Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: DocuSigned by: Chen Lin 67E464B5830C497...

Chen Lin
Name

11/4/2024 | 3:40 AM EST
Date

Title Modeling Search Trends and Search Interests for Health

Author Chen Lin

Degree Doctor of Philosophy

Program Computer Science and Informatics

Approved by the Committee

DocuSigned by:
EugeneAgichtein
C81F95707F73454...

EugeneAgichtein
Advisor

DocuSigned by:
Carl Yang
577C2A6BC14A45D...

Carl Yang
Committee Member

DocuSigned by:
Joyce C Ho
34E5DE72DFD0456...

Joyce C Ho
Committee Member

DocuSigned by:
414A03C3BA644EB...

Jeremy Sarnat
Committee Member

Committee Member

Committee Member

Accepted by the Laney Graduate School:

_____

Kimberly Jacob Arriola, Ph.D, MPH
Dean, James T. Laney Graduate School

_____

Date

Modeling Search Trends and Search Interests for Health

By

Chen Lin
M.Sc., Emory University, GA, 2023
M.Sc., Georgia Institute of Technology, GA, 2017
B.A., Peking University, Beijing, 2016

Advisor: Eugene Agichtein, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2024

Abstract

Modeling Search Trends and Search Interests for Health
By Chen Lin

Tracking pollution exposure and infectious disease exposure is a significant public health challenge. While online search trends offer valuable insights for monitoring health-related issues, accurately capturing and predicting users' health concerns remains difficult. This challenge is compounded by the fact that users' search behavior changes frequently and varies across different sessions. Integrating additional data sources, such as geographical and environmental data, allows models to better account for external factors that compensate search data, providing a more comprehensive understanding of public health trends. Existing methods, although useful, do not fully leverage advanced neural networks and integrate multifaceted data for symptom prediction and intent categorization. To address this, my research develops a fundamental approach that leverages multimodal data integration and contextual learning across sequential, graph-based, and large language models, each tailored to specific health-related forecasting and intent categorization tasks.

The primary research questions addressed here are as follows:

RQ1: How can multimodal data sources, including online search trends, be effectively leveraged to forecast health symptoms related to environmental factors (e.g., air pollution) and infectious diseases?

RQ2: To enhance understanding of search trends and deliver responsive health information services, how can search intent for health-related queries be identified and anticipated by analyzing user interaction data, such as user click logs?

These research questions are addressed by processing various data sources using multimodal sequential learning, leveraging online search trends and integrating multi-modal data for enhanced forecasting models. Additionally, I employ unsupervised learning with limited annotation data to identify health-related search intents and apply supervised learning to anticipate health service seekers' needs by modeling queries with both consistent and varying intents across different sessions. To achieve this, the dissertation introduces key contributions such as integrating semantic information from search queries with search trends, utilizing cross-location information for improved pandemic forecasting, and presenting a novel fine-tuning method for adapting large language models to interpret health-related queries.

The proposed methods, validated on diverse real-world datasets, demonstrate significant advancements in health search modeling. These innovations directly contribute to better pollution exposure symptom monitoring, pandemic forecasting, and the accurate interpretation of complex user search behaviors. As a result, this work offers practical solutions to real-world challenges in public health surveillance and health information systems, ultimately paving the way for more responsive, data-driven health search services that can better meet users' evolving healthcare needs.

Modeling Search Trends and Search Interests for Health

By

Chen Lin
M.Sc., Emory University, GA, 2023
M.Sc., Georgia Institute of Technology, GA, 2017
B.A., Peking University, Beijing, 2016

Advisor: Eugene Agichtein, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2024

Acknowledgments

I would like to thank my advisor, Dr. Eugene Agichtein, for his invaluable guidance and support throughout my graduate studies. I am also grateful to my committee members, Dr. Joyce Ho, Dr. Carl Yang, and Dr. Jeremy Sarnat, for their constructive feedback and insightful suggestions, which have been instrumental in shaping my research. I extend my heartfelt thanks to my colleagues and friends at Emory University for their continuous encouragement and to my family for their unwavering love and support. Special appreciation goes to my lab mates in the EmoryIR Lab and my co-workers from Kaiser Permanente, whose support and camaraderie have enriched my academic journey. Additionally, I would like to acknowledge the project members of CGAP and my research collaborators for their contributions and cooperation. Finally, I am profoundly thankful for the financial support provided by the National Institutes of Health (NIH), Kaiser Permanente, and Emory University, without which this research would not have been possible. This funding has enabled me to pursue my research goals and contribute to advancing health information systems.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

With the rapid development of the Web, over half of the U.S. adults used the Internet to look for health or medical information in the latter half of 2022 [15]. Therefore, online search activity is a critical resource for researchers to understand the health-seeking behavior of the public. Consequently, it is necessary to model user search activity for understanding people's health needs in a crowd perspective [92]. Traditionally, crowd-based online health models relies on the social media and search engine logs to predict the health needs of the public or individuals, which could be caused by infectious diseases, environmental hazards or other specific health needs. However, those models are search pattern-based and ignore both the query semantic information and users' feedback, which may lead to a poor prediction performance. Given the vast amount of effort in understanding semantic information and users' interaction data with artificial intelligence, search models that can serve the public with more accurate health event and health need prediction are feasible and expected. In this dissertation, I discuss the novel online search-based health models, which are the search trend and search intent prediction models.

1.1 Online Health Monitoring

Online health monitoring focuses on utilizing user-generated data, such as web searches, social media posts, and global search data, to assess the public's health needs. This approach helps track emergent risks to public health, such as infectious disease outbreaks. For instance, search terms like "cough" and "flu" have been used to track disease outbreaks, demonstrating a correlation between search activity and the timing of disease spread.

The relationship between search terms like "cough" and "flu" can illustrate how these queries help in tracking disease outbreaks. Figure 1.1 illustrates the correlation between search trends for "cough" and "flu" As shown in the figure, there is a notable correlation between these search trends, suggesting that search trend signals can be effectively used for modeling health monitoring trends. By analyzing search trends, we can accurately detect the timing and severity of flu outbreaks, which enables proactive prevention measures. This demonstrates the utility of Google Trends data in real-time health monitoring and supporting public health initiatives.

In addition to users' search trends, other modalities of information can significantly enhance the modeling of user search behavior for online health monitoring. For example, geographic and mobility information are two critical modalities that can provide valuable insights into user search patterns. People's movement can significantly influence the spread of infectious diseases. Understanding users' geographic locations can help analyze health trends across different regions. Certain diseases may spread more prevalently in specific areas. Geographic data can reveal these patterns, offering valuable insights for public health monitoring. At the same time, studying people's movement behavior provides crucial clues about disease transmission. Information such as the range, frequency, and routes of human mobility can help predict and monitor the spread of diseases. For example, if a population frequently travels to a known epidemic area, the risk of infection in their home region may increase. Inte-

