Information Processing and Management 59 (2022) 102811

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Unifying telescope and microscope: A multi-lens framework with open data for modeling emerging events

Yunhe Feng *, Chirag Shah

Information School, University of Washington, WA, USA

ARTICLE INFO

Keywords:
Open data
Data retrieval
Data fusion
Model fusion
Google Trends
Twitter

ABSTRACT

Open data is becoming ubiquitous as governments, companies, and even individuals have the option to offer more or less unrestricted access to their non-sensitive data. The benefits of open data, such as accessibility and transparency, have motivated and enabled a large number of research studies and applications in both academia and industry. However, each open data only offers a single perspective, and its potential inherent limitations (e.g., demographic biases) may lead to poor decisions and misjudgments. This paper discusses how to create and use multiple digital lenses empowered by open data, including census data (macro lens), search logs (meso lens), and social data (micro lens), to investigate general real-world events. To reveal the unique angles and perspectives brought by each open lens, we summarize and compare the underpinning open data from eleven dimensions, such as utility, data volume, dynamic variability, and demographic fairness. Then, we propose an easy-to-use and generalized open data driven framework, which automatically retrieves multi-source data, extracts features, and trains machine learning models for the event specified by answering what, when, and where questions. With low labor efforts, the framework's generalization and automation capabilities guarantee an instant investigation of general events and phenomena, such as disasters, sports events, and political activities. We also conduct two case studies, i.e., the COVID-19 pandemic and Great American Eclipse (see Appendix), to demonstrate its feasibility and effectiveness at different time granularities.

1. Introduction

In this section, we first briefly introduce the research background and present the research motivation. Then we highlight our research objectives and contributions.

1.1. Background

Open data is playing an increasingly important role in many applications and services for social good, such as disaster management, policy making, public opinion investigation, social innovation, and economic growth (Desouza & Smith, 2014; Janssen, Charalabidis, & Zuiderwijk, 2012; McCombs & Valenzuela, 2020; Napoli & Karaganis, 2010; Ortmann, Limbu, Wang, & Kauppinen, 2011).

Using open data in understanding natural and social phenomena is a very common method in various scientific disciplines including social and behavioral sciences, information science, and political science, as well as media studies and economics.

* Corresponding author.
E-mail address: yunhe@uw.edu (Y. Feng).

https://doi.org/10.1016/j.ipm.2021.102811
Received 17 May 2021; Received in revised form 17 October 2021; Accepted 2 November 2021
Available online 13 December 2021

Furthermore, open data has become ubiquitous and easily accessible due to the advances of information and communications technology. Besides traditional open data (e.g., census data), search engines and social media offer new sources of open information to the public. For example, Google provides a daily search trends index for given keywords and locations based on the aggregated people's search behaviors.[1] Similarly, Twitter officially allows the harvest of public tweet streams on a large scale.[2]

1.2. Motivation

It is obvious that both traditional and emerging open data have intrinsic and exclusive features, providing unique perspectives but suffering from constraints and bottlenecks at the same time. For example, census data covers a large-scale general population but fails to reflect the monthly and daily changes. Google Trends represent the normalized search interests but not the real search volumes.[3] Social media data, such as tweets, contains heterogeneous information but may introduce demographic biases (Ribeiro, Benevenuto, & Zagheni, 2020; Wang et al., 2019). These facts make it difficult to ensure a robust and accurate event investigation if only applying an individual open data source. The choice of open data can have a profound impact on what one can learn and derive. It is the same with what the folk tale "The Blind Men and the Elephant" teaches us — different perspectives lead to distinct points of view as the elephant could be recognized as a wall, snake, spear, tree, fan, or rope, depending upon where the blind men had touched.

When investigating a phenomenon, we think choosing a data source can be seen as choosing a lens for observations in physical sciences; a telescope and a microscope both allow us to observe, but two very different worlds. Here, we consider lenses that cover three levels of observations: macro, meso, and micro. A macro lens can allow us to look at a phenomenon from a distance, covering a large area, but not being very precise. A micro lens, on the other hand, can provide a more specific picture but may be prone to localized fluctuations. A meso lens falls in between these two. While each of these lenses has its relative pros and cons, scientists make choices about which one to use when a more meaningful picture emerges through a careful combination of some or all of these lenses. It seems to be challenging to identify appropriate data source candidates to build these lenses. Thanks to the ubiquitous open data, it offers an excellent opportunity to define and enable multiple lenses to look at events of interest through different eyes.

However, this notion of integrating multiple open-data lenses in a generalized and effortless way is still under-explored. To bridge such research gaps, in this article, we summarize and compare the characteristics of different open data on eleven aspects, including demographic biases, potential ethical concerns, accessibility, and others. Furthermore, we propose a universal and easy-to-use framework, incorporating multi-source open data retrieval, feature extraction, and the training and fusion of machine learning models, to investigate events of interest. Specifically, we instantiate census data as the macro lens because it offers an overall picture of a large area and a large population. Social media data serves as the micro lens to examine the detailed and diverse features of individuals in a timely manner. The aggregated search engine data, such as Google Trends, is selected as a meso lens because it usually summarizes daily searching patterns generated by a relatively large group of users. Our framework only requires users, who can be researchers and practitioners in industry, academia, and government, to provide event keywords, timelines, and locations by simply answering what, when, and where questions. According to the users' inputs, our framework retrieves open data from government websites, search engines, and social media respectively, and then conducts feature engineering and builds models automatically.

1.3. Research objectives

In this article, we aim to design, implement, and evaluate an easy-to-use open data driven framework that is able to model general real-world events through the unified open lenses of census data (macro lens), search logs (meso lens), and social data (micro lens) at different time granularities.

1.4. Contributions

Our approach leads to several major contributions. First, our approach is transparent, collaboration-friendly, and free from privacy issues. Today, people, media, and governments have become more careful about data privacy (Altman, Wood, O'Brien, Vadhan, & Gasser, 2015; Goldfarb & Tucker, 2012; Klasnja, Consolvo, Choudhury, Beckwith, & Hightower, 2009; Liu & Carter, 2018), making it hard to access private data even for research purposes. The European Union (EU) enforced General Data Protection Regulation (GDPR) to protect private data and to prevent illegal private data collections (Voigt & Von dem Bussche, 2017). Under this scenario, open data, which is easier to obtain and with less potential ethical concerns, becomes more important and sheds light on investigating any general topics instantly and transparently. In addition, as open data is transparent to everyone, our framework can facilitate reproducibility and replicability in scientific research. Finally, the proposed framework motivates and boosts the collaboration of researchers across industries, academics, and governments, because open data can be shared among diverse agencies without raising ethical concerns and copyright or other legal issues.

Second, the proposed framework's generalization and automation capabilities guarantee an instant investigation of general events and phenomena, such as disasters, sports events, and political activities, with low labor efforts. As long as researchers know which

[1] https://trends.google.com/trends/?geo=US.
[2] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.
[3] https://support.google.com/trends/answer/4365533?hl=en.

event to explore, when and where the event occurs, they can configure the related keywords, time periods, and regions of interest to launch their studies. For example, to study the 2020 United States presidential election, researchers only need to specify a group of related keywords (e.g., 2020 election USA), the election day of November 3, 2020, and the entire United States' geographic scope. Then our framework sets up the macro (census data), meso (search logs), and micro (social media data) lenses according to the given keywords, locations, and times. The detailed operations of data retrieval, feature extraction, and model training are hidden as black boxes without requiring any knowledge or effort from users. We believe our end-to-end approach can facilitate cross-domain research and benefit a broad scientific community because it lowers the bar of information seeking and deployment of machine learning models.

Third, instead of relying on a single lens with open data, our framework provides two fusion based (i.e., feature fusion and model fusion based) training mechanisms to unify multiple open-data lenses and merge different viewpoints, allowing researchers and practitioners to determine how to fuse the extracted demographic information, search behavior signals, and social media features. The first method concatenates all features extracted from different open data to train one machine learning model. The other method trains individual models using each lens independently and then combines them to rebuild an integrated model. The two mechanisms provide researchers with flexible options to consolidate perspectives of the studied events from multiple angles. When raw data or processed features from third-party data providers are available, both mechanisms are compatible with those external features. When only a model trained on private data is accessible, the model fusion based mechanism is also extensible to integrate it with other open data driven individual models.

Lastly, to demonstrate the usability and effectiveness of the proposed framework, we take the COVID-19 pandemic and Solar Eclipse of August 21, 2017 (see Appendix) as case studies to estimate how COVID-19 progressed across U.S. states and when the total eclipse occurred using individual and collective lenses. To be specific, we use COVID-19 (eclipse) related words as keywords to collect multi-source open data, including census data, Google Trends, and Twitter data, in 51 (12) U.S. states from April 4 to May 9, 2020 (on August 21, 2017). The census data, as a macro lens, provides an overall demographic distribution pattern, which covers all populations across U.S. states. Google Trends data, as a meso lens, indicates the aggregated search interests of the COVID-19 pandemic and the solar eclipse from millions of U.S. Google search engine users. Twitter data, as a micro lens, enables to take a closer look at the individuals' attitudes and behaviors on COVID-19 and the solar eclipse. Then we adopt the two aforementioned mechanisms to perform data fusion and to train seven types of regression models. For estimating daily confirmed COVID-19 cases and deaths, the best performance of models trained through our framework can beat those trained on expert-generated datasets in more than one-fourth of all 50 U.S. states and D.C. More importantly, our approach requires fewer labor efforts to collect, preprocess, and organize data and less domain knowledge to create data features. For the eclipse case study, we find that multi-lens models outperform any single-lens-driven models in 33% U.S. states. We believe the proposed framework is generalizable enough to study a wide range of real-world events and social phenomena using publicly available data.

2. Related work

For each type of open data incorporated in our framework, we summarize the literature regarding its applications and societal impacts. Accordingly, these works can be categorized into three groups: census data, open search logs, and open social media data. In addition, we highlight the difference between our research work and existing ones.

2.1. Census data

Census data usually encompasses a wide range of participants across large geographic areas. According to the U.S. Census Bureau,[4] census data can be used in up to 50 ways, such as decision making at all levels of government, forecasting future transportation needs for all segments of the population, and directing funds for services for people in poverty. In academia, census data is regularly used to represent a population base that samples are meant to represent. It provides a strong default to compare alternatives against and a baseline that approximations can be based on. As some of the following studies show, it can also be combined with other forms of data to offer a more accurate view than either source could alone.

Census data is widely used to ensure representative samples are chosen from the population. For example, Twenge and Joiner (2020) compared prevalence of anxiety and depressive disorders before and during the COVID-19 pandemic based on U.S. Census Bureau-administered national samples. Ogorzalek, Piston, and Puig (2020) merged income survey data with census data to examine the relationships between local income and vote choice among white voters. Linked with census data, national surveys were used to conduct a cross-sectional study to explore the relationship between neighborhood characteristics and the engagement in telehealth among older adults (Okoye, Mulcahy, Fabius, Burgdorf, & Wolff, 2021). Rahman et al. (2020) measured sentiment analysis using both census and Twitter data concerning reopening the country during the COVID-19 pandemic.

