, 2014.

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.

X. Glorot, A. Bordes, and Y. Bengio. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proceedings of the 28th International Conference on Machine Learning*, pages 513–520, 2011.

R. T. Gluskin, M. A. Johansson, M. Santillana, and J. S. Brownstein. Evaluation of Internet-based Dengue Query Data: Google Dengue Trends. *PLOS Neglected Tropical Diseases*, 8(2):e2713, 2014.

Y. Goldberg. A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

G. H. Golub and C. Reinsch. Singular Value Decomposition and Least Squares Solutions. *Numerische Mathematik*, 14(5):403–420, 1970.

J. Gomide, A. Veloso, W. Meira Jr, V. Almeida, F. Benevenuto, F. Ferraz, and M. Teixeira. Dengue Surveillance based on a Computational Model of Spatio-Temporal Locality of Twitter. In *Proceedings of the 3rd International Web Science Conference*, page 3, 2011.

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep Learning*, volume 1. MIT Press, 2016.

R. S. Gottfried. *Black death*. Simon & Schuster, 2010.

S. Greenwood, A. Perrin, and M. Duggan. Social media update 2016. *Pew Research Center*, 11:83, 2016.

I. Guyon and A. Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182, 2003.

J. Han, X. Tian, G. Yu, and F. He. Disclosure Pattern of Self-Labeled People Living with HIV/AIDS on Chinese Social Networking Site: An Exploratory Study. *Cyberpsychology, Behavior, and Social Networking*, 19(8):516–523, 2016.

C. Harrison, M. Jorder, H. Stern, F. Stavinsky, V. Reddy, H. Hanson, H. Waechter, L. Lowe, L. Gravano, and S. Balter. Morbidity and Mortality Weekly Report. *MMWR. Morbidity and Mortality Weekly Report*, 63(20):441–445, 2014.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer, 2009.

N. Heaivilin, B. Gerbert, J. E. Page, and J. L. Gibbs. Public Health Surveillance of Dental Pain via Twitter. *Journal of Dental Research*, 90(9):1047–1051, 2011.

R. Heffernan, F. Mostashari, D. Das, M. Besculides, C. Rodriguez, J. Greenko, L. Steiner-Sichel, S. Balter, A. Karpati, P. Thomas, et al. New York City Syndromic Surveillance Systems. *Morbidity and Mortality Weekly Report*, pages 25–27, 2004.

K. M. Hiller, L. Stoneking, A. Min, and S. M. Rhodes. Syndromic Surveillance for Influenza in the Emergency Department - A Systematic Review. *PLOS One*, 8 (9):e73832, 2013.

A. E. Hoerl and R. W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970.

J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. Cross-Language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308, 2013.

V. G. Iannacchione. The Changing Role of Address-based Sampling in Survey Research. *Public Opinion Quarterly*, 75(3):556–575, 2011.

D. T. Jamison, J. G. Breman, A. R. Measham, G. Alleyne, M. Claeson, D. B. Evans, P. Jha, A. Mills, and P. Musgrove. *Disease Control Priorities in Developing Countries*. The World Bank, 2006.

J. Jiang and C. Zhai. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, 2007.

F. Jin, W. Wang, L. Zhao, E. R. Dougherty, Y. Cao, C.-T. Lu, and N. Ramakrishnan. Misinformation Propagation in the Age of Twitter. *IEEE Computer*, 47(12):90–94, 2014.

A. K. Johnson and S. D. Mehta. A Comparison of Internet Search Trends and Sexually Transmitted Infection Rates using Google Trends. *Sexually Transmitted Diseases*, 41(1):61–63, 2014.

N. P. Johnson and J. Mueller. Updating the Accounts: Global Mortality of the 1918-1920 "Spanish" Influenza Pandemic. *Bulletin of the History of Medicine*, 76(1): 105–115, 2002.