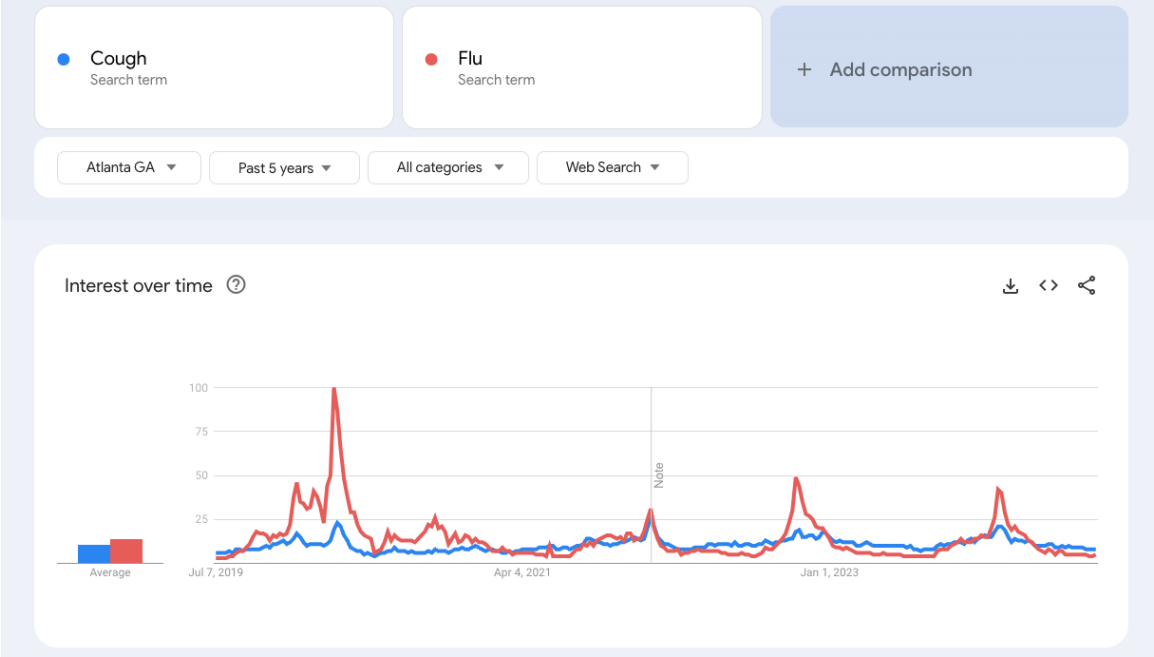


Figure 1.1: Correlation between search trends for "cough" and "flu" indicating potential for real-time health monitoring

grating additional information modalities such as geographic and mobility data can substantially enhance the modeling of user search behavior for online health monitoring. A well-known example of this is the online monitoring of COVID-19. As shown in Figure 1.2, the mobility and geoinformation graph is crucial for monitoring the spread and impact of COVID-19. This graph tracks the movement patterns of populations and correlates them with geographic information to provide insights into the virus's transmission dynamics.

By analyzing geographic locations and movement patterns, public health officials can gain valuable insights into the spread of infectious diseases across different regions. This data can reveal specific areas where certain diseases are more prevalent and help predict and monitor the transmission of diseases based on human mobility patterns. Thus, understanding and incorporating these modalities can significantly improve public health monitoring and response strategies.

In addition to leveraging search trends and geographic data for monitoring public

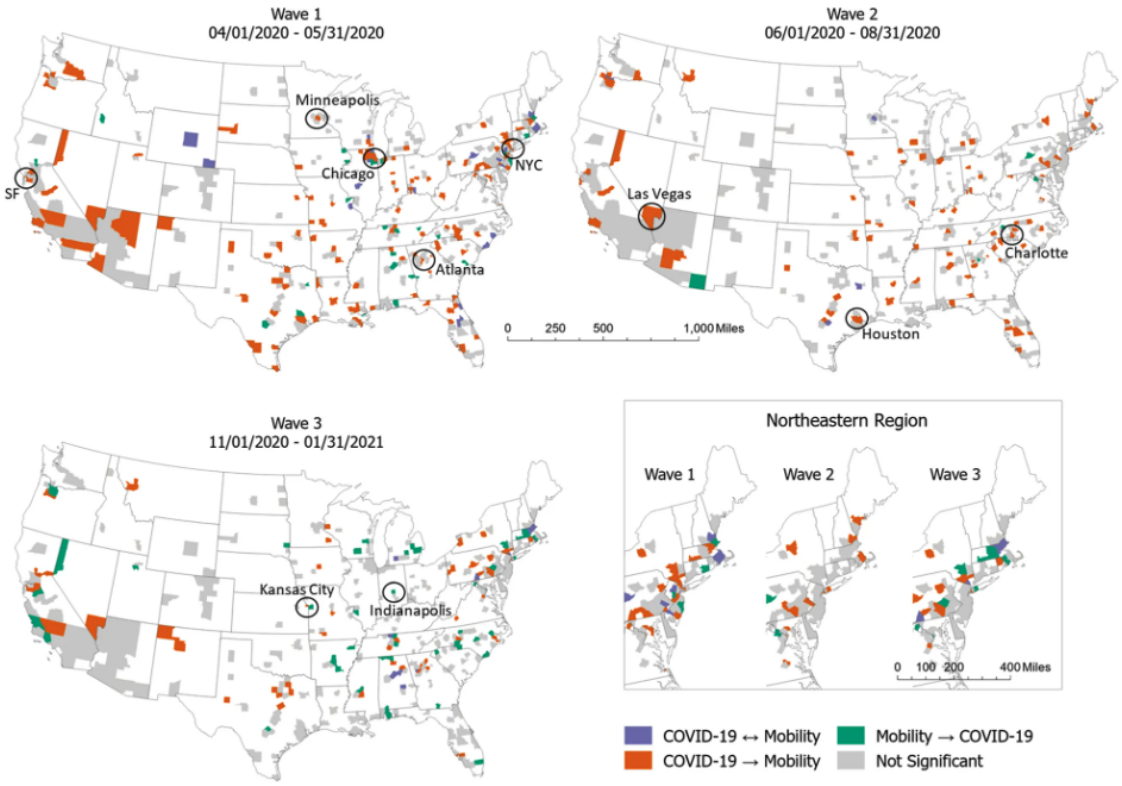


Figure 1.2: An example illustrating the importance of the mobility and geoinformation graph for COVID-19 monitoring [59]

health risks, it is equally important to enhance the understanding of individual health search queries. Health-related search intent recognition is particularly challenging due to the ambiguity of medical queries, where the same search term may have different meanings depending on the context. For instance, users may search for symptoms, treatments, or medical services with varied expectations, making it crucial for search engines to accurately interpret these queries. Traditional methods often struggle with this complexity, especially when queries are brief or vague. By incorporating query representation learning and session context, health search intent recognition can be significantly improved.

The connection between search trends and search intent modeling lies in their combined ability to offer a comprehensive view of public health needs. Search trends capture the "when" and "where" of health-related interests, revealing temporal and

spatial patterns that can be used to forecast public health trends. In contrast, search intent modeling delves into the "why" and "what" behind individual searches, uncovering the underlying needs or motivations that drive search behaviors. By combining trend analysis with intent recognition, this research enables a deeper understanding of not only what health issues are emerging but also why people are seeking specific health information at a given time. This integration allows for a robust methodology that can not only track health symptoms but also dynamically adjust to users' informational needs, thereby supporting proactive, responsive health monitoring systems.

1.2 Challenges in Health Search and Trend Modeling

To apply multiple sources to improve online health monitoring, there are several challenges and major steps involved.