Census data also plays an important role in identifying and correcting demographic bias existing in other types of data. Mislove, Lehmann, Ahn, Onnela, and Rosenquist (2011) is one of the first to report the demographic differences between Twitter data and the real-world population. Jiang, Li, and Ye (2019) studied county-level population biases on Twitter by modeling and mapping the relationships between different demographic/socioeconomic factors and geo-tagged Twitter users. Ribeiro et al. (2020) investigated demographic differences between Facebook users and census data, and proposed correction factors to mitigate the existing demographic biases. Cui and He (2021) corrected sampling and socio-demographics bias of social media data by linking people's Twitter account with their Facebook accounts.

[4] https://2020census.gov/content/dam/2020census/materials/partners/2019-03/ccc-guide-d-1280.pdf.

2.2. Open search logs

Anonymous data from aggregated search logs is able to portray public interest in a topic during a given period of time; it also enables the prediction of emerging trends in real life. For example, an increase in searches for "unemployment benefits" would indicate a rise in the unemployment rate.

In the past decade, open search logs, such as Google Trends and Baidu Index,[5] have been studied extensively in a wide range of areas, such as tourism demand estimation (Bokelmann & Lessmann, 2019; Feng, Li, Sun and Li, 2019; Höpken, Eberle, Fuchs, & Lexhagen, 2019), stock price prediction (Liu, Peng, Hu, Dong, & Zhang, 2019; Maneejuk & Yamaka, 2019; Salisu, Ogbonna, & Adediran, 2021; Wilcoxson, Follett, & Severe, 2020), and unemployment rate and employment growth forecasts (Borup & Schütte, 2020; Mihaela, 2020; Mulero & García-Hiernaux, 2021; Nagao, Takeda, & Tanaka, 2019). In addition, Google Trends search data can be used in many applications for social good. For example, Chai et al. (2019) proposed an early warning system of suicide by combining data from Google Trends and features extracted from media reporting on suicide news. Thompson, Wilby, Matthews, and Murphy (2021) utilized Google Trends as a tool to evaluate flooding in places where formal hydrometeorological data were scarce.

Open search logs are also adopted widely in the field of health and medicine to assess public awareness and monitor the spread of many diseases, such as COPD (Boehm et al., 2019), measles (Santangelo et al., 2019), and HIV (Mahroum et al., 2019). After the outbreak of the COVID-19 pandemic, a large amount of literature chose Google Trends as a tool to study its spread from diverse perspectives, e.g., symptoms, risks, prevention, treatment, and impacts on humans. Cherry et al. (2020) used Google Trends to track the reduced sense of smell and taste (one typical symptom of COVID-19), determining its association with confirmed COVID-19 cases. Ayyoubzadeh, Ayyoubzadeh, Zahedi, Ahmadi, and R Niakan Kalhori (2020) used linear regression and long short-term memory (LSTM) models on data gathered from Google Trends to estimate the number of COVID-19 cases in Iran. Husnayain, Fuad, and Su (2020) demonstrated that Google Trends could potentially define the proper timing and location of risk communications for affected populations. Hong, Lawrence, Williams Jr, and Mainous III (2020) relied on Google Trends to measure population-level interest in telehealth and found the current telehealth capacity could not meet the increased population demand. Google Trends also enables studies that focus on well-being and mental health caused by lockdowns (Brodeur, Clark, Fleche, & Powdthavee, 2020, 2021) and physical distancing (Knipe, Evans, Marchant, Gunnell, & John, 2020).

2.3. Open social media data

Social media data enables an instant perspective on the public's immediate reaction to an event or a topic by analyzing users' online emotions, attitudes, and opinions. Some social media platforms (e.g., Twitter and Weibo) allow anyone to access unprotected postings by default, generating a great opportunity to collect and explore open social media data. Note that open social media data may suffer from various biases (e.g., political bias (Chun et al., 2019), media bias (Ribeiro et al., 2018), and gender bias (Usher, Holcomb, & Littman, 2018)) and noisy data generated by bots (Gilani, Farahbakhsh, Tyson, Wang, & Crowcroft, 2017).

Open social media data has demonstrated its capability and effectiveness in event and topic analysis across many fields, including politics (AlDayel & Magdy, 2021; Ruz, Henrìquez, & Mascareño, 2020; Stamatelatos, Gyftopoulos, Drosatos, & Efraimidis, 2020), natural phenomena (Feng et al., 2019), business (Choi, Yoon, Chung, Coh, & Lee, 2020; Ge, Qiu, Liu, Gu, & Xu, 2020), and disaster management (Lifang, Zhiqiang, Zhang, & Hong, 2020; Ragini, Anand, & Bhaskar, 2018; Zahra, Imran, & Ostermann, 2020). AlDayel and Magdy (2021) surveyed different methodologies and applications of stance detection on social media. Through Twitter follower network analysis, Stamatelatos et al. (2020) studied and assessed the possibility of deriving political affinity of particular entities. Based on five million English eclipse-mentioning tweets, Feng, Lu et al. (2019) discovered trending topics, monitored public sentiment, and identified human mobility patterns during the 2017 Solar Eclipse that crossed the United States. A recent work (Choi et al., 2020) conducted a systematic review on open social media analytics-based business intelligence studies. To explore social media reposting behaviors after natural disasters, Lifang et al. (2020) reported emotional responses of the public and how emotional factors and influential users affected the number of reposts. Ragini et al. (2018) collected Twitter data concerning floods in the India–Pakistan regions and proposed classification models to learn the needs of the people during the period of disaster. Zahra et al. (2020) proposed an approach to identify eyewitness messages on Twitter during disasters automatically.

Similar to open search logs, open social media data makes many studies about the COVID-19 pandemic possible. Jahanbin, Rahmanian, et al. (2020) combined Twitter data and web news to predict the COVID-19 outbreak. As a platform to express one's attitudes and emotions, Twitter generated high-quality data for tracking public perception (Boon-Itt & Skunkan, 2020; Dyer & Kolic, 2020; Saleh, Lehmann, McDonald, Basit, & Medford, 2021), mental health (Guntuku et al., 2020; Valdez, Ten Thij, Bathina, Rutter, & Bollen, 2020) and psychological fears (Singh et al., 2020) during the COVID-19 pandemic. Geo-tagged COVID-19 tweets enabled the investigation of human mobility dynamics (Huang, Li, Jiang, Li, & Porter, 2020). Another important research topic based on Twitter data is misinformation analysis (Kouzy et al., 2020; Sharma, Seo, Meng, Rambhatla, & Liu, 2020; Singh et al., 2020) and detection (Al-Rakhami & Al-Amri, 2020; Memon & Carley, 2020). Singh, Bansal et al. (2020) took a first look at the spatio-temporal dynamics of COVID-19 misinformation spread on Twitter. Memon and Carley (2020) reported COVID-19 misinformed communities demonstrated more denser and more organized than informed ones. Al-Rakhami and Al-Amri (2020) proposed an ensemble-learning-based framework for verifying the credibility of a vast number of tweets.

[5] https://index.baidu.com/.

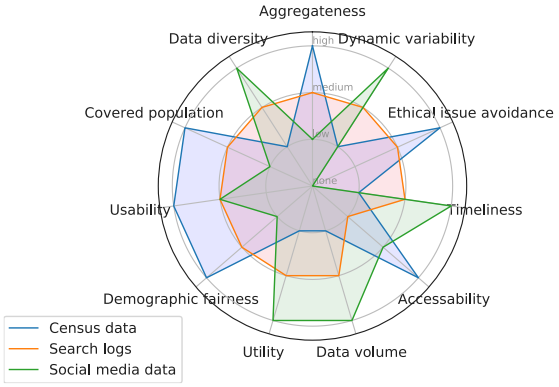


Fig. 1. Characteristics of census data, search logs data, and social media data.

2.4. Difference between our work and existing ones

This paper adds to the literature on exploring emerging events with open data in the following ways. First, instead of relying on a single (Bangwayo-Skeete & Skeete, 2015; Choi et al., 2020; Park, Lee, & Song, 2017; Zahra et al., 2020) or two (Choy, Cheong, Laik, & Shung, 2011, 2012; Rahman et al., 2020) open data sources, our framework incorporates three open lenses driven by data at different granularity levels. Second, unlike most existing approaches (AlDayel & Magdy, 2021; Feng, Lu et al., 2019; Ragini et al., 2018; Stamatelatos et al., 2020) focusing on specific tasks, our framework takes versatility and usability into account. In addition, we provide both feature fusion and model fusion based training mechanisms, whereas most of the existing works (Ayyoubzadeh et al., 2020; Choi & Varian, 2012; Ge et al., 2020; Park et al., 2017) only consider one of them.

3. Methodology

In this section, we first compare the characteristics of three open lenses driven by census data, search logs, and socia media data. Then we describe the overview of the proposed multi-lens framework, followed by the detailed designs and implementations of open data retrieval, feature extraction, and model training.

3.1. Characteristic comparisons of open lenses

It is obvious that different types of open data have unique characteristics, demonstrating intrinsic strengths and limitations at the same time. For example, census data covers a large number of populations and remains unchanged for years. By contrast, the Google Trends data changes daily, and Twitter data keeps dynamic in a real-time manner.

To determine the three open lenses' characteristics for comparison, we first refer to data quality assessment related literature and review commonly used data quality dimensions. Specifically, we reduce the 40 traditional quality dimensions for general data (Sidi et al., 2012), seven for open government data (Vetrò et al., 2016), and 19 for open data from the industry (Hassine & Clément, 2020), to nine important characteristics by removing trivial dimensions (e.g., data decay) and combining similar dimensions (e.g., timeliness and freshness). In addition, we add two new dimensions from the perspective of responsible data, namely ethical issue avoidance and demographic fairness, to develop eleven characteristics for comparison in total.

We investigate and summarize the characteristics of the three forms of open data on eleven aspects, as shown in Fig. 1. For instance, as census data contain almost all populations within a given region, its aggregation degree is high. Most, if not all, Internet users rely on search engines to filter and access information online, which leads to a medium aggregation for search logs data. Only registered and active users contribute to social media content, implying a low data aggregation. Accessibility of census data is high because it can be downloaded directly from government websites. But the accessibility of search log data is low because log data is usually protected, and only anonymous samples or aggregated data are available to the public. The social media data can be retrieved using official social media APIs, requiring authentication as a developer or researcher, and thus its accessibility is rated medium. For data diversity, we think social media data have a high amount of diversity because of multimedia content, while the other two lenses have a very limited number of data types.