R. Juric, I. Kim, H. Panneerselvam, and I. Tesanovic. Analysis of ZIKA Virus Tweets: Could Hadoop Platform Help in Global Health Management? In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

P. Koehn and J. Schroeder. Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, 2007.

R. Kohavi and G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

J. Krumm, N. Davies, and C. Narayanaswami. User-Generated Content. *IEEE Pervasive Computing*, 7(4):10–11, 2008.

B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1785–1792, 2011.

A. Lamb, M. J. Paul, and M. Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, 2013.

V. Lampos and N. Cristianini. Tracking the Flu Pandemic by Monitoring the Social Web. In *Proceedings of the 2nd International Workshop on Cognitive Information Processing*, pages 411–416, 2010.

V. Lampos and N. Cristianini. Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4): 72, 2012.

V. Lampos, D. Preoţiuc-Pietro, and T. Cohn. A User-Centric Model of Voting Intention from Social Media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 993–1003, 2013.

V. Lampos, N. Aletras, D. Preoţiuc-Pietro, and T. Cohn. Predicting and Characterising User Impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 405–413, 2014.

V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen. Advances in Nowcasting Influenza-like Illness Rates using Search Query Logs. *Scientific Reports*, 5: 12760, 2015.

V. Lampos, B. Zou, and I. J. Cox. Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance. In *Proceedings of the 26th International Conference on World Wide Web*, pages 695–704, 2017.

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.

S. Lee, J. Zhu, and E. P. Xing. Adaptive Multi-task Lasso: With Application to eQTL Detection. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 1306–1314, 2010.

K. L. Lentine, M. A. Schnitzler, K. C. Abbott, K. Bramesfeld, P. M. Buchanan, and D. C. Brennan. Sensitivity of Billing Claims for Cardiovascular Disease Events among Kidney Transplant Recipients. *Clinical Journal of the American Society of Nephrology*, 4(7):1213–1221, 2009.

O. Levy and Y. Goldberg. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 302–308, 2014.

O. Levy, Y. Goldberg, and I. Ramat-Gan. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the 18th Conference on Computational Natural Language Learning*, pages 171–180, 2014.

O. Levy, Y. Goldberg, and I. Dagan. Improving Distributional Similarity with

Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

O. Levy, A. Søgaard, and Y. Goldberg. A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 765–774, 2017.

Y. Li and C. Hu. A method for tracking flu trends through weibo. *International Journal of Database Theory and Application*, 9(5):91–100, 2016.

Z. Li, T. Liu, G. Zhu, H. Lin, Y. Zhang, J. He, A. Deng, Z. Peng, J. Xiao, and S. Rutherford. Dengue Baidu Search Index Data can Improve the Prediction of Local Dengue Epidemic: A Case Study in Guangzhou, China. *PLOS Neglected Tropical Diseases*, 11(3):e0005354, 2017.

H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*, volume 454. Springer, 2012.

J. Liu, S. Ji, and J. Ye. Multi-task Feature Learning via Efficient $\ell_{2,1}$-Norm Minimization. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 339–348, 2009.

P. Liu, X. Qiu, and X. Huang. Recurrent Neural Network for Text Classification with Multi-task Learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 2873–2879, 2016a.

Y. Liu, Q. Mei, D. A. Hanauer, K. Zheng, and J. M. Lee. Use of Social Media in the Diabetes Community: An Exploratory Analysis of Diabetes-related Tweets. *Journal of Medical Internet Research Diabetes*, 1(2):e4, 2016b.

M. Lukasik, T. Cohn, and K. Bontcheva. Classifying Tweet Level Judgements of Rumours in Social Media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2590–2595, 2015.

S. F. Magruder, S. H. Lewis, A. Najmi, and E. Florio. Progress in Understanding and Using Over-The-Counter Pharmaceuticals for Syndromic Surveillance. *Morbidity and Mortality Weekly Report*, pages 117–122, 2004.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain Adaptation: Learning Bounds and Algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

J. R. Massey and J. Frank. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

B. Matérn. *Spatial Variation*. Springer, 1986.

S. F. McGough, J. S. Brownstein, J. B. Hawkins, and M. Santillana. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLOS Neglected Tropical Diseases*, 11(1):e0005295, 2017.