1.2.1 Challenges in Modeling Search Trend for Health

Handling multi-modal data is crucial in health-related data (such as environmental data and search data) processing, especially given the common occurrence of incomplete or missing data from sources such as geographical networks or health databases. In many real-world scenarios, information and data are multi-modal, and features from different modalities are often seamlessly used together for classification or regression purposes. However, multi-modal sequential data is frequently incomplete due to broken sensors, failed data transmission, or low sampling rates. This challenge is exacerbated when trying to integrate multiple sources of data, requiring sophisticated machine learning models that can mitigate the effect of noise, inaccuracies, and incomplete data.

- There is the need to collect data from various sources like geographical location, mobile networks, social media, and health databases. For specific health events, the data collection process can be complex and time-consuming and multiple sources of data need to be integrated.

- I need to deal with incomplete, inaccurate, or noisy data which can lead to poor prediction performance. It is essential to employ advanced techniques to mitigate and reduce the impact of noise and inaccuracies in the data to ensure reliable and effective predictions.

- I need to develop innovative and advanced machine learning models that can effectively integrate and analyze data from multiple sources. These models should be able to handle the complexity and heterogeneity of the data and provide accurate and reliable predictions.

- I need to integrate the knowledge and expertise of domain experts to have a better understanding of the data and the health events being monitored. Specifically, I need to have a deep understanding of online search activity at individual online search activity and population levels to effectively model and predict health needs.

By addressing these challenges, I can improve the effectiveness of online health monitoring systems and provide more accurate predictions and insights into health trends and events. The specific objective of this study is to develop better, more sensitive measures of exposure and response to the myriad suite of atmospheric pollutants via online surveillance, using the Web search interest data as a proxy for the physical world experience.

1.2.2 Challenges in Health Search Intent Recognition

Another major challenge is the complexity of understanding user intent in health-related search queries. Search behavior related to health often spans multiple topics and can be ambiguous, making it difficult for models to capture the user's exact health needs, especially when considering session context or multi-label search classification. Accurately recognizing user health-related search intent presents another challenge. Queries often have ambiguous intent, and it is important to develop models capable of understanding these complex searches.

Building on the complexity of understanding user intent in health-related search queries, another significant challenge lies in the ambiguity and multi-dimensionality of these queries. Health-related searches often contain a blend of technical medical jargon and everyday language, making them difficult to interpret without adequate context. Additionally, these queries can span multiple intent categories, such as symptom inquiries, treatment research, or healthcare provider searches, requiring more sophisticated multi-label classification models to capture the full range of user needs. The same query can also hold different meanings based on the user's session history or current context, which further complicates the recognition of search intent. For example, a query like "online visit" may relate to scheduling a healthcare appointment in one instance, while in another, it could refer to seeking information about telemedicine platforms. These challenges highlight the need for models that can not only manage multi-label intent recognition but also incorporate session-based context to improve prediction accuracy and better meet users' health information needs.

1.3 Research Questions

Based on the above introductions, online search trends and search interest analysis are important because they provide insights into the health information-seeking be-

haviors of large populations. As the internet becomes the primary source for health information, search data helps researchers understand people's health interests and questions. Online search trends and interests hold great promise for health-related research purposes, as they can reveal patterns and preferences that may not be easily accessible through traditional methods. By utilizing this abundant data, researchers can more effectively meet the needs of those seeking health services, create focused interventions, and track the development of epidemics.

Understanding and modeling search trends and user behavior involves several challenges in the process. The first step is collecting search data from search engines, which includes the search query, time of the search, and user's geographic location. People search for different terms depending on the diseases being researched, with the biggest challenge being the retrieval of an exhaustive list of search queries for the disease of interest and the imputation of missing values. Data preprocessing is the second step, involving converting search terms to semantic representations and normalizing search trends for comparison across temporal and spatial resolutions. The biggest challenge in this stage is handling the co-linearity in search queries and generating and preprocessing multi-modal data as a supplement to search data. Building models on search trends and logs is the third step, where statistical methods, machine learning, and deep learning models are applied to identify trends and patterns in the data. The main challenge during this stage is incorporating multi-modal data to improve model performance and compensate for the disadvantages of online search data.

In parallel with analyzing search trends data, analyzing search click logs data is also crucial for understanding users' health interests, but it also comes with considerable challenges. User click logs data includes search terms and corresponding clicked URLs and documents. One challenge comes from the fact that search queries are often short and varied, making it difficult to interpret the semantic meaning behind

them. Additionally, health-related search intents are fine-grained, and the challenge exists in capturing the hierarchical structure of the intents from short search queries. Another challenge arises from modeling the large amount of search queries, where some queries have consistent search intents across different sessions, while others exhibit varying intents in different sessions. During the modeling process, I need to consider how to accurately interpret queries with consistent search intents as well as those with varying intents. Given the total amount of search queries is large, it is also essential to employ memory and computation-efficient algorithms to achieve my aims.

Based on the challenges as mentioned above, I propose two key research questions in this dissertation.

- RQ1. How can multimodal data sources, including online search trends, be effectively leveraged to forecast health symptoms related to environmental factors (e.g., air pollution) and infectious diseases?
    - RQ1.1. How can I process the data sources with missing values for sequential activity modeling?
    - RQ1.2. How can I leverage the online search trends and environmental multi-modal data and validate their performance on the forecasting model?
    - RQ1.3. How can I leverage the online search trends and infectious disease-related multi-modal data and validate their performance on the forecasting model?
- RQ2. To enhance understanding of search trends and deliver responsive health information services, how can search intent for health-related queries be identified and anticipated by analyzing user interactions, such as user click logs?
    - RQ2.1. How can I effectively identify and understand the intents of health-related search logs using unsupervised learning techniques when there is

limited annotation data?

- RQ2.2. How can I apply supervised learning approaches to anticipate the search needs of health service seekers and effectively model queries with both consistent and varying search intents across different sessions?

This dissertation addresses the aforementioned research questions by proposing solutions to enhance models for search trend analysis and health-related query comprehension. First, I address RQ1.1 by utilizing information from different modalities for multimodal seqenuential learning with missing values, as described in section 3. Second, I address RQ1.2 by integrating semantic information from search queries with search trends, enhancing prediction outcomes using a multimodal learning approach for data augmentation, as described in section 3. Third, I address RQ1.3 by incorporating cross-location information to improve prediction and enable timely forecasting of infectious diseases, as described in section 4. Finally, I present a novel fine-tuning method for adapting large language models (LLMs) to interpret health-related queries. I also use clustering and classification models to identify search intents and utilize session context to model queries with varying search intents across different sessions, addressing RQ2.1 and RQ2.2 as described in section 5.

1.4 Contributions and Dissertation Structure

The dissertation is structured into six major sections. The first section reviews related work, including existing studies of online health monitoring and recent advancements. The second second section dives deeper into the background of search trend and interest analysis, including the challenges and opportunities in the field. The next three sections explore the application of my proposed solutions to the real-world research questions, each containing an overview, methodology, experiments, and discussion. The final section presents the conclusion and future work of the dissertation.

This dissertation presents a comprehensive framework for improving health monitoring and intent recognition through advanced data-driven modeling techniques. At its core, the fundamental approach involves leveraging multimodal data integration and contextual learning, spanning across sequential, graph-based, and large language models, each tailored to address specific health-related forecasting and intent categorization tasks.