3.2. Multi-lens framework overview

Aiming at establishing automatic end-to-end analysis flows for general events, the proposed multi-source open data framework consists of four components: event description, open data retrieval, feature extraction, and model training. As shown in Fig. 2, the first step is to describe the event under investigation by answering the following three questions: (i) which event it is; (ii) when it happens; and (iii) where it happens. These questions help determine event relevant keywords, time periods, and locations that guide

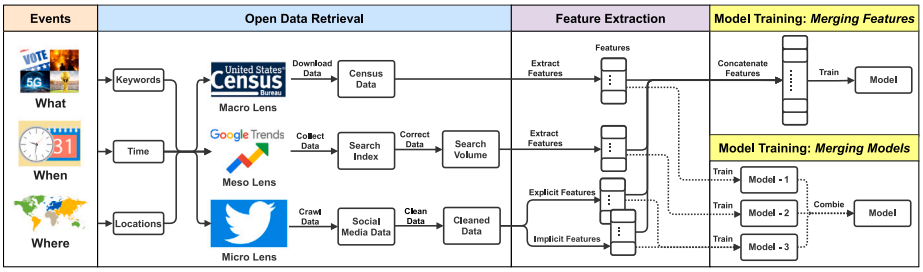


Fig. 2. Overview of the multi-lens framework with open data. For an event to be investigated, researchers and practitioners need to determine what it is, when, and where it occurs by specifying the keywords, time periods, and locations. Three open data sources, i.e., census data, Google search logs, and social media, are used to harvest data to serve as macro, meso, and micro lenses respectively. For each data source, we conduct feature engineering to extract features. Finally, feature fusion and model fusion based model training mechanisms are designed to estimate results.

the data collection. Specifically, the demographic data with the inferred time granularity (e.g., months and years) and geographic granularity (e.g., counties and states) are crawled from government census websites. Given keywords and locations, Google Trends allows both real-time and archived aggregated search behavioral data retrieval. Similarly, Twitter, on which the posts are publicly available by default, offers official streaming and search APIs to filter and harvest data by keywords and regions in real-time and offline manners.

After collecting data, we preprocess Google Trends data and tweets to improve the data reliability and accuracy by correcting data biases and filtering out bot generated content. Then feature engineering is performed on the three types of open data to prepare features for model training. For example, we extract population distributions of ages and races from census data, Google search indices from search logs, and semantic embeddings from tweets. Finally, two model training mechanisms are designed and implemented: (i) feature fusion based mechanism: training an overall model based on concatenated three lens features; and (ii) model fusion based mechanism: training a new regression model based on the regressors trained on individual open data independently. We describe the details of each step in the following subsections.

3.3. Open data retrieval

In this subsection, we present how to collect open data including census data, Google Trends data, and Twitter data. After collecting data, we preprocess Google Trends data and tweets to improve the data reliability and accuracy by correcting data biases and filtering out bots generated content.

3.3.1. Census data

As a traditional type of open data, census data is usually maintained and released at governmental websites, where the public can freely select and download data according to their needs. In our study, we collect the 2019 state population data by characteristics from the United States Census Bureau.[6]

3.3.2. Google trends data

Google Trends measures Google search interest in a particular topic in both real-time and offline manners. Given a time, location, and topic of interest, Google Trends returns a scaled and normalized index, ranging from 0 to 100 based on the topic's proportion to all searches on all topics, to represent the popularity of a specified topic.[7] Google Trends allows a single search term for one query. But the proposed framework is flexible to support more than one Google Trends index generated by multiple queries for one topic by treating each index as a single feature. Note that the index of search trends fails to reflect the actual search volume due to fluctuations in total search volumes from time to time. The value zero does not indicate no search volumes, but very low search volumes, i.e., the lowest interest within the examined period, since the data are normalized. Moreover, a value of 100 indicates the highest interest within the selected time frame.

To mitigate such data inaccuracy caused by normalization, we propose an algorithm to infer the actual search volume of a query at a location named *place* from its Google Trends index by adjusting the population-weighted search popularity of *place* generated within locations other than *place*. The main idea is to estimate the total search volumes at the location of *place*, which serves as a coefficient to correct the Google Trends index, by observing the search popularity of the term *place* at other locations. We take the estimation of search volume of one state in the U.S. as an example to explain how our data correction algorithm works. To be specific, the search volume V_s^t of *keyword* in state s at time t is calculated as follows.

$$V_s^t = I_s^t * \frac{\sum_{i \neq s}^{S} L_{is}^t * P_i^t}{\sum_{i \neq s}^{S} P_i^t} \tag{1}$$

[6] https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-detail.html.

[7] https://trends.google.com/trends/?geo=US.

where I_s^t is Google Trends index of *keyword* in state s at time t; L_{is}^t represents the Google Trends index of the term *state* s in state i at time t; P_i^t is the population of state i at time t; and the S is the set of all 50 states and the District of Columbia. For example, we infer the search volume of *Super Bowl* in Alabama on February 2, 2020. We first retrieve the Google search indices of the term *Alabama* from other 49 states and the District of Columbia on February 2, 2020. Then, we aggregate all these search indices of *Alabama* by normalizing them based on state populations. Finally, we use the above normalized value as a coefficient to correct the daily search index of *Super Bowl* in Alabama. In this paper, we adopt pytrend,[8] a Python-based Google Trends API tool, to collect two types of daily Google Trends data (i.e., I_s^t and L_{is}^t in Eq. (1)).

3.3.3. Twitter data

Twitter is chosen as the social media open data source in our study for the following reasons. First, with 192 million daily active users,[9] and 500 million newly generated tweets per day[10] Twitter is one of the most widely used social networks worldwide. Its popularity makes it a good crowdsourcing platform to collect people's first-hand data even from sparsely populated regions. Second, different from many other social networks, Twitter provides official streaming APIs[11] allowing developers to specify keywords and locations to sample and collect real-time data. The APIs capture and save tweets as JSON files, which contain an attribution of location indicating where tweets are posted. Thus, we can parse and extract tweet locations of interest for further analysis. Third, tweets are visible to everyone by default, reducing potential ethical issues and privacy violations in data collection.

Regarding the Twitter data, it is reasonable and necessary to clean the collected tweets because malicious bots generate a large amount of noisy data. Following Ljubešić and Fišer (2016), two types of Twitter accounts are regarded as bots in our study. The first one is the account that tweets very frequently and constantly for a long time, such as posting hundreds of tweets per day for months without stopping. The other type of bots demonstrates a regular and periodic tweeting behavior, e.g., tweeting every 30 min or three hours. We believe the detection and removal of bots will make it more precise to crowdsource human opinions and attitudes towards the event under study.

In addition to removing bots, many rudimentary natural language preprocessing techniques are deployed on raw data to clean up messy tweets. Specifically, we first tokenize raw tweets into individual components for further processing. Then we fix unintentional typing errors like typos and misspellings in raw tweets. Next, we make tweets more structured by removing redundant information such as @mentions, #hashtags, URLs, non-alphabetic words, and stop words. Finally, we lemmatize tweeting words to reduce them to their base forms, reducing the total number of unique words for further analysis.

3.4. Feature engineering

It is non-trivial to design and extract features of each lens because raw open data, especially social media data, is usually not well-structured and unsuitable for training machine learning models.

3.4.1. Census data

For census data, we mainly concentrate on demographic features, including population distributions with respect to ages and races. Take the state-level census data in the U.S. as an example. We reorganize the raw state population information, which is downloaded from the United States Census Bureau,[12] by both age and ethnicity. We project each state population into 18 five-year age groups ranging from under 5 years to 85 years and over to reduce data dimensions. Also, we aggregate state populations by race and produce six more population features: White, Black or African American, American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander, and Two or More Races. Thus, we extract 24 population features (18 age features and 6 race features) from census data for further model training.

3.4.2. Google trends data

As Google Trends data is well-structured and low-dimensional, we directly take the estimated search volume using Eq. (1) as the feature for model training. Note that besides Google Trends information, Eq. (1) has also incorporated population data.

3.4.3. Twitter data

Social media data usually needs extensive feature engineering to convert unstructured information into meaningful features. Both explicit and implicit Twitter features are constructed in the proposed framework. We take the most straightforward feature, the count of tweets posted in given periods and regions, as the explicit feature.

For implicit features, we consider sentiment and semantic distributions of tweets collected during the same time slot. Sentiment features mainly consist of emotion polarity and subjectivity of tweets. For each tweet, we use TextBlob[13] to calculate its polarity (a float within the range $[-1.0, 1.0]$) and subjectivity (a float within the range $[0.0, 1.0]$) scores. Then we build two distributions of

[8] https://pypi.org/project/pytrends/.
[9] https://s22.q4cdn.com/826641620/files/doc_financials/2020/q4/FINAL-Q4'20-TWTR-Shareholder-Letter.pdf.
[10] https://www.dsayce.com/social-media/tweets-day/.
[11] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data.
[12] https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-detail.html.
[13] https://textblob.readthedocs.io/en/dev/.

Table 1
Fine-tuned parameters for regression models.

Models	Parameters
Bayesian ridge	alpha_1, alpha_2, lambda_1, lambda_2, normalize
Ridge	alpha, normalize
Lasso LARS	alpha, normalize
Random Forest	n_estimators, bootstrap
SVR	kernel, C, epsilon
Linear	fit_intercept, normalize
K-Neighbors	n_neighbors, weights

sentiment polarity and subjectivity to represent implicit feature values by binning continuous emotion scores into discrete buckets, respectively.

We create a word embedding representation of aggregated tweets to construct semantic features. After removing mentions, hashtags, URLs, and stop words, we tokenize and lemmatize tweets. Then we conduct term frequency–inverse document frequency (TF–IDF) to figure out the important tweet tokens with a TF–IDF weight larger than a given threshold. On these tokens, we adopt the word2vec GloVe model (Pennington, Socher, & Manning, 2014), which is pre-trained on 2 billion tweets and 27 billion tokens, to infer a word embedding vector. Finally, we sum up TF–IDF weighted GloVe word embeddings of the selected tweet tokens to represent the semantic feature values.

3.5. Model training and evaluation metrics

We propose two model training mechanisms to incorporate insights from multiple open data sources. Our framework also supports automatic parameter tuning to train machine learning models.

3.5.1. Two model training mechanisms

Feature fusion based model training mechanism concatenates all features extracted from multiple open-data lenses to train an individual model. It is intuitive to merge and represent the multiple open-data perspectives through the combination of these features. We standardize all feature values before applying regression or classification models due to varying scales in different open lenses.

Rather than combining features of multi-source open data, model fusion based training mechanism combines models trained on single-source open data. The reason why we design such a training procedure is twofold. First, the best fitting model for each open lens can be found independently, implying more flexibility in model selection. Also, it becomes easy to incorporate external models pre-trained on third-party data. For example, we can extend our framework by integrating models trained on Microsoft Bing non-public search logs, without requesting access to the company's data or threatening customer privacy. If the research question is formatted as a regression problem, we take ideas of meta-analysis to estimate the final output. If it is a classification question, majority voting ensemble methods can be used to determine the final class label.