D. J. McIver and J. S. Brownstein. Wikipedia Usage Estimates Prevalence of Influenza-like Illness in the United States in Near Real-Time. *PLOS Computational Biology*, 10(4):e1003581, 2014.

T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur. Recurrent Neural Network based Language Model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, 2010.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, 2013a.

T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting Similarities among Languages for Machine Translation. *arXiv preprint arXiv:1309.4168*, 2013b.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013c.

T. Mikolov, W.-T. Yih, and G. Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013d.

G. J. Milinovich, G. M. Williams, A. C. A. Clements, and W. Hu. Internet-Based Surveillance Systems for Monitoring Emerging Infectious Diseases. *The Lancet Infectious Diseases*, 14(2):160–168, 2014.

M. Miller, T. Banerjee, R. Muppalla, W. Romine, and A. Sheth. What are People Tweeting about Zika? An Exploratory Study Concerning its Symptoms, Treatment, Transmission, and Prevention. *Journal of Medical Internet Research Public Health and Surveillance*, 3(2), 2017.

M.-F. Moens, J. Li, and T.-S. Chua. *Mining User Generated Content*. Chapman & Hall/CRC, 2014.

A. Mogadala and A. Rettinger. Bilingual Word Embeddings from Parallel and Non-parallel Corpora for Cross-language Text Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–702, 2016.

M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi, and S. Kumar. Google Correlate Whitepaper. `https://www.google.com/trends/correlate/whitepaper.pdf`, 2011.

P. Mook, C. Joseph, P. Gates, and N. Phin. Pilot Scheme for Monitoring Sickness Absence in Schools during the 2006/07 Winter in England: Can these Data Be Used as a Proxy for Influenza Activity? *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, 12(12):E11–2, 2007.

E. M. Morgan, C. Snelson, and P. Elison-Bowers. Image and Video Disclosure of Substance Use on Social Media Websites. *Computers in Human Behavior*, 26(6): 1405–1411, 2010.

A. S. Mosa, I. Yoo, and J. C. Parker. Online Electronic Data Capture and Research Data Repository System for Clinical and Translational Research. *Missouri Medicine*, 112(1):46–52, 2015.

N. Mrkšić, I. Vulić, D. Ó. Séaghdha, I. Leviant, R. Reichart, M. Gašić, A. Korhonen, and S. Young. Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-lingual Constraints. *arXiv preprint arXiv:1706.00374*, 2017.

A. J. Ocampo, R. Chunara, and J. S. Brownstein. Using Search Queries for Malaria Surveillance, Thailand. *Malaria Journal*, 12(1):390, 2013.

M. Odlum and S. Yoon. What can We Learn about the Ebola Outbreak from Tweets? *American Journal of Infection Control*, 43(6):563–571, 2015.

Y. Ofran, O. Paltiel, D. Pelleg, J. M. Rowe, and E. Yom-Tov. Patterns of Information-Seeking for Cancer on the Internet: An Analysis of Real World Data. *PLOS One*, 7(9):e45921, 2012.

O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, 2013.

S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain Adaptation via Transfer Component Analysis. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, pages 1187–1192, 2009.

G. Park, H. A. Schwartz, and J. C. Eichstaedt. Automatic Personality Assessment through Social Media Language. *Journal of Personality and Social Psychology*, 108(6):934–952, 2015.

M. J. Paul and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, volume 20, pages 265–272, 2011.

M. J. Paul and M. Dredze. Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models. In *Proceedings of 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 168–178, 2013.

M. J. Paul and M. Dredze. Discovering Health Topics in Social Media using Topic Models. *PLOS One*, 9(8):e103408, 2014.

M. J. Paul and M. Dredze. Social Monitoring for Public Health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183, 2017.

M. J. Paul, M. Dredze, and D. Broniatowski. Twitter Improves Influenza Forecasting. *PLOS Currents Outbreaks*, 2014.

M. J. Paul, M. Dredze, D. A. Broniatowski, and N. Generous. Worldwide Influenza Surveillance through Twitter. In *AAAI Workshop: WWW and Public Health Intelligence*, 2015.