1.4.1 Cross-modal Memory Fusion Network for Multimodal Sequential Learning with Missing Values

As I mentioned, leveraging complementary information from multiple modalities can significantly enhance the imputation process. The first step is to effectively model the intra-modality and inter-modality dynamics. I decompose this challenge into two sub-problems:

- 1. Modeling the recurrent dynamics within a single modality.
- 2. Leveraging cross-modal interactions to enhance the imputation process.

Existing methods primarily focus on the recurrent dynamics in one modality, often neglecting the complementary properties of other modalities. To address this, I propose a novel method called Cross-modal Memory Fusion Network (CMFN). CMFN explicitly learns both modal-specific and cross-modal dynamics to impute missing values in multimodal sequential learning tasks.

To validate the effectiveness of the proposed method, I conduct experiments on two benchmark datasets. Our results demonstrate that CMFN outperforms state-of-the-art methods, showcasing its potential to better handle missing values in complex multimodal datasets.

In this research, my contributions include:

- I introduce an innovative approach to impute missing values by leveraging both intra-modality and inter-modality dynamics.
- I demonstrate the superior performance of CMFN through comprehensive experiments on benchmark datasets.

The details of these contributions are described in Chapter 3.

1.4.2 Detecting Elevated Air Pollution Levels from Web Search Data

To address the challenge of RQ1.2, I propose an effective model to "nowcast" observed elevated pollution levels using search interest data, which is publicly available in near real-time from major search engines. Unlike previous efforts that tracked infectious diseases with search data, nowcasting air pollution presents distinct challenges due to non-specific symptoms but also opportunities to combine crowd-based observation data with physical measurements such as temperature and humidity.

The first step is to develop a composite model that effectively integrates physical measurements and search volume data. I decompose this challenge into two sub-problems:

- 1. Designing a search term Dictionary Learner to effectively capture relevant search interest data.
- 2. Combining this data with physical measurements using a Long-Short Term Memory (LSTM) model to nowcast pollution levels.

I propose a search-term Dictionary Learner-Long-Short Term Memory (DL-LSTM) composite model to combine physical measurements and search volume data for pollution-related terms. This model aims to provide real-time predictions (nowcasting) on whether air pollution is elevated on a specific day.

To validate the effectiveness of my approach, I explore several sequence classification models and evaluate my method in predicting three common and harmful pollutants: Ozone, Nitrogen Dioxide, and $PM_{2.5}$, across 10 major metropolitan areas in the USA. Our results demonstrate that incorporating search interest data using the DL-LSTM model significantly improves prediction performance for all three pollutants, suggesting promising novel applications for tracking global physical phenomena using online search data.

In this research, my contributions include:

- I introduce a novel method to nowcast elevated pollution levels by integrating search interest data with physical measurements.
- I demonstrate the superior performance of the DL-LSTM model through comprehensive experiments on multiple pollutants and metropolitan areas.

The details of these contributions are described in Chapter 3.

1.4.3 Modeling of Web Search Activity for Real-time Pandemic Forecasting using Graph Neural Network

To address the challenge of RQ 1.3, I leverage web search activity for pandemic forecasting. I propose a novel Self-supervised Message-Passing Neural Network (SMPNN) framework for modeling local and cross-location dynamics in pandemic forecasting. I decompose this challenge into two sub-problems:

- 1. Modeling local and cross-location dependencies using web search data.
- 2. Enhancing prediction accuracy through self-supervised learning and graph-generated features.

In previous research, regularized linear models have been effective in predicting the spread of respiratory illnesses like COVID-19 but are limited to specific locations and

do not incorporate neighboring areas' data. Our SMPNN framework addresses these issues by utilizing a Message-Passing Neural Network (MPNN) module to learn cross-location dependencies through self-supervised learning, improving local predictions with graph-generated features.

To validate the effectiveness of my proposed method, I compare the SMPNN framework with state-of-the-art statistical and deep learning models using COVID-19 data from England and the US. The results demonstrate that the SMPNN model outperforms other models, achieving up to a 6.9% improvement in prediction accuracy and lower prediction errors during the early stages of disease outbreaks.

In this research, my contributions include:

- I introduce a novel SMPNN framework to model local and cross-location dynamics for pandemic forecasting.
- I demonstrate the superior performance of the SMPNN model through comprehensive experiments using COVID-19 data from multiple regions.

This approach represents a significant advancement in disease surveillance and forecasting, providing a novel methodology, datasets, and insights that combine web search data and spatial information. The proposed SMPNN framework offers a promising avenue for modeling the spread of pandemics, leveraging both local and cross-location information, and has the potential to inform public health policy decisions.

The details of these contributions are described in Chapter 4.

1.4.4 Enhancing Healthcare Search Intent Recognition with Query Representation Learning and Session Context

To address the challenge of RQ 2.1 and 2.2, I propose novel methods to enhance query representation learning and session-based intent classification.

First of all, healthcare search queries often have multiple intents, resulting in ambiguous or divergent click behavior. This complexity necessitates improved query representation methods. I address this by aggregating similar queries via clustering and introducing a novel loss function to capture the multifaceted nature of health search queries. This scenario can be effectively modeled by leveraging user interaction data. However, different interactions and session contexts require distinct handling methods. Therefore, the feedback forms must be carefully considered when improving healthcare search intent recognition.

Generally, I need to consider two forms of data:

- 1. User interaction data, such as clicks from search logs.
- 2. Session context, to capture the intent within specific search sessions.

I propose a novel approach to quantify the ambiguity of health queries and the misalignment between global search intents and those discerned from individual sessions by introducing the concordance rate score. I also demonstrate a simple and effective method for incorporating my learned query representation into contextual, session-based search intent classification.

To validate the effectiveness of my approach, I conduct extensive experiments and analysis on two real-world search log datasets: a Health Search (HS) dataset and the publicly available TripClick dataset. Our results demonstrate that my method not only improves the intrinsic clustering metrics for query representation learning but also enhances accuracy for subsequent search intent classification tasks.

In this research, my contributions include:

- I introduce a novel approach to enhance query representation learning by aggregating similar queries and capturing their multifaceted nature.

- I demonstrate the superior performance of my method through comprehensive experiments on real-world datasets, improving both clustering metrics and search intent classification accuracy.

The details of these contributions are described in Chapter 5.

Chapter 2

Related Work

This work of this dissertation focuses on building efficient machine learning models for health search data and I have three major goals:

- to effectively impute the missing data in the search series data so that the missing data can be used for training the model.
- to incorporate semantic and geographical information into search trend modeling.
- to predict individual search query intent based on the search sessions.

These goals rely on processing online search data, analyzing search trend models, aggregating information from multiple sources, and understanding search queries. In this chapter, I review related work on these previous research:

- Online health monitoring.
- Search trend modeling.
- Multi-modal data aggregation.
- Search query understanding.

2.1 Online Health Monitoring

Online crowd surveillance has been recently used as a means of tracking emergent risks to public health, as described in studies by [10, 40, 49]. This involves the collection of online search queries to document changes in symptoms related to infectious diseases such as influenza [9, 100], Ebola [41], Lyme disease [103], and dengue fever [16].