3.5.2. Model parameter tuning

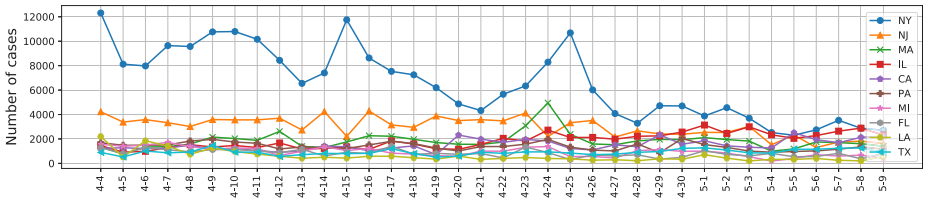
The two proposed mechanisms are flexible to support automatic parameter tuning when training regression and classification models. For a regression-based research problem, we train and evaluate the following regression models: Bayesian ridge regression, ridge regression, lasso LARS regression, random forest regression, Support Vector regression (SVR), linear regression, and k-neighbors regression. For each regression, we tune their parameters to find the best parameter combination with five-fold cross validation automatically. Because we implement all these regression models based on the Python scikit-learn machine learning library, we follow the parameter names in scikit-learn functions and list the tuned parameters of each regression model in Table 1. Similar parameter tuning is also supported for a classification-based research question.

3.5.3. Evaluation metrics

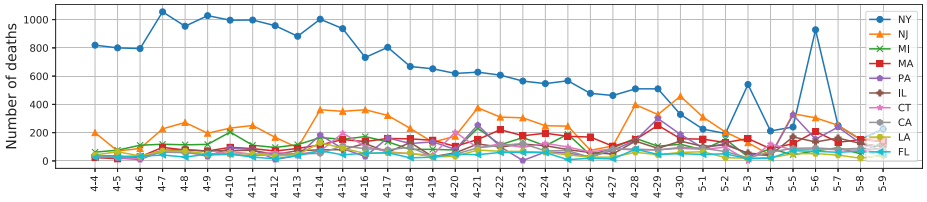
The proposed framework is compatible with various evaluation metrics to measure the performance of trained regression and classification models. For regression models, metrics such as R Square (R^2), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) can be selected to calculate the accuracy of models. To evaluate classification models, many evaluation options, such as Accuracy, Recall, F1-score, Categorical Cross-Entropy, and Area under the ROC Curve (AUC) (Green, Swets, et al., 1966; Hanley & McNeil, 1982) are available.

4. Evaluation

In this section, we take the COVID-19 pandemic as a case study to demonstrate the feasibility and effectiveness of our open data driven multi-lens framework. Another case study of the Solar Eclipse of August 21, 2017 is reported in Appendix B. We also qualitatively compare the proposed framework with existing works from six dimensions: data fusion, model fusion, timeliness, geo-awareness, generalizability, and automation.



(a) The number of daily confirmed COVID-19 cases in the top 10 states



(b) The number of daily reported COVID-19 deaths in the top 10 states

Fig. 3. The number of daily COVID-19 cases and deaths across U.S. states. We show the top 10 states with the most cases and deaths from April 4, 2020 to May 9, 2020 respectively. Among them, New York state and New Jersey reported most COVID-19 cases and deaths during that time.

4.1. COVID-19 case study

This subsection presents how to use the proposed framework to estimate COVID-19 daily cases and deaths in 50 U.S. states and D.C. in 2020.

4.1.1. Background and problem statement

As the COVID-19 pandemic swept over the world, it has had a widespread impact on people's daily lives and the whole society (Feng & Zhou, 2020; Singh & Singh, 2020; Torales, O'Higgins, Castaldelli-Maia, & Ventriglio, 2020). In our case study, we train regression models using open data to estimate the daily COVID-19 confirmed cases and deaths in U.S. states. The number of cases and deaths can be further used to calculate the case fatality ratio (deaths over cases), which is important to explore how the course of the pandemic progresses according to Johns Hopkins University[14] and World Health Organization.[15] We choose The New York Times COVID data repository[16] as the ground truth of reported COVID-19 cases and deaths. The top 10 states with most confirmed cases and deaths from April 4, 2020 to May 9, 2020 are illustrated in Figs. 3(a) and 3(b) respectively.

4.1.2. COVID dataset

We collect the census data from United States Census Bureau websites. We adopt pytrends[17] to collect two types of daily Google Trends data (i.e., I_s^t and L_{is}^t in Eq. (1)) across U.S. states from April 4, 2020 to May 9, 2020. Specifically, we use *COVID* and each state's name as keywords to collect the daily COVID Google Trends indices in all U.S. states. Then we calculate the daily search volumes of COVID using Eq. (1) for each state and regard them as feature values.

We use Twitter's Streaming APIs to collect geo-tagged COVID-19 related tweets across U.S. states. Considering people may pick up diverse words when tweeting the COVID-19 pandemic, we use a set of keywords, including *COVID19*, *COVID–19*, *coronapocalypse*, *Coronavid19*, *Covid_19*, *COVID-19*, *coronavirus*, *wuhan*, *corona*, and *nCoV*, to sample related tweets. Inspired by Ljubešić and Fišer (2016), Twitter accounts that generate more than 1500 tweets (tweeting around 50 times per day in our dataset) are recognized as bots. In addition, for users having more than 300 collected postings in our dataset, we analyze the time interval between two successive tweets. If the three most popular posting intervals dominate more than 90% of their all intervals, we treat such a user as a bot. After removing bots, we get 340,757 tweets in total across 50 states and D.C. from April 4, 2020 to May 9, 2020. The number of tweets collected in each state is summarized in Fig. 4. Note that some confounding variables (e.g., weather and local hospital capacity) that are not covered by the collected open data may also affect the daily COVID-19 cases and deaths.

[14] https://coronavirus.jhu.edu/data/mortality.
[15] https://www.who.int/news-room/commentaries/detail/estimating-mortality-from-covid-19.
[16] https://github.com/nytimes/covid-19-data.
[17] https://pypi.org/project/pytrends/.

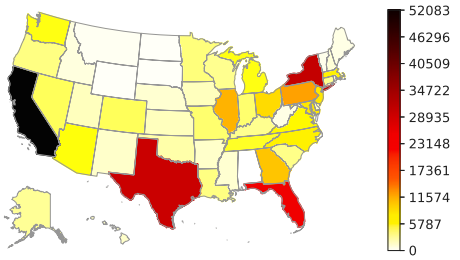


Fig. 4. The number of collected COVID-19 tweets in each state (April 4–May 9, 2020).

4.1.3. Model training and testing

Following Section 3.4, we extract features from multi-source open data as input, and treat the number of daily COVID-19 cases or deaths as output, to train and evaluate regression models. We split the dataset into 80% training and 20% testing data randomly to calculate the normalized RMSE.

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}}{\bar{y}} \tag{2}$$

where n is the number of testing samples; y_i is the ground truth; $\hat{y}$ is the estimated value by regression models; and $\bar{y}$ is the mean of all y_i.

4.1.4. Estimating daily COVID-19 cases and deaths in California

We take California as an example to demonstrate the estimation of daily confirmed COVID-19 cases and deaths using our framework. We summarize the normalized RMSE results of estimating the daily COVID-19 cases and deaths in California between April 4, 2020 and May 9, 2020 with different regression models and features in Fig. 5. Besides California, we illustrate the performance of estimating daily COVID-19 cases and deaths in moderate and least populous states, which demonstrate higher normalized RMSEs during the first wave of COVID-19, in Appendix A.

For both COVID-19 daily case and death estimations, we evaluate the performance of feature fusion based (see Figs. 5(a) and 5(c)) and model fusion based (see Figs. 5(b) and 5(d)) training mechanisms. The former merges features extracted from multi-source data to train models, while the latter concatenates the outputs of models that are trained on each open data independently as the new input features to train a final model. As mentioned in Section 3.5.2, we cover seven regression models in our experiments. For each regression model, we show their performance on the individual lens and merged lenses.

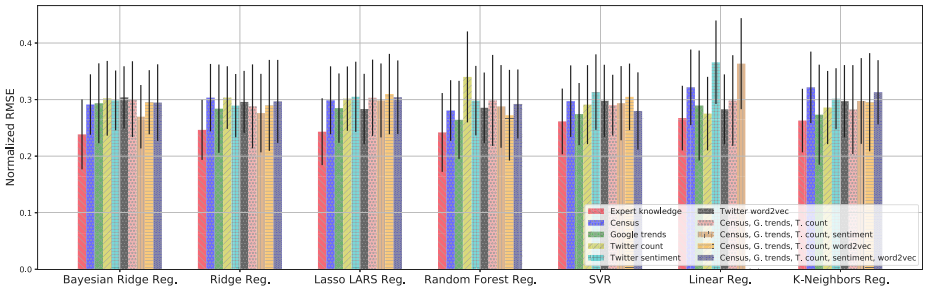
The first bar from the left of each regression model represents the results based on an expert-generated COVID-19 search trends symptoms dataset,[18] which includes aggregated, anonymized search trends for more than 400 symptoms, signs and health conditions, such as cough, fever and difficulty breathing. The rightmost four bars demonstrate the merged multi-source open data, i.e., census data, Google Trends data, and Twitter data, with various feature combinations. The rest demonstrate results of single-sourced open data with an individual feature. For Twitter data, we evaluate its performances with features of tweet counts, sentiment scores, and semantic embeddings (word2vec) respectively.

As expected, different features lead to varying accuracies even with the same regression models to investigate the same event (see the grouped bars labeled with the same regression models). In addition, the same feature demonstrates inconsistent performances using different regression models. For example, in Fig. 5(b), the feature combination of census data, Google Trends, and Twitter count, outperforms others on the k-neighbors regression, but this does not hold true for the other six models. We also observe that some features demonstrate relatively high normalized RMSEs and standard deviations when training regression models (see the blank bars, e.g., the merged features on the linear regression in Figs. 5(a) and 5(c)).

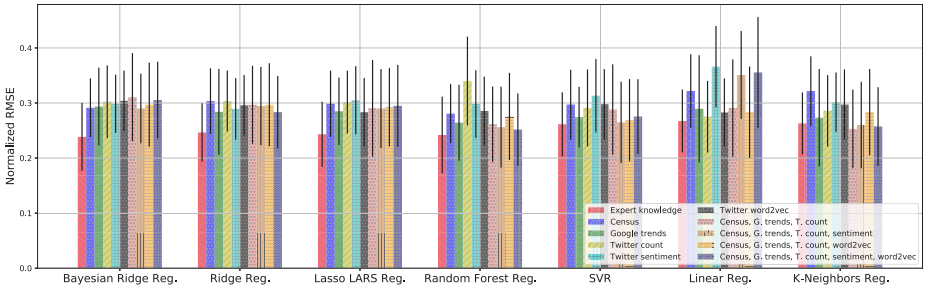
4.1.5. Individual lens versus merged lens

Since different types of open data have unique but incomplete characteristics, we intend to incorporate them together to investigate the event of interest. We think multiple open data lenses are more likely to reflect stable and real COVID-19 pandemic statuses than a single lens. Fig. 6 illustrates the open data that achieves the best performance when estimating daily COVID-19 cases and deaths with the two model training mechanisms in each U.S. state. Feature fusion based training mechanism demonstrates that merged lenses have a higher regression accuracy in 21 out of 50 states and D.C. for daily COVID-19 case and death estimations respectively (see Figs. 6(a) and 6(c)). For model fusion based training mechanism, the merged lens outperforms single lenses in 21 states and 14 states for estimating daily COVID-19 cases and deaths respectively (see Figs. 6(b) and 6(d)).