M. J. Paul, R. W. White, and E. Horvitz. Search and Breast Cancer: On Episodic Shifts of Attention over Life Histories of an Illness. *ACM Transactions on the Web*, 10(2):13, 2016.

K. Pearson. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

C. Pelat, C. Turbelin, A. Bar-Hen, A. Flahault, and A.-J. Valleron. More Diseases Tracked by using Google Trends. *Emerging Infectious Diseases*, 15(8):1327, 2009.

D. Pelleg, E. Yom-Tov, and Y. Maarek. Can you Believe an Anonymous Contributor? On Truthfulness in Yahoo! Answers. In *Proceedings of the 2012*

ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, pages 411–420, 2012.

P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47 (11):1443–1448, 2008.

S. Pollett, W. J. Boscardin, E. Azziz-Baumgartner, Y. O. Tinoco, G. Soto, C. Romero, J. Kok, M. Biggerstaff, C. Viboud, and G. W. Rutherford. Evaluating Google Flu Trends in Latin America: Important Lessons for the next Phase of Digital Disease Detection. *Clinical Infectious Diseases*, page ciw657, 2016.

T. Preis and H. S. Moat. Adaptive Nowcasting of Influenza Outbreaks using Google Searches. *Royal Society Open Science*, 1(2):140095, 2014.

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), 2015.

P. Prettenhofer and B. Stein. Cross-language Text Classification using Structural Correspondence Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, 2010.

R. Priedhorsky, D. Osthus, A. R. Daughton, K. R. Moran, N. Generous, G. Fairchild, A. Deshpande, and S. Y. Del Valle. Measuring Global Disease with Wikipedia: Success, Failure, and a Research Agenda. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1812–1834, 2017.

K. W. Prier, M. S. Smith, C. Giraud-Carrier, and C. L. Hanson. Identifying Health-related Topics on Twitter. In *Proceedings of the 2011 International Conference on Social Computing, Behavioral-cultural Modeling, and Prediction*, pages 18–25, 2011.

K. Purcell, L. Rainie, and J. Brenner. Search Engine Use 2012. *Pew Internet & American Life Project Survey*, 2012.

S. Ram, W. Zhang, M. Williams, and Y. Pengetnze. Predicting Asthma-Related Emergency Department Visits using Big Data. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1216–1223, 2015.

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

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

J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, and H. Herrman. Social Media and Suicide Prevention: A Systematic Review. *Early Intervention in Psychiatry*, 10(2):103–121, 2016.

L. Rossignol, C. Pelat, B. Lambert, A. Flahault, E. Chartier-Kastler, and T. Hanslik. A Method to Assess Seasonality of Urinary Tract Infections based on Medication Sales and Google trends. *PLOS One*, 8(10):e76020, 2013.

S. Ruder. A Survey of Cross-Lingual Embedding Models. *arXiv preprint arXiv:1706.04902*, 2017.

M. Santillana, E. O. Nsoesie, S. R. Mekaru, D. Scales, and J. S. Brownstein. Using Clinicians' Search Query Data to Monitor Influenza Epidemics. *Clinical Infectious Diseases*, 59(10):1446–1450, 2014a.

M. Santillana, D. W. Zhang, B. M. Althouse, and J. W. Ayers. What can Digital Disease Detection Learn from (An External Revision to) Google Flu Trends? *American Journal of Preventive Medicine*, 47(3):341–347, 2014b.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2): 461–464, 1978.

J. Shaman and A. Karspeck. Forecasting Seasonal Outbreaks of Influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

S.-Y. Shin, T. Kim, D.-W. Seo, C. H. Sohn, S.-H. Kim, S. M. Ryoo, Y.-S. Lee, J. H. Lee, W. Y. Kim, and K. S. Lim. Correlation between National Influenza Surveillance Data and Search Queries from Mobile Devices and Desktops in South Korea. *PLOS One*, 11(7):e0158539, 2016.