In recent years, there has been an increasing amount of literature on predicting the concentration of air pollutants using machine learning models [19, 136, 76, 99]. The highly non-linear relationship between air pollutants and corresponding influential factors has been explored using Support Vector Machines [143], Artificial Neural Networks [133], and ensemble learning methods [8]. Most recently, Chen *et al.* and Zhao *et al.* build models to forecast daily AQI (air quality index) based on recurrent neural networks [136, 19]. Although these studies have successfully established frameworks to predict and monitor air pollution based on historical pollutant concentrations, they become infeasible when and where those data are unavailable. Online surveillance-based approaches are promising ways to overcome this limitation since user-generated content such as social media or online search behavior is publicly available in near real-time. For many years, online surveillance models have been applied to various syndromic surveillance researches [145, 120]. Over the past decade, most research using online search behavior has emphasized the use of Google Trends [54, 14]. Although previous studies have used online search data for various prediction purposes [103, 16], there is limited work on using online search data for monitoring ambient air pollution.

These methods have the potential to provide public health and medical professionals with benefits over traditional health surveillance and environmental epidemiology in their ability to capture both personal exposures and response dynamics at more sensitive spatial and temporal scales [40]. Despite the promise of these approaches, only a limited number of studies have examined how crowd-surveillance approaches can be

used to track exposure and, less frequently, response to non-infectious environmental-mediated disease processes [30, 33, 104]. Recent efforts to quantify the global burden of disease attributable to outdoor and indoor air pollution have increased public awareness on the severity of this public health crisis worldwide [27].

Urban air pollution serves as a critical test case for evaluating online surveillance methods for non-infectious environmental risks. Recent estimates attribute over 3 million excess premature deaths to ambient fine particulate matter ($PM_{2.5}$) and ozone (O_3), collectively placing these pollutants among the most substantial environmental contributors to global disease burden [27]. Traditional indicators of air pollution exposures, namely concentrations measured at ambient monitoring sites, have been widely used to assess health effects associated with air pollution in epidemiological studies. However, using ambient monitoring measurements as proxies for exposure may significantly underestimate health responses and potential risks, particularly for individuals who do not live near monitoring sites [130, 102, 70]. Moreover, ambient monitoring is designed to provide data on outdoor povllutant concentrations, which may not accurately reflect health-relevant exposures for individuals who spend most of their time indoors or have preexisting biological susceptibility to air pollution. Several recent studies have explored the use of smartphones in distributed air pollution sensing networks, enabling users to record and upload local air pollution data to create crowd-sourced, geospatially-refined pollution maps [30, 33, 104]. These studies demonstrate the feasibility of crowd-generated participation in projects predicated on urban air pollution awareness.

Linear models such as regularized regression have been successfully applied to various multivariate time series prediction tasks, including the estimation of disease rates from social media or online search data [60, 63]. For online search data, search queries can be semantically correlated and the co-linear predictors in the input space could cause model consistency problems, which is common in text regression tasks [61, 62].

Previous researches use elastic net [146] as the regression function and suggest that ℓ_2-norm could address the model consistency problem [63, 145]. However, it is not able to explore non-linear relationships across different time steps and explore the semantic correlations across multiple search terms. Long short-term memory units (LSTM) [50] are recurrent neural networks (RNNs) models designed for sequence modeling, which could learn the non-linear relationship in time series data [37]. Although LSTM models have achieved great success in many fields e.g. neural machine translation [118] and speech recognition [44], there is limited work on exploring how to incorporate semantic information of search queries with search volumes as the input for multivariate time series prediction.

Multivariate sequential learning with missing values

Studies have explored multi-modal perception for face-to-face communication [126, 127], and many models have been developed to impute missing values in multivariate sequential data using either local statistics or recurrent dynamics [18, 106, 12, 122]. While these methods have had success in single-modality datasets, they are not naturally adaptable to multi-modal data where different modalities may have varying missing rates. Previous research has explored multi-modal sequential learning approaches such as early fusion, late fusion, and models that learn both intra-modality and inter-modality dynamics in an end-to-end manner [127, 128, 71]. However, the challenge of imputing missing values across multiple modalities remains underexplored.

A variety of imputation methods such as statistical imputation (e.g., mean, median), EM-based imputation [86], K-nearest neighborhood [38] and tensor factorization [20] have been applied to estimate missing values. However, these approaches fail to model the sequential pattern of data and are independent of the training process, which often leads to sub-optimal results. To tackle this issue, recent studies

[12, 18, 106] propose end-to-end frameworks that jointly estimate missing values and make the prediction. For example, Che *et al.* [18] introduced the GRU-D model to impute missing values in a single modality using the linear combination of statistical features, which is under strong assumptions that missing values could be learned by assigning weights between the last observed value and statistical mean value.

Multi-modal sequential learning

Previous studies dealing with multi-modal sequential data have largely focused on three major types of models as mentioned in section 1. The third category of models [71, 128, 129] relies on collapsing the time dimension from sequences by learning a temporal representation for each of the different modalities. Memory fusion network (MFN) [128] is one of these models, which uses a special attention mechanism called the Delta-memory Attention Network (DMAN) and a Multi-view Gated Memory to identify the cross-modal interactions. Experiments show that these models [127, 128, 129] achieve remarkable success on a variety of tasks, including multi-modal sentiment analysis and emotion recognition; however, none of them can handle input with missing values in one or more modalities.

2.2 Health Search Trends Modeling

In this section, I describe related work using search trends for disease predictions and emphasize the need for proposing advanced neural architectures for modeling search trends for health. Given the context, the major baselines of the GFT model, regularized autoregressive models, I emphasize why developing advanced neural architectures, i.e. Recurrent Neural Network (RNN) Models and Graph Neural Network (GNN) Models for modeling search trends is necessary.

There has been increasing interest in using signals from online search activity

to predict infectious diseases such as seasonal influenza, the H1N1 pandemic, and COVID-19 [43, 137, 21, 108]. For example, Figure 4.1 shows the correlation between COVID-19-related symptom search activity and daily confirmed cases during the early stages of the pandemic.