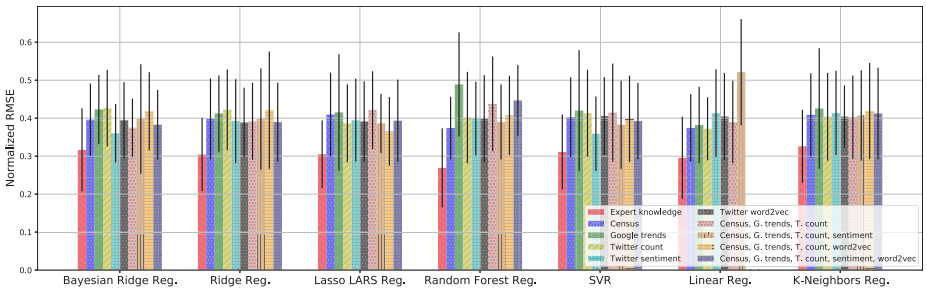
[18] https://pair-code.github.io/covid19_symptom_dataset/?country=NZ.



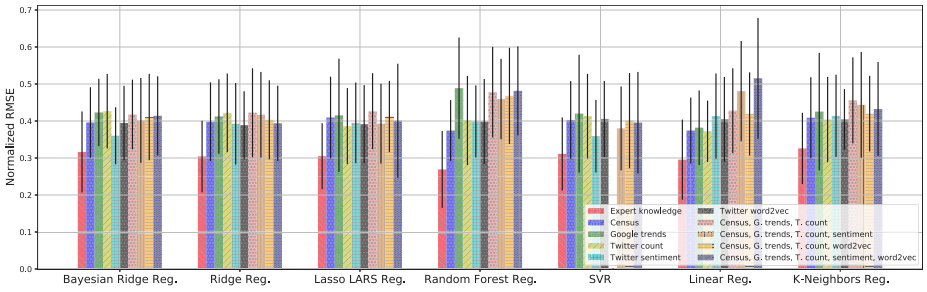
(a) Feature fusion based training mechanism - COVID-19 daily cases in California



(b) Model fusion based training mechanism - COVID-19 daily cases in California



(c) Feature fusion based training mechanism - COVID-19 daily deaths in California

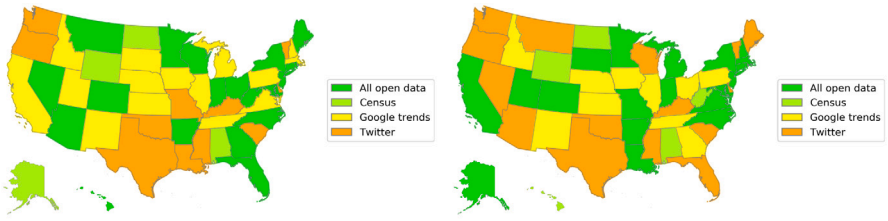


(d) Model fusion based training mechanism - COVID-19 daily deaths in California

Fig. 5. Regression results of daily COVID-19 cases and deaths in California with different model training mechanisms and feature fusions. The most left bar (colored as red) represents the results based on expert knowledge. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

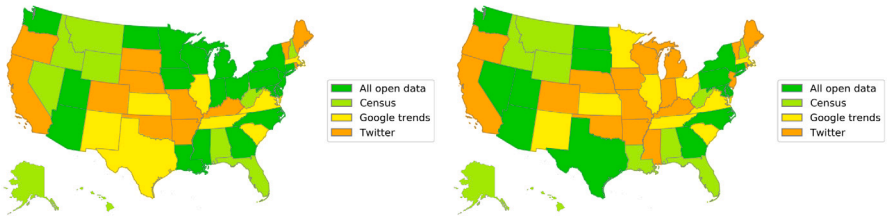
4.1.6. Expert knowledge versus open lenses

To show the proposed framework's effectiveness, we compare the best normalized RMSEs achieved by open lenses with that by an expert-generated dataset that includes more than 400 fine-tuned COVID-19 symptoms and related values. Using feature fusion based training mechanism, the best performance of open lenses can beat expert knowledge in 15 and 18 out of 50 states and D.C. for daily COVID-19 case and death estimations respectively (see Figs. 7(a) and 7(c)). For model fusion based training mechanism,



(a) Feature fusion based training mechanism - - daily cases

(b) Model fusion based training mechanism - daily cases



(c) Feature fusion based training mechanism - - daily deaths

(d) Model fusion based training mechanism - daily deaths

Fig. 6. Comparison of single and merged lenses in 50 U.S. states and the District of Columbia. Each state is colored by the lens that achieves the best performance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Qualitative comparison with existing works. ✓(Yes), ✗(No), ✢(Somewhat)

Approach	Data fusion	Model fusion	Real-time	Geo-aware	Generalized	Automatic
Ogorzalek et al. (2020)	✓	✗	✗	✓	✗	✢
Rahman et al. (2020)	✓	✗	✓	✓	✓	✓
Thompson et al. (2021)	✗	✗	✓	✓	✢	✓
Ayyoubzadeh et al. (2020)	✗	✓	✓	✢	✓	✓
Feng, Lu et al. (2019)	✗	✗	✓	✓	✓	✓
Singh, Singh et al. (2020)	✗	✗	✢	✢	✓	✓
Multi-lens framework	✓	✓	✓	✓	✓	✓

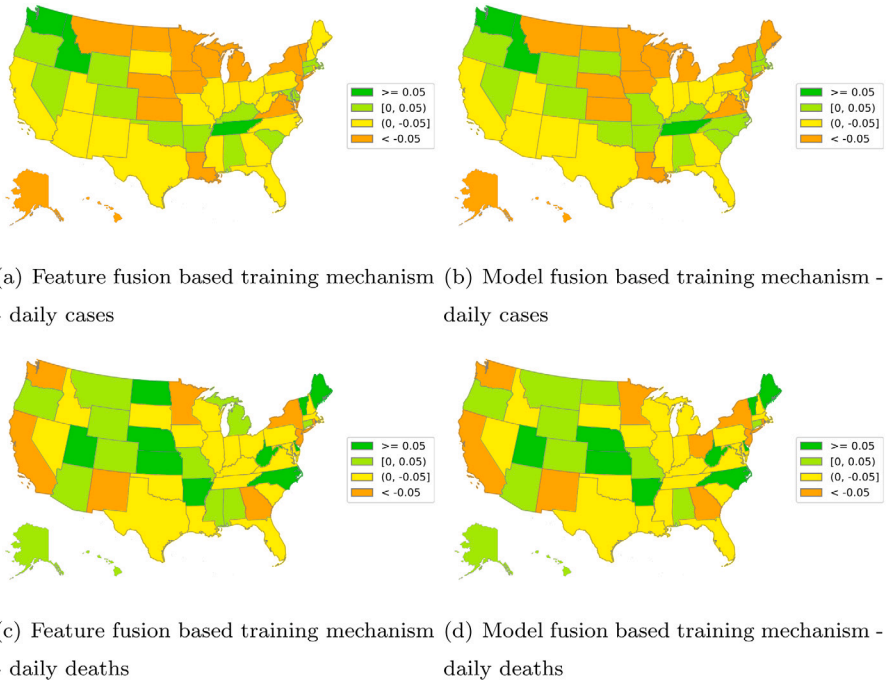
the open lenses perform better in 19 states and 16 states for estimating daily COVID-19 cases and deaths respectively (see Figs. 7(b) and 7(d)). We think our findings are of significant importance because our open data is transparent and public. Also, our open lenses are able to be used in real time and enable a timely estimation of the event. In addition, our findings demonstrate that the merged open data lenses are valuable and competitive with manually created datasets (here, Google's COVID-19 symptoms data with 400 fine-tuned features). Most importantly, they also require less labor effort than the human-generated datasets, making it practical to analyze new events as they emerge.

4.2. Qualitative comparison with existing works

In this subsection, we compare the proposed multi-lens framework with existing studies qualitatively. To be specific, we investigate six dimensions, namely data fusion, model fusion, timeliness, geo-awareness, generalizability, and automation, across existing studies and our framework, as shown in Table 2. Survey data based studies (e.g., Ogorzalek et al., 2020) are more likely to suffer from poor timeliness, low generalizability and automation. We also notice that not all works built upon Google Trends and social media data, such as Ayyoubzadeh et al., 2020 and Singh, Singh et al., 2020, take fine-grained locations explicitly. Many existing works, like Thompson et al., 2021 and Feng, Lu et al., 2019, do not take data fusion and model fusion into account. In contrast, the proposed multi-lens framework unifies diverse perspectives of multiple open data sources and offers timely and automatic investigations for general events and topics.

5. Discussion and implications

In this section, we first present the limitations on the accessibility of open data incorporated in our framework. Then we discuss our framework's implications in cross-disciplinary research and its compatibility with the models pre-trained on private datasets.



(a) Feature fusion based training mechanism - daily cases

(b) Model fusion based training mechanism - daily cases

(c) Feature fusion based training mechanism - daily deaths

(d) Model fusion based training mechanism - daily deaths

Fig. 7. The normalized RMSE difference of expert knowledge and open lenses in 50 U.S. states and the District of Columbia. Each state is colored by a RMSE range and the green color indicates open lenses outperforms expert knowledge. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1. Limitation on the accessibility of search logs and social media data

As commercial search engines, e.g., Google Search, and Microsoft Bing, harvest most search logs, it is difficult for outsiders to access huge amounts of desired search data due to business and customer privacy concerns. Therefore, the proposed framework has to rely on rough and low-precision search data aggregated and released by these companies. Although we have proposed an algorithm to improve the data quality of search logs, our framework will work more effectively and robustly if precise search data, such as Google search volumes, are available. Unlike Twitter, many other popular social networking platforms, e.g., Facebook, and Instagram, do not provide official streaming APIs to sample platform-wide social media content.

Their underlying business models are different from that of Twitter, which determines their generated social media data is not shared with the public. However, besides Twitter, if other social media data sources (even if they are anonymous and aggregated) are available, our framework can mitigate the sparsity of data in least-densely populated regions, promising a more generalized estimation and prediction of the events under investigation.

5.2. Applications and implications in cross-disciplinary research

The openness and high usability of the proposed framework make it easy to be adopted in cross-disciplinary research studies. First, all data used in the framework is open data that is accessible to researchers with very low efforts. Second, the entire framework, including data collection, feature extraction, and model training, is open-source, allowing researchers to redevelop, revise, and customize each component according to their needs. Third, it is easy-to-use for those who have minimal programming skills because they only need to answer the three W's questions (what, when, and where) of an event to be investigated. Then our open-source data collecting scripts will automatically download the corresponding census data, search logs, and social media data. When the data is ready, feature extractions and model training in the following steps can also be performed automatically.