A. Signorini, A. M. Segre, and P. M. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the US during the Influenza A H1N1 Pandemic. *PLOS One*, 6(5):e19467, 2011.

S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. 2016.

T. T. P. Souza and T. Aste. A Nonlinear Impact: Evidences of Causal Effects of Social Media on Market Prices. *arXiv preprint arXiv:1601.04535*, 2016.

C. Stefansen. Google Flu Trends Gets a Brand New Engine. *Google Research Blog*, 2014.

X. Sun, J. Ye, and F. Ren. Real Time Early-stage Influenza Detection with Emotion Factors from Sina Microblog. In *Proceedings of the 5th Workshop on South and Southeast Asian Natural Language Processing*, pages 80–84, 2014.

X. Sun, J. Ye, and F. Ren. Detecting Influenza States based on Hybrid Model with Personal Emotional Factors from Social Networks. *Neurocomputing*, 210:257–268, 2016.

X. Sun, F. Ren, and J. Ye. Trends Detection of Flu Based on Ensemble Models with Emotional Factors from Social Networks. *IEEJ Transactions on Electrical and Electronic Engineering*, 12(3):388–396, 2017.

C. C. Tam. 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.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

K. Torkkola. Feature Extraction by Non-Parametric Mutual Information Maximization. *Journal of Machine Learning Research*, 3(Mar):1415–1438, 2003.

L. Torrey and J. Shavlik. Transfer Learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*, page 242, 2009.

S. Towers, S. Afzal, G. Bernal, N. Bliss, S. Brown, B. Espinoza, J. Jackson, J. Judson-Garcia, M. Khan, and M. Lin. Mass Media and the Contagion of Fear: The Case of Ebola in America. *PLOS One*, 10(6):e0129179, 2015.

M.-F. Tsai and C.-J. Wang. Financial Keyword Expansion via Continuous Word Vector Representations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1453–1458, 2014.

A. Valdivia and S. Monge-Corella. Diseases Tracked by using Google Trends, Spain. *Emerging Infectious Diseases*, 16(1):168, 2010.

P. Velardi, G. Stilo, A. E. Tozzi, and F. Gesualdo. Twitter Mining for Fine-grained Syndromic Surveillance. *Artificial Intelligence in Medicine*, 61(3):153–163, 2014.

I. Vulić and M.-F. Moens. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, pages 719–725, 2015a.

I. Vulić and M.-F. Moens. Monolingual and Cross-lingual Information Retrieval Models based on (Bilingual) Word Embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372, 2015b.

I. Vulić and M.-F. Moens. Bilingual Distributed Word Representations from Document-Aligned Comparable Data. *Journal of Artificial Intelligence Research*, 55:953–994, 2016.

H. Wackernagel. Geostatistics. *Wiley StatsRef: Statistics Reference Online*, 2014.

M. Wagner, V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content. *Journal of Medical Internet Research*, 19(12):e416, 2017.

M. Wagner, V. Lampos, I. J. Cox, and R. Pebody. The Added Value of Online User-Generated Content in Traditional Methods for Influenza Surveillance. *Scientific Reports*, 8(13963), 2018.

M. M. Wagner, A. W. Moore, and R. M. Aryel. *Handbook of Biosurveillance*. Elsevier, 2011.

X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, 2009.

S. Wang, M. J. Paul, and M. Dredze. Exploring Health Topics in Chinese Social Media: An Analysis of Sina Weibo. In *Proceedings of the 2014 AAAI Workshop on the World Wide Web and Public Health Intelligence*, volume 31, page 59, 2014.

Z. Wang, P. Chakraborty, S. R. Mekaru, J. S. Brownstein, J. Ye, and N. Ramakrishnan. Dynamic Poisson Autoregression for Influenza-like Illness Case Count Prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2015.

S. Weisberg. *Applied Linear Regression*, volume 528. John Wiley & Sons, 2005.

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A Survey of Transfer Learning. *Journal of Big Data*, 3(1):9, 2016.