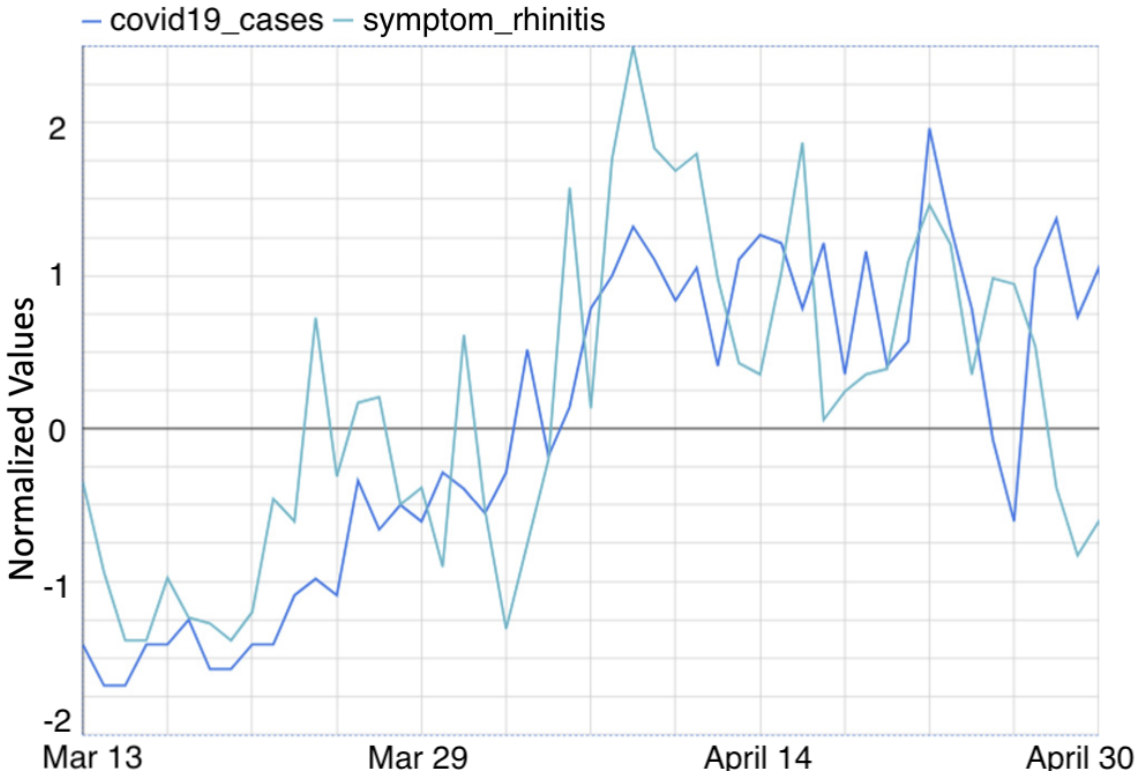


Figure 2.1: Two time series from normalized google search volumes of "Rhinitis" and normalized daily confirmed cases in Norfolk, UK for March to May 2020.

Studies have found that search terms related to symptoms like "fever" and "cough" can be robust indicators of COVID-19 incidence [91, 123]. Traditional location-specific regression models, like the Google Flu Trends method [43], have faced criticism for their accuracy, especially in the presence of media-driven spikes in search activity [28].

Recent advancements have proposed using autoregressive models with Elastic Net regularization to analyze disease-relevant search queries [108, 144, 64]. Graph Neural Networks (GNNs) have also been proposed to explore cross-location dependencies

for predicting infectious diseases [32, 90, 139], highlighting the potential for further improvement in disease forecasting using search data.

2.2.1 Google Flu Trends (GFT)

Google Trends is a widely used web-based epidemic signal for monitoring and predicting outbreaks of infectious diseases. It provides a simple and cost-effective way to track public interest in various topics and keywords related to infectious diseases, thereby offering a unique opportunity to monitor early warning signals of emerging disease outbreaks. Over the past few decades, research on infectious disease prediction using Google Trends has been validated and has shown promising results [43, 13, 107, 84].

Google Flu Trends (GFT) was an early study utilizing Google Trends to predict infectious diseases, showcasing the potential of search data for forecasting flu-like illnesses. Ginsberg et al. introduced GFT in 2009, employing a linear model to estimate the log-odds of Influenza-like illness (ILI) physician visits based on the log-odds of ILI-related search queries [43]. The GFT model is given by:

$$\text{logit}(P) = \beta_0 + \beta_1 \,\text{logit}(Q) + \epsilon$$

where P is the percentage of Influenza-like illness (ILI) physician visits, Q is the ILI-related query fraction computed in previous steps, β_0 is the intercept, β_1 is the coefficient, ϵ is the error term. The study highlighted the accuracy of Google Trends in tracking temporal and geographic patterns of flu-like illnesses by monitoring search volumes for specific keywords like “flu” and “influenza.” The authors also discovered that Google Trends data could predict the onset of flu-like illnesses one to two weeks earlier than traditional surveillance systems, such as the FluView program of the US Centers for Disease Control and Prevention.

Despite its initial success, GFT eventually faced issues with false positives and over-reporting due to the disparity between search data and actual disease occurrence. Consequently, Google Flu Trends was shut down in 2015. The primary cause of GFT's declining prediction accuracy was the increasing divergence between its algorithm and model, and the actual flu outbreaks. GFT's model used time-series data of keyword searches related to flu, such as "flu symptoms" and "cold medicine" to predict flu outbreaks. Initially, the algorithm and model were based on previous flu trends. However, as people's search behavior and habits evolved, the model's data bias increased, resulting in a weaker correlation with actual flu outbreaks [66].

The above discussion shows that the GFT model is powerful for describing temporal and geographic patterns for search trends and selecting search queries for prediction, but faces a detrimental performance drop because lack of semantic learning of search terms and failure to capture the change when people's search behavior and habits evolved. Therefore, I propose to use a model capable of learning the semantic correlation between the search terms and making timely forecasting when the search pattern occurs. I propose to adopt the embeddings from pre-trained language models and design specific RNN and GNN models for the forecasting tasks. But first, I need to introduce why exploring the correlation between search terms is important and introduce the RNN and GNN models for time series forecasting task.

2.2.2 Regularized Autoregressive Models

Learning from the limitations of the GFT model, subsequent research investigated the use of regularized autoregressive linear models to select search queries by penalizing the model coefficients [108, 144, 64]. Given that linear regression models can exhibit large prediction errors, potentially due to correlations within the data, Lampos et al. enhance the modeling approaches by examining the Elastic Net regression solver, incorporating selected queries into a nonlinear Gaussian Process (GP) regres-

sion method, and augmenting query-only predictions with an autoregressive model to leverage prior knowledge about the disease [63]. The authors clustered queries, learned a prediction function for each cluster, and summed these functions to obtain the final prediction result to avoid the interference of abnormal queries on the results. This study demonstrated that using the autoregressive Gaussian Process regression model for time series can more accurately predict the development trend of influenza.

The Elastic Net regression objective function is given by:

$$\hat{\beta}_{ElasticNet} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2 \ ,$$

where where y is the target variable, X is the input feature matrix, β is the vector of coefficients, λ_1 and λ_2 are regularization parameters. Elastic Net combines the advantages of feature selection and coefficient shrinkage. The GP model is given by:

$$y = \sum_{j=1}^{m} X_j\beta_j + f(x) + \epsilon,$$

where m is the number of linear regressors, X_j is the input feature matrix for the j-th regressor, β_j is the vector of coefficients for the j-th regressor, $f(x)$ is a GP prior, and ϵ is the error term. The objective is to estimate the coefficients, β_j, and the latent function, $f(x)$, to minimize the residual sum of squares (RSS) while considering the uncertainty in the function estimation.

Along with the development of regularized autoregressive models, the application of Google Trends data has been extended to predict various diseases beyond influenza. For instance, [4] explored the prediction of dengue fever outbreaks in Brazil by employing Ridge regression and Lasso regression models. These models utilized Google search queries and meteorological data to forecast dengue cases. The Lasso regression model outperformed the Ridge regression model, underlining the significance of feature selection in predicting dengue fever outbreaks. In another study, [134] examined

the prediction of norovirus outbreaks in the United Kingdom using linear and regularized linear models, such as linear regression, Ridge regression, Lasso regression, and Elastic Net models. The results demonstrated that Elastic Net and Lasso regression models offered the best predictive performance, showcasing the benefits of combining L_1 and L_2 penalties in this particular context.

The above discussion shows that understanding and minimizing the effect of correlation in search trends is crucial for improving forecasting performance, but current studies are limited to machine learning models without considering semantic information for search queries. Therefore, I propose to use a model aware of the semantic embeddings of search terms to minimize the influence of correlation in the data. Specifically, I introduce a matrix multiplication operation and propose to learn the semantic representation from the training process.