We think the proposed framework benefits a broad research community from different disciplines and domains. For example, researchers in public transportation can take advantage of our framework to explore the emerging transportation systems, e.g., shared dockless electric scooter, in given cities and time periods. The presented framework can also be used to monitor and analyze time-sensitive social and political events, for example, the 2020 United States presidential election. Recall that Google Trends data is updated daily and Twitter data can be retrieved in a real time manner. Neither of the two examples require intensive efforts in collecting data nor huge budgets.

5.3. Compatibility with pre-trained models using private datasets

In many scenarios, it is impractical to call for companies and organizations to make their raw data open access, even for research purposes. Instead, it is more acceptable and reasonable to inquire whether they can release pre-trained models minimizing the risk of customers' personal information leak. On the one hand, the raw data never leaves company devices, and the access to the raw information is restricted to authorized personnel only. On the other hand, companies can design and provide Application Programming Interface (API) services allowing external users to submit queries to retrieve pre-trained model outputs regarding specified time periods and regions.

We can integrate such a pre-trained model into the proposed framework by treating it as an individual model in the model fusion based training mechanism (see Fig. 2). Along with other models trained independently on open data, we concatenate their outputs as the new input features to train a final model. Thus, the lens learned from private data is merged seamlessly into the existing workflow, making our approach more robust at no cost of privacy violations and ethical issues. From another perspective, agencies that hold exclusive data can also leverage our framework partially or entirely to incorporate open lenses into their internal model training operations.

6. Conclusion

Open data is playing an increasingly important role in many applications and services for social good, such as decision making and public opinion investigation. The choice of open data can have a profound impact on what one can learn and derive. Choosing a data source can be seen as choosing a lens for observations in physical sciences; a telescope and a microscope both allow us to observe, but two very different worlds. Here, we consider lenses that cover three levels of observations: macro, meso, and micro. A macro lens can allow us to look at a phenomenon from a distance, covering a large area, but not being very precise. A micro lens, on the other hand, can provide a more specific picture but may be prone to localized fluctuations. A meso lens falls in between these two. While each of these lenses has relative pros and cons, scientists make choices about which one to use when a more meaningful picture emerges through a careful combination of some or all of these lenses.

Although many studies integrating multi-source open data have been presented continuously over the past decade, most of them focus on less than three open data sources and are not designed for effortlessly investigating universal events. To bridge such research gaps, in this article, we summarize and compare the characteristics of different open data on eleven aspects, including demographic biases, potential ethical concerns, accessibility, and others. Furthermore, we propose a universal and easy-to-use framework, incorporating multi-source open data retrieval, data feature extraction, and machine learning model training and fusion, to investigate events of interest. Specifically, our framework only requires users, who can be researchers and practitioners in industry, academia, and government, to provide event keywords, timelines, and locations by simply answering what, when, and where questions. According to the users' inputs, our framework retrieves open data from government websites, search engines, and social media respectively, and then conducts feature engineering and builds models automatically. We take the COVID-19 pandemic (a social phenomenon with a time granularity of one day) and Solar Eclipse of August 21, 2017 (a natural phenomenon with a time granularity of 30 min) as case studies to demonstrate the usability and effectiveness of the proposed framework.

CRediT authorship contribution statement

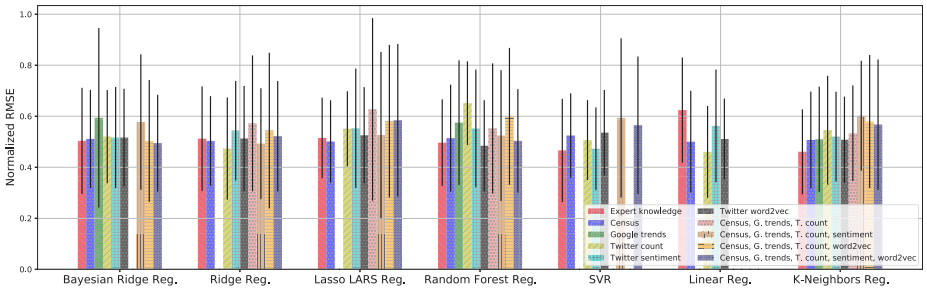
Yunhe Feng: Conceptualization, Methodology, Software, Data curation, Validation, Writing – original draft. **Chirag Shah:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Appendix A. Estimating daily COVID-19 cases and deaths in less populous states

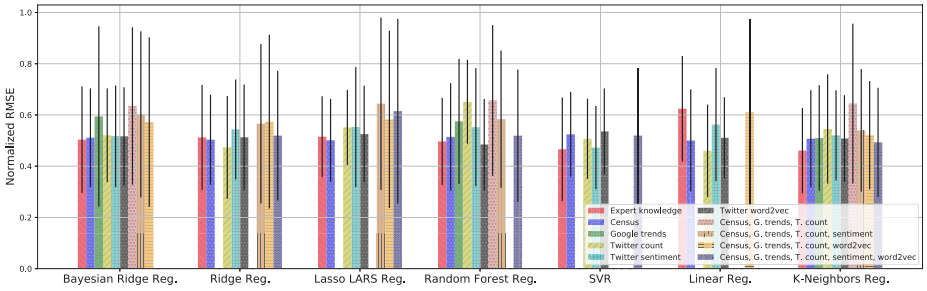
Besides the most populous state, i.e., California, we illustrate regression results of estimating daily COVID-19 cases and deaths in two additional states – Kentucky and Wyoming – with smaller populations. Kentucky ranks as 26th most populous state among 50 U.S. states and D.C. and Wyoming is the least populous state. The results of Kentucky and Wyoming are illustrated in Figs. A.9 and A.8 respectively.

Appendix B. Eclipse case study

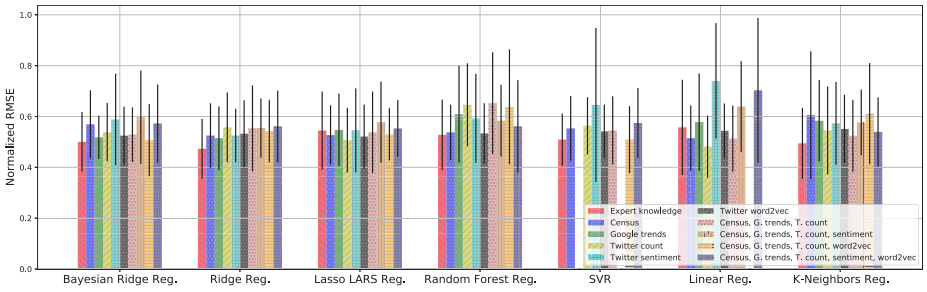
This subsection presents a case study of Total Solar Eclipse 2017 to investigate natural phenomena. It implies the potential usage of the proposed framework in other natural events such as natural hazards and disasters.



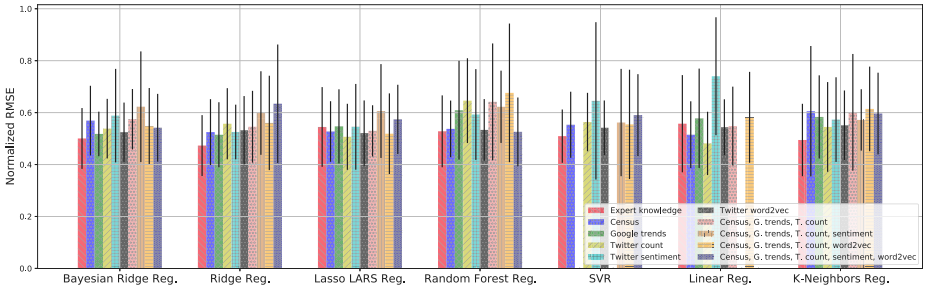
(a) Feature fusion based training mechanism - COVID-19 daily cases in Kentucky



(b) Model fusion based training mechanism - COVID-19 daily cases in Kentucky

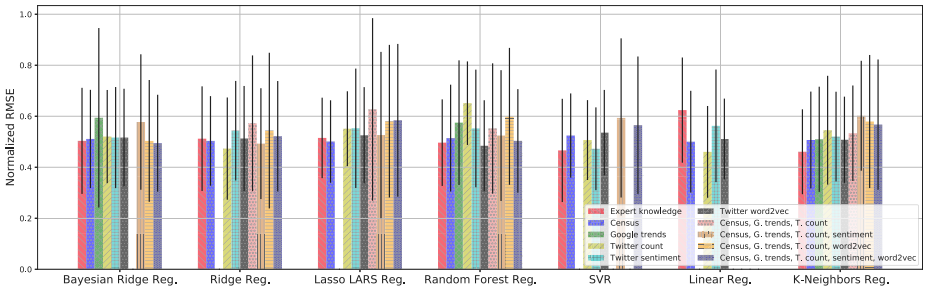


(c) Feature fusion based training mechanism - COVID-19 daily deaths in Kentucky

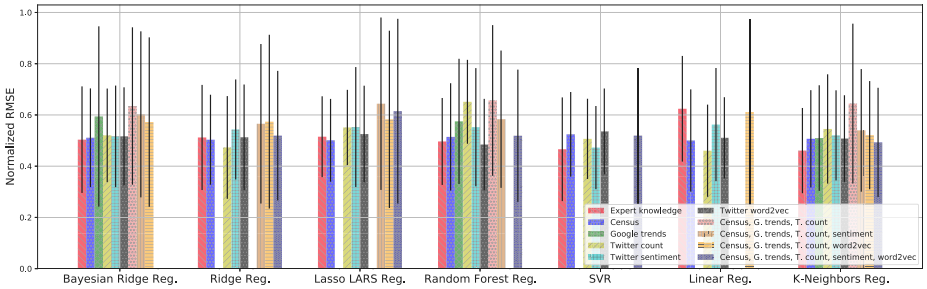


(d) Model fusion based training mechanism - COVID-19 daily deaths in Kentucky

Fig. A.8. Regression results of daily COVID-19 cases and deaths in Kentucky with different model training mechanisms and feature fusions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Feature fusion based training mechanism - COVID-19 daily cases in Wyoming



(b) Model fusion based training mechanism - COVID-19 daily cases in Wyoming

Fig. A.9. Regression results of daily COVID-19 cases in Wyoming with different model training mechanisms and feature fusions. We did not train regression models for daily deaths in Wyoming because only seven COVID-19 deaths were reported during April 4, 2020 and May 9, 2020.

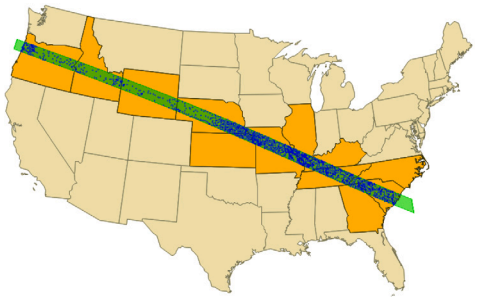


Fig. B.10. Path of the eclipse shadow across the U.S.[19]; The narrow track is the path of totality; small blue dots inside the narrow track represent cities. The 12 states where total solar eclipse was visible are highlighted (Feng, Lu et al., 2019). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

B.1. Background and problem statement

The coast-to-coast natural phenomenon of Solar Eclipse of August 21, 2017 crossed the continental U.S. (see Fig. B.10), attracting wide attention from the entire continent. According to The Washington Post,[20] nine in ten adults in the U.S. watched this total eclipse. We adopt the proposed framework to estimate when the totality occurred in 12 states lying within the path of totality.