J. G. Wheeler. 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.

R. Wong, J. K. Harris, M. Staub, and J. M. Bernhardt. Local Health Departments Tweeting about Ebola: Characteristics and Messaging. *Journal of Public Health Management and Practice*, 23(2):e16–e24, 2017.

World Health Organization. *AIDS Epidemic Update, December 2006*. World Health Organization, 2006.

World Health Organization. *WHO: Ebola Response Roadmap Situation Report 24 December 2014*, 2014.

J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao, F. Wang, and H. Hao. Short Text Clustering via Convolutional Neural Networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, pages 62–69, 2015.

W. Xu, Z.-W. Han, and J. Ma. A Neural Network based Approach to Detect Influenza Epidemics using Search Engine Query Data. In *Proceedings of 2010 International Conference on Machine Learning and Cybernetics*, volume 3, pages 1408–1412, 2010.

S. Yang, M. Santillana, and S. C. Kou. Accurate Estimation of Influenza Epidemics using Google Search Data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478, 2015.

T. Yano, N. A. Smith, and J. D. Wilkerson. Textual Predictors of Bill Survival in Congressional Committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 793–802, 2012.

A. Yates and N. Goharian. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. In *Proceedings of the 35th European Conference on Information Retrieval*, pages 816–819, 2013.

M. Ybarra and M. Suman. Reasons, Assessments and Actions Taken: Sex and Age Differences in Uses of Internet Health Information. *Health Education Research*, 23(3):512–521, 2008.

W. K. Yih, K. S. Teates, A. Abrams, K. Kleinman, M. Kulldorff, R. Pinner, R. Harmon, S. Wang, and R. Platt. Telephone Triage Service Data for Detection of Influenza-like Illness. *PLOS One*, 4(4):e5260, 2009.

E. Yom-Tov, R. W. White, and E. Horvitz. Seeking Insights about Cycling Mood Disorders Via Anonymized Search Logs. *Journal of Medical Internet Research*, 16(2):e65, 2014.

S. D. Young, W. Yu, and W. Wang. Toward Automating HIV Identification: Machine Learning for Rapid Identification of HIV-related Social Media Data. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 74(Suppl 2):S128, 2017.

Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein. Monitoring Influenza Epidemics in China with Search Query from Baidu. *PLOS One*, 8(5):e64323, 2013.

F. Zhang, J. Luo, C. Li, X. Wang, and Z. Zhao. Detecting and Analyzing Influenza Epidemics with Social Media in China. In *Proceedings of the 2014 Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 90–101, 2014a.

W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji. Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1475–1484, 2015.

W. Zhang, S. Ram, M. Burkart, and Y. Pengetnze. Extracting Signals from Social

Media for Chronic Disease Surveillance. In *Proceedings of the 6th International Conference on Digital Health*, pages 79–83, 2016.

Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial Landmark Detection by Deep Multi-Task Learning. In *European Conference on Computer Vision*, pages 94–108, 2014b.

L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-Task Learning for Spatio-Temporal Event Forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1503–1512, 2015.

P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

A. Zhila, W.-T. Yih, C. Meek, G. Zweig, and T. Mikolov. Combining Heterogeneous Models for Measuring Relational Similarity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1000–1009, 2013.

J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling Disease Progression via Fused Sparse Group Lasso. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1095–1103, 2012.

J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan. Heterogeneous Domain Adaptation for Multiple Classes. In *Artificial Intelligence and Statistics*, pages 1095–1103, 2014.

Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang. Heterogeneous Transfer Learning for Image Classification. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 1304–1309, 2011.

K. Zickuhr and L. Rainie. Wikipedia, Past and Present. *Pew Internet & American Life Project Survey*, 13, 2012.

P. Ziegler. *The Black Death*. Faber & Faber, 2013.

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

B. Zou, V. Lampos, and I. J. Cox. Multi-Task Learning Improves Disease Models from Web Search. In *Proceedings of the 2018 World Wide Web Conference*, pages 87–96, 2018.

H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2):301–320, 2005.

W. Y. Zou, R. Socher, D. Cer, and C. D. Manning. Bilingual Word Embeddings for Phrase-based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, 2013.