2.2.3 Recurrent Neural Network (RNN) Models

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models are RNN models that have been applied in various research domains, including disease prediction. LSTMs, introduced by Hochreiter et al., address the vanishing gradient problem faced by traditional RNNs by incorporating memory cells and gating mechanisms [50]. GRUs, proposed by Cho et al., are a simplified variant of LSTMs that combine the input and forget gates into a single update gate [23]. Both LSTM and GRU models have shown promise in handling sequential data, making them well-suited for time series-based disease prediction tasks [24, 3]. For example, Cho et al. applied GRU deep learning methods for predicting heart failure (HF) based on sequential clinical records. They found that GRU models outperformed traditional machine learning algorithms in predicting the risk of HF, emphasizing the advantages of RNN models in medical prognosis tasks [24]. Aiken et al. applied a GRU model for influenza prediction with the inclusion of real-time flu-related Internet search data

and show that GRU leads to a lower prediction error in their experiments [3]. LSTM models are composed of LSTM units. An LSTM unit consists of three gates, namely the input gate (i_t), the forget gate (f_t), and the output gate (o_t), as well as a memory cell (c_t). The LSTM unit updates can be given as:

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{ic}x_t + b_{ic} + W_{hc}h_{t-1} + b_{hc}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

where x_t denotes the input at time t, h_t represents the hidden state at time t, W and b are the weight matrices and bias terms, σ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication. GRU models are a simplified variant of LSTMs. They combine the input and forget gates into a single update gate (z_t) and use a reset gate (r_t) to control the degree of influence from the previous hidden state. The GRU unit updates can be given as:

$$
\begin{aligned}
z_t &= \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{t-1} + b_{hz}) \\
r_t &= \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}) \\
n_t &= \tanh(W_{in}x_t + b_{in} + r_t \odot (W_{hn}h_{t-1} + b_{hn})) \\
h_t &= (1 - z_t) \odot n_t + z_t \odot h_{t1}
\end{aligned}
$$

where x_t denotes the input at time t, h_t represents the hidden state at time t, W and b are the weight matrices and bias terms, σ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication.

The above discussion emphasizes using RNNs for time series forecasting, but there is a lack of studies building RNN models for search trends. Therefore, I propose to use this neural network architecture for building a forecasting model using RNNs, specifically, I propose several variants of standard RNNs to test their performance on our specific tasks.

2.2.4 Graph Neural Network (GNN) Models

Graph neural networks (GNNs) have demonstrated great potential in various applications for disease prediction, particularly due to their ability to model the spatial and temporal dependencies in disease transmission. Recent studies have showcased the effectiveness of GNNs in capturing complex relationships and improving epidemic forecasting accuracy. For example, Deng et al. proposed using a cross-location attention module in the graph message passing models for long-term influenza-like illness [32]. By modeling the spatial and temporal dependencies in disease transmission, they were able to improve the accuracy of epidemic forecasting. Real-time data and adaptive GNN models have also gained increasing attention in epidemic forecasting due to their ability to quickly adapt to changing epidemiological conditions [42]. The message-passing neural network (MPNN) framework, introduced by Panagopoulos et al. [90], can quickly adapt to real-time disease data and make epidemic forecasting with a limited amount of training data. In addition, Xie et al. proposed modeling spatial transmission with graph neural networks for pandemic forecasting with local and global encoding modules. By incorporating spatial information into their model, they were able to improve the accuracy of their predictions [119]. An MPNN model consists of two main components: (1) message passing, where each node receives messages from its neighbors and updates its hidden state based on the received messages and (2) readout, where a readout function is applied to the hidden states of all nodes

to produce a graph-level output, and the MPNN learning process can be given as:

$$m^{(t)}v = \sum u \in N(v) M_t(h_u^{(t-1)}, h^{(t-1)}v, euv)$$

$$h_v^{(t)} = U_t(h_v^{(t-1)}, m_v^{(t)})$$

$$y = R(h^{(T)}vv \in V)$$

where $h_v^{(t)}$ is the hidden state of node v at iteration t, $m^{(t)}v$ is the aggregated message for node v at iteration t, $N(v)$ denotes the neighbors of node v, euv is the edge feature between nodes u and v, M_t and U_t are the message function and update function at iteration t, respectively, y is the output, $h_v^{(T)}$ is the hidden state of node v at the final iteration T, and R is the readout function.

The above discussion emphasizes using GNNs for modeling graph information (geographical graph for search trends) for better representation learning and forecasting with limited annotated data, but there is a lack of studies on applying GNNs to improve the forecasting of search trends based on geographical information. Therefore, I propose the develop GNN models on search trends data in different geographical locations for timely forecasting task.

2.2.5 Time Series Forecasting with Noise and Missingness

Multivariate time series prediction task is crucial in various domains, including traffic flow forecasting, air pollution forecasting and medical anlysis [36]. Real-world time series data often come with different levels of noise and missingness, which can adversely impact the performance of prediction models.

Previous research has applied two primary methodologies to handle noise and missingness in multivariate time series forecasting: preprocessing data with imputation methods and jointly imputing data with forecasting tasks. For preprocessing data with imputation methods, previous studies have applied statistical imputation (e.g.,

mean, median), EM-based imputation [86], K-nearest neighborhood [38] and tensor factorization [20] to handle noise and missingness in data. However, these approaches often ignore the sequential pattern of data and the information in predicted labels, which often leads to sub-optimal results. The second approach is to jointly impute the data with forecasting tasks. Several studies have proposed end-to-end frameworks that jointly estimate missing values and make the prediction [12, 18, 106]. However, most of them rely on either strong statistical assumptions that missing values could be learned by assigning weights between the last observed value or ignores the correlation in multivariate data. Previous work on multi-modal sequential learning has proved that jointly modeling the data from different modalities leads to better prediction performance for the down-stream tasks, which benefits from learning the representation across different modalities [71, 128, 129]. Therefore, I propose to learn the joint representation across different modalities and use the joint representation to handle the noise and missingness in data in an end-to-end sequential learning framework.

2.2.6 Time Series Forecasting with Foundation Models

TimesFM[29], developed by Google, is a decoder-only foundation model designed for large-scale time series forecasting. Pretrained on over 100 billion time points from diverse data sources, including Google Trends and Wikipedia pageviews, TimesFM demonstrates robust zero-shot capabilities across various benchmarks. However, health-related time series data, particularly for tasks like pollution monitoring and disease forecasting, often exhibit unique temporal patterns and seasonality that general-purpose models may not capture effectively. By training models from scratch, this research tailors model architectures specifically to health-related data needs, ensuring adaptability and precision in capturing domain-specific trends.

2.3 Search Query Understanding

In this section, I first introduce the related studies on fine-tuning large language models (LLMs) for identifying health-related queries, then I introduce the related studies on clustering and classifying health-related intents using unsupervised and supervised approaches.

Search engines in the health domain rely heavily on identifying user intent to deliver relevant results. However, medical search queries are often ambiguous, making intent prediction challenging, especially without context [53, 115]. Prior studies have shown that implicit user feedback (e.g., clicks) can be useful for learning query representations [132], and co-click queries can serve as weak supervision for identifying similar search intent.

Several approaches, such as contrastive learning, have been applied to co-click query pairs to improve intent classification, but these methods may struggle when faced with ambiguous or multi-label health queries [125]. It is essential to develop more robust representation learning methods to handle these complex queries effectively.