We formulate the estimation of solar totality occurring time (with a time granularity of 30 min) as a regression problem. The open data within half an hour, e.g., 8:00 am–8:30 am and 9:30 am–10:00 am, are aggregated together. Take Oregon as an example. The solar totality was visible in Oregon between local time 10:15am and 10:27am on August 21, 2017. Thus we assign a $y = 0$ to the time slot 10:00 am–10:30 am for Oregon representing the totality occurred within this time slot, $y = 1$ to the time slot 9:30 am–10:00 am representing totality would occur 30 min later, and $y = 2$ to the time slot 9:00 am–9:30 am representing totality would occur one hour ($2 * 30$ min) later. Besides the totality time slot, we consider 14 such time slots for each state.

[19] http://robslink.com/SAS/democd94/eclipse_2017.htm.

[20] https://wapo.st/2Ok4G2D.

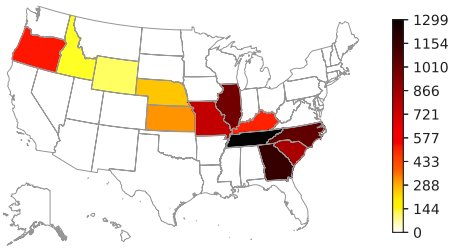


Fig. B.11. The number of eclipse-mentioned tweets in each state (August 21, 2017).

B.2. Eclipse dataset

As mentioned in Section 3.3.1, we download official census data from United States Census Bureau websites. We collect two types of hourly Google Trends data for states of interest. The first type of data is the Google Trends indices with a keyword of *eclipse*, i.e., I_s^t in Eq. (1). The other type is the L_{is}^t, the Google Trends indices inside state i with a keyword of state s. Then we estimate the search volume using Eq. (1).

We use Twitter's Streaming APIs to crawl English tweets containing keyword *eclipse* from the entire duration of the event (from Aug. 20, 2017 to Aug. 24, 2017). Following the bot detection approach (Ljubešić & Fišer, 2016), we conceived the two types of Twitter users as bots: (i) those who post more than 100 eclipse-tagged tweets per day; (ii) those who post more than 25 eclipse-tagged tweets per day and the top three most frequent posting intervals cover at least their 90% tweets. In our study, we focus on 8183 tweets posted from 12 states within 15 half an hour time slots before and during the totality time. The state-wise tweet counts are illustrated in Fig. B.11.

B.3. Model training and testing

We extract features from census data, Google Trends data, and Twitter data, as described in Section 3.4. The corresponding ground truth is set as mentioned in Appendix B.1. When training and evaluating regression models, we split the dataset into 80% training and 20% testing data randomly to calculate the normalized RMSE. We repeat the training process 1000 times to calculate the average normalized RMSE and its standard deviation.

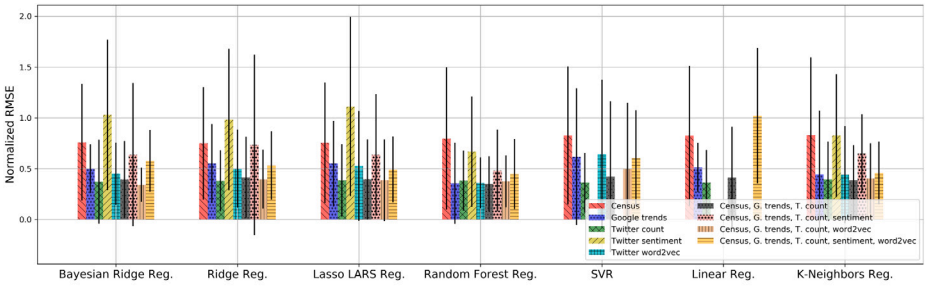
B.4. Eclipse totality forecast in Oregon

Both feature fusion and model fusion based training mechanisms (see Fig. 2) are flexible to adopt diverse regression models including Bayesian ridge regression, ridge regression, Lasso LARS regression, support vector regression (SVR), linear regression, random forest regression, and k-neighbors regression. We summarize the normalized RMSE results of estimating the eclipse totality in Oregon between 3:00 am and 10:30 am on August 21, 2017 with different regression models and features in Fig. B.12. To make the figures more readable, the bars with a standard deviation larger than 2 are not plotted.

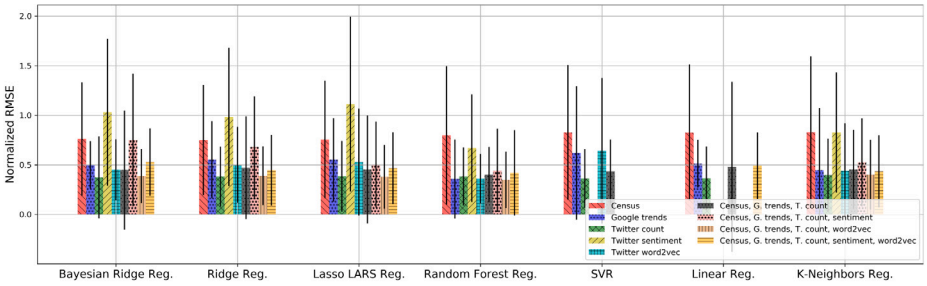
It is reasonable that the census lens demonstrates a poor performance on different regression models adopted in the two model training mechanisms because it always keeps unchanged from time to time. We can also observe that the social data lens empowered by Twitter sentiment shows a high normalized RMSE, indicating it contributes little to the model's precision. In contrast, Twitter count and two merged lenses (census data-Google Trends-Twitter count and census data-Google Trends-Twitter count-Twitter word embedding) perform very well.

B.5. Individual lens versus merged lens

Since different types of open data have unique but incomplete characteristics, we intend to incorporate them together to investigate the event of interest. We think multiple open data lenses are more likely to be stable and accurate than a single lens. Fig. B.13 highlights open lenses achieving the best performance in the 12 U.S. states lying within the path of totality. Feature fusion and model fusion based training mechanisms demonstrate that merged lenses achieve a higher regression accuracy in 33.33% and 41.67% states respectively. Google Trends lens performs the best in 58.33% and 50% states in Figs. B.13(a) and B.13(b). Note that Google Trends lens also incorporates the census data when inferring its search volume from the search index.

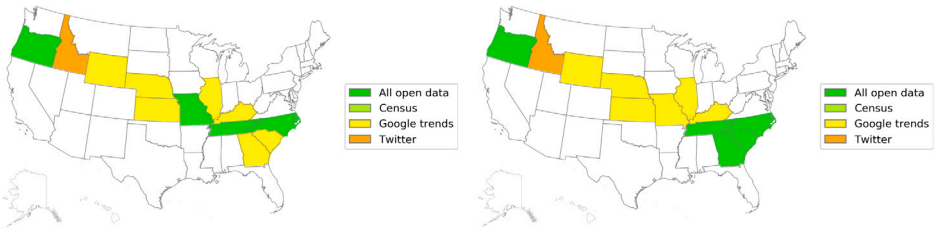


(a) Feature fusion based training mechanism - eclipse totality forecast in Oregon



(b) Model fusion based training mechanism - eclipse totality forecast in Oregon

Fig. B.12. Regression results of eclipse totality forecast in Oregon with different model training mechanisms and feature fusions.



(a) Feature fusion based training mechanism (b) Model fusion based training mechanism

Fig. B.13. Comparison of single and merged lenses in 12 states on the path of the eclipse totality. Each state is colored by the lens that achieves the best performance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Al-Rakhami, M. S., & Al-Amri, A. M. (2020). Lies kill, facts save: detecting COVID-19 misinformation in twitter. *IEEE Access*, *8*, 155961–155970.

AlDayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management*, *58*(4), Article 102597.

Altman, M., Wood, A., O'Brien, D. R., Vadhan, S., & Gasser, U. (2015). Towards a modern approach to privacy-aware government data releases. *Berkeley Technology Law Journal*, *30*(3), 1967–2072.

Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & R Niakan Kalhori, S. (2020). Predicting COVID-19 incidence through analysis of google trends data in Iran: Data mining and deep learning pilot study. *JMIR Public Health Surveill*, *6*(2).

Bangwayo-Skeete, P., & Skeete, R. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, *46*, 454–464.

Boehm, A., Pizzini, A., Sonnweber, T., Loeffler-Ragg, J., Lamina, C., Weiss, G., et al. (2019). Assessing global COPD awareness with Google Trends. *European Respiratory Journal*, *53*(6).

Bokelmann, B., & Lessmann, S. (2019). Spurious patterns in Google Trends data-An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management*, *75*, 1–12.

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, *6*(4), Article e21978.

Borup, D., & Schütte, E. C. M. (2020). In search of a job: Forecasting employment growth using google trends. *Journal of Business & Economic Statistics*, 1–15.

Brodeur, A., Clark, A. E., Fleche, S., & Powdthavee, N. (2020). Assessing the impact of the coronavirus lockdown on unhappiness, loneliness, and boredom using Google Trends. arXiv preprint arXiv:2004.12129.

Brodeur, A., Clark, A. E., Fleche, S., & Powdthavee, N. (2021). COVID-19, lockdowns and well-being: Evidence from Google Trends. *Journal of Public Economics*, *193*, Article 104346.

Chai, Y., Luo, H., Zhang, Q., Cheng, Q., Lui, C. S., & Yip, P. S. (2019). Developing an early warning system of suicide using Google Trends and media reporting. *Journal of Affective Disorders*, *255*, 41–49.
Cherry, G., Rocke, J., Chu, M., Liu, J., Lechner, M., Lund, V. J., et al. (2020). Loss of smell and taste: a new marker of COVID-19? Tracking reduced sense of smell during the coronavirus pandemic using search trends. *Expert Review of Anti-Infective Therapy*, *18*(11), 1165–1170.
Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *Economic Record*, *88*(s1), 2–9.
Choi, J., Yoon, J., Chung, J., Coh, B.-Y., & Lee, J.-M. (2020). Social media analytics and business intelligence research: A systematic review. *Information Processing & Management*, *57*(6), Article 102279.
Choy, M., Cheong, M. L. F., Laik, M. N., & Shung, K. P. (2011). A sentiment analysis of Singapore presidential election 2011 using Twitter data with census correction.
Choy, M., Cheong, M., Laik, M. N., & Shung, K. P. (2012). US presidential election 2012 prediction using census corrected Twitter model.
Chun, S. A., Holowczak, R. D., Dharan, K., Wang, R., Basu, S., & Geller, J. (2019). Detecting political bias trolls in Twitter data. In *WEBIST* (pp. 334–342).
Cui, Y., & He, Q. (2021). Inferring Twitters' socio-demographics to correct sampling bias of social media data for augmenting travel behavior analysis. *Journal of Big Data Analytics in Transportation*, 1–16.
Desouza, K. C., & Smith, K. L. (2014). Big data for social innovation. *Stanford Social Innovation Review*, *12*(3), 38–43.
Dyer, J., & Kolic, B. (2020). Public risk perception and emotion on Twitter during the Covid-19 pandemic. *Applied Network Science*, *5*(1), 1–32.
Feng, Y., Li, G., Sun, X., & Li, J. (2019). Forecasting the number of inbound tourists with Google Trends. *Procedia Computer Science*, *162*, 628–633.
Feng, Y., Lu, Z., Zheng, Z., Sun, P., Zhou, W., Huang, R., et al. (2019). Chasing total solar eclipses on twitter: Big social data analytics for once-in-a-lifetime events. In *2019 IEEE global communications conference (GLOBECOM)* (pp. 1–6). IEEE.
Feng, Y., & Zhou, W. (2020). Is working from home the new norm? an observational study based on a large geo-tagged covid-19 twitter dataset. arXiv preprint arXiv:2006.08581.
Ge, Y., Qiu, J., Liu, Z., Gu, W., & Xu, L. (2020). Beyond negative and positive: Exploring the effects of emotions in social media during the stock market crash. *Information Processing & Management*, *57*(4), Article 102218.
Gilani, Z., Farahbakhsh, R., Tyson, G., Wang, L., & Crowcroft, J. (2017). Of bots and humans (on twitter). In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017* (pp. 349–354).
Goldfarb, A., & Tucker, C. (2012). Shifts in privacy concerns. *American Economic Review*, *102*(3), 349–353.
Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics, Vol. 1*. New York: Wiley.
Guntuku, S. C., Sherman, G., Stokes, D. C., Agarwal, A. K., Seltzer, E., Merchant, R. M., et al. (2020). Tracking mental health and symptom mentions on twitter during covid-19. *Journal of General Internal Medicine*, *35*(9), 2798–2800.
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.
Hassine, S. B., & Clément, D. (2020). Open data quality dimensions and metrics: State of the art and applied use cases. In *International conference on business information systems* (pp. 311–323). Springer.
Hong, Y.-R., Lawrence, J., Williams Jr, D., & Mainous III, A. (2020). Population-level interest and telehealth capacity of US hospitals in response to COVID-19: cross-sectional analysis of Google search and national hospital survey data. *JMIR Public Health and Surveillance*, *6*(2), Article e18961.
Höpken, W., Eberle, T., Fuchs, M., & Lexhagen, M. (2019). Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden. *Information Technology & Tourism*, *21*(1), 45–62.
Huang, X., Li, Z., Jiang, Y., Li, X., & Porter, D. (2020). Twitter, human mobility, and COVID-19. arXiv preprint arXiv:2007.01100.
Husnayain, A., Fuad, A., & Su, E. C.-Y. (2020). Applications of google search trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. *International Journal of Infectious Diseases*, *95*, 221–223.
Jahanbin, K., Rahmanian, V., et al. (2020). Using Twitter and web news mining to predict COVID-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, *13*(8), 378.
Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258–268.
Jiang, Y., Li, Z., & Ye, X. (2019). Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and Geographic Information Science*, *46*(3), 228–242.
Klasnja, P., Consolvo, S., Choudhury, T., Beckwith, R., & Hightower, J. (2009). Exploring privacy concerns about personal sensing. In *International conference on pervasive computing* (pp. 176–183). Springer.
Knipe, D., Evans, H., Marchant, A., Gunnell, D., & John, A. (2020). Mapping population mental health concerns related to COVID-19 and the consequences of physical distancing: a Google trends analysis. *Wellcome Open Research*, *5*.
Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., et al. (2020). Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, *12*(3).
Lifang, L., Zhiqiang, W., Zhang, Q., & Hong, W. (2020). Effect of anger, anxiety, and sadness on the propagation scale of social media posts after natural disasters. *Information Processing & Management*, *57*(6), Article 102313.
Liu, D., & Carter, L. (2018). Impact of citizens' privacy concerns on e-government adoption. In *Proceedings of the 19th annual international conference on digital government research: Governance in the data age* (pp. 1–6).
Liu, Y., Peng, G., Hu, L., Dong, J., & Zhang, Q. (2019). Using Google trends and Baidu index to analyze the impacts of disaster events on company stock prices. *Industrial Management & Data Systems*.
Ljubešić, N., & Fišer, D. (2016). A global analysis of emoji usage. In *Proceedings of the 10th web as corpus workshop* (pp. 82–89).
Mahroum, N., Bragazzi, N. L., Brigo, F., Waknin, R., Sharif, K., Mahagna, H., et al. (2019). Capturing public interest toward new tools for controlling human immunodeficiency virus (HIV) infection exploiting data from Google Trends. *Health Informatics Journal*, *25*(4), 1383–1397.
Maneejuk, P., & Yamaka, W. (2019). Predicting contagion from the US financial crisis to international stock markets using dynamic copula with google trends. *Mathematics*, *7*(11), 1032.
McCombs, M., & Valenzuela, S. (2020). *Setting the agenda: Mass media and public opinion*. John Wiley & Sons.
Memon, S. A., & Carley, K. M. (2020). Characterizing covid-19 misinformation communities using a novel twitter dataset. arXiv preprint arXiv:2008.00791.
Mihaela, S. (2020). Improving unemployment rate forecasts at regional level in Romania using Google Trends. *Technological Forecasting and Social Change*, *155*, Article 120026.
Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, N. J. (2011). Understanding the demographics of twitter users.
Mulero, R., & García-Hiernaux, A. (2021). Forecasting Spanish unemployment with Google Trends and dimension reduction techniques. *SERIEs*, 1–21.
Nagao, S., Takeda, F., & Tanaka, R. (2019). Nowcasting of the US unemployment rate using Google Trends. *Finance Research Letters*, *30*, 103–109.
Napoli, P. M., & Karaganis, J. (2010). On making public policy with publicly available data: The case of US communications policymaking. *Government Information Quarterly*, *27*(4), 384–391.
Ogorzalek, T., Piston, S., & Puig, L. G. (2020). Nationally poor, locally rich: Income and local context in the 2016 presidential election. *Electoral Studies*, *67*, Article 102068.
Okoye, S. M., Mulcahy, J. F., Fabius, C. D., Burgdorf, J. G., & Wolff, J. L. (2021). Neighborhood broadband and use of telehealth among older adults: Cross-sectional study of national survey data linked with census data. *Journal of Medical Internet Research*, *23*(6), Article e26242.

Ortmann, J., Limbu, M., Wang, D., & Kauppinen, T. (2011). Crowdsourcing linked open data for disaster management. In *Proceedings of the terra cognita workshop on foundations, technologies and applications of the geospatial web in conjunction with the ISWC* (pp. 11–22). Citeseer.

Park, S., Lee, J., & Song, W. (2017). Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data. *Journal of Travel & Tourism Marketing*, *34*(3), 357–368.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Ragini, J. R., Anand, P. R., & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, *42*, 13–24.

Rahman, M. M., Ali, G. G. M. N., Li, X. J., Paul, K. C., & Chong, P. H. (2020). Twitter and census data analytics to explore socioeconomic factors for post-COVID-19 reopening sentiment.

Ribeiro, F. N., Benevenuto, F., & Zagheni, E. (2020). How biased is the population of facebook users? Comparing the demographics of facebook users with census data to generate correction factors. In *12th ACM conference on web science* (pp. 325–334).

Ribeiro, F. N., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., et al. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth international AAAI conference on web and social media*.

Ruz, G., Henrìquez, P., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. *Future Generation Computer Systems*, *106*, 91–104.

Saleh, S. N., Lehmann, C. U., McDonald, S. A., Basit, M. A., & Medford, R. J. (2021). Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter. *Infection Control & Hospital Epidemiology*, *42*(2), 131–138.

Salisu, A. A., Ogbonna, A. E., & Adediran, I. (2021). Stock-induced Google trends and the predictability of sectoral stock returns. *Journal of Forecasting*, *40*(2), 327–345.

Santangelo, O., Provenzano, S., Piazza, D., Giordano, D., Calamusa, G., & Firenze, A. (2019). SHORT PAPER Digital epidemiology: assessment of measles infection through Google Trends mechanism in Italy. *Annali di Igiene: Medicina Preventiva e di Comunita*, *31*, 385–391.

Sharma, K., Seo, S., Meng, C., Rambhatla, S., & Liu, Y. (2020). Covid-19 on social media: Analyzing misinformation in twitter conversations. arXiv preprint arXiv:2003.12309.

Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In *2012 international conference on information retrieval & knowledge management* (pp. 300–304). IEEE.

Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., et al. (2020). A first look at COVID-19 information and misinformation sharing on Twitter. arXiv preprint arXiv:2003.13907.

Singh, J., & Singh, J. (2020). COVID-19 and its impact on society. *Electronic Research Journal of Social Sciences and Humanities*, *2*.

Singh, P., Singh, S., Sohal, M., Dwivedi, Y. K., Kahlon, K. S., & Sawhney, R. S. (2020). Psychological fear and anxiety caused by COVID-19: Insights from Twitter analytics. *Asian Journal of Psychiatry*, *54*, Article 102280.

Stamatelatos, G., Gyftopoulos, S., Drosatos, G., & Efraimidis, P. S. (2020). Revealing the political affinity of online entities through their Twitter followers. *Information Processing & Management*, *57*(2), Article 102172.

Thompson, J. J., Wilby, R. L., Matthews, T., & Murphy, C. (2021). The utility of Google Trends as a tool for evaluating flooding in data-scarce places. *Area*.

Torales, J., O'Higgins, M., Castaldelli-Maia, J. M., & Ventriglio, A. (2020). The outbreak of COVID-19 coronavirus and its impact on global mental health. *International Journal of Social Psychiatry*, *66*(4), 317–320.

Twenge, J. M., & Joiner, T. E. (2020). US Census Bureau-assessed prevalence of anxiety and depressive symptoms in 2019 and during the 2020 COVID-19 pandemic. *Depression and Anxiety*, *37*(10), 954–956.

Usher, N., Holcomb, J., & Littman, J. (2018). Twitter makes it worse: Political journalists, gendered echo chambers, and the amplification of gender bias. *The International Journal of Press/Politics*, *23*(3), 324–344.

Valdez, D., Ten Thij, M., Bathina, K., Rutter, L. A., & Bollen, J. (2020). Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of twitter data. *Journal of Medical Internet Research*, *22*(12), Article e21418.

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325–337.

Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (DDPR). In *A practical guide, Vol. 10* (1st ed.). Cham: Springer International Publishing, Article 3152676.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., et al. (2019). Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference* (pp. 2056–2067).

Wilcoxson, J., Follett, L., & Severe, S. (2020). Forecasting foreign exchange markets using Google trends: Prediction performance of competing models. *Journal of Behavioral Finance*, *21*(4), 412–422.

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information Processing & Management*, *57*(1), Article 102107.