2.3.1 Large Language Models for Health

Large language models, such as BERT and its variants, have demonstrated impressive results in various NLP tasks, including understanding user intents behind health-related search queries. These models have led to the development of domain-specific information retrieval systems and improvements in clinical NLP tasks. BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that has significantly improved the state-of-the-art across various natural language processing (NLP) tasks by leveraging bidirectional context and pre-training on a large corpus [34]. Since its introduction, several BERT variants and extensions have been

proposed, such as RoBERTa [79], ALBERT [65], and ELECTRA [26], which have all shown remarkable results in NLP tasks, including search query understanding.

In the health domain, there has been considerable research on understanding user intents behind health-related search queries using LLMs. Domain-specific information retrieval systems like PubMed [87] have played a crucial role in assisting researchers, clinicians, and patients in finding relevant information. Several studies have explored the application of large language models like BERT for this purpose. For example, Lee et al. [67] introduced BioBERT, a BERT model pre-trained on a large biomedical corpus, which demonstrated significant improvements in various biomedical NLP tasks compared to the original BERT. Similarly, Huang et al. [52] proposed ClinicalBERT, pre-trained on clinical notes, to enhance the performance of clinical NLP tasks.

In addition to BERT and its variants, there are also studies to propose LLMs trained specifically for health text understanding. For example, Yang et al. [121] developed GatorTron, a large clinical language model trained from scratch on over 90 billion words of text, including more than 82 billion words of de-identified clinical text. They systematically evaluated GatorTron on five clinical NLP tasks, demonstrating significant improvements in accuracy for tasks such as natural language inference and medical question answering, which can be applied to medical AI systems to enhance healthcare delivery.

The above discussion emphasizes the developing of LLMs for health-related text, but there is a lack of studies evaluating those models for evaluating their performance on short text (i.e. short web search phrases). Therefore, I propose to first directly apply and evaluate the performance of different LLMs in health domains for health-related search query clustering. I'll talk about the clustering technique in the following section in order to propose our method.

2.3.2 Search Query Clustering and Classification

Search query clustering and classification have gained increasing attention in recent years, with the application of pretrained language models like BERT and RoBERTa to understand user search intent and enhance search engine performance. Pretrained language models have significantly improved various natural language processing tasks. They have been increasingly employed to cluster and classify search queries to better comprehend user intent and optimize search engine performance [34, 79].

In the context of health-related search queries, several studies have demonstrated the potential of LLMs in retrieving and understanding health-related search queries and improving performance in clustering and classification tasks. Roberts et al. provided an overview of the TREC-COVID information retrieval shared task, which aimed at retrieving relevant information for COVID-19-related questions using large-scale language models [98]. The shared task demonstrated the potential of large language models in retrieving and understanding health-related search queries. Reimers et al. experimented with contextualized word embedding methods, ELMo and BERT, to enhance open-domain argument search. They achieved impressive results in both argument classification and clustering tasks, substantially improving the state-of-the-art performance on multiple datasets [95]. Furthermore, MacAvaney et al. explored the use of BERT and other large language models for classifying search queries concerning search result diversification. They demonstrated that these models could effectively capture query intent, leading to improved search result diversification [82].

In addition to using LLMs embeddings, existing studies also attempt to fine-tune these models with downstream clustering and classification tasks to achieve better performance. Zhang et al. present Deep Aligned Clustering, an innovative method for discovering new intents in dialogue systems by leveraging limited known intent data. The approach uses pre-trained BERT model, k-means clustering, and an alignment strategy to improve robustness and outperform state-of-the-art methods

on two benchmark datasets [131]. Wang et al. introduce MEDIC, a few-shot learning method for medical search query intent recognition, addressing the challenges posed by short, noisy, and fine-grained medical queries. By leveraging co-click queries from user search logs as weak supervision and designing a new query encoder that combines semantic, syntactic, and generic knowledge, the method effectively recognizes intents in a real medical search query dataset [115]. In addition, Zhang et al. leverages large-scale Bing search logs and user clicks to learn a distributed representation space for user intent in search queries and demonstrate the effectiveness of using search log data as weak supervision to fine-tune the embeddings of search queries [132].

The above discussion emphasizes the usage of LLMs in the health domain and fine-tuning them for clustering and classification, but there is a lack of studies on fully exploring the user behavior and search logs for understanding search intents. In order to introduce how to explore the massive search log data, we'll first introduce the general search intent prediction study in web search in the next section.

2.3.3 Search Intent Prediction in Web Search and Conversational Agents

Search intent understanding is a crucial task in web search and search personalization [132, 46]. By understanding users' search intent, search engines can provide more relevant and personalized search results. Discovering new search intents and uncovering sub-intents can provide users with potential queries and explore additional search results for users, thereby improving the overall user experience [78, 131].

Prior research has applied three primary methodologies to understand search queries: analyzing query phrases, examining metadata such as clickthrough data and analyzing search contextual information. Analyzing query phrases requires the researcher to build models on search terms, structure and sequences. Leveraging the linguistic parsing structures from query phrases is one of the earliest method

of analyzing search contextual information [77]. Previous studies applied different machine learning and deep learning methods on learning from search query phrases and achieved great success [58]. Most recently, given the rapid development of NLP models, more and more studies start to use pertained LLMs for query phrase representation and achieve significant improvement on this task [105]. The biggest challenge of using LLMs for search query understanding is because of their short length and lack of context. Existing studies have explored retrieval augmentation techniques to improve query understanding but this technique leads to increased latency of LMs [105].

To accomplish propose to learn multiple representations for queries with multiple intents and apply hierarchical K-means clustering methodology to better understand the structure of search intents at different levels. Another approach to understanding search query intent is through user behavior data. One of earliest approach is to incorporate user click feedback to the ranking process to improve the performance of real web search [1]. Techniques such as click models and learning to rank (LTR) models have achieved great success in the past decades [25, 2]. For the purpose of search query intent understanding, recent studies have been using co-click data weak supervision to better understand users' search intent and consider query logs and user's search history as contextual information to understand search intents [132, 78]. However, there is still lack of research to fully explore the contextual information in the log data and different co-click supervision methods to generalize the search intent representation learning.

In recent years, conversational agents, such as chatGPT [88], LLaMA [109] and Gemini [31], have gained increasing popularity across a wide range of applications, from customer support to healthcare, driving the demand for more sophisticated and accurate systems. One critical component for improving these systems is fine-grained intent prediction, which has become essential not only for enhancing user interaction

but also for minimizing issues such as hallucinations and improving the automation of workflows .

Fine-tuning conversational agents for intent recognition involves adapting models to capture nuanced user intents within specific tasks, such as information retrieval and dialogue systems. Techniques like Intent-based Prompt Calibration (IPC) have been introduced to enhance prompt engineering by refining user intent during conversations [68]. IPC iteratively adjusts prompts based on user input, using synthetic boundary cases to optimize prompts. This ensures the model better aligns with user expectations, even when handling ambiguous queries or diverse conversational goals, which is especially relevant in health or specialized search systems where user intent can vary significantly across sessions (as seen in fine-tuning efforts using large LLMs)

Retrieval-augmented generation (RAG) is another approach that combines retrieval mechanisms with generative models to improve the conversational agent's ability to respond to queries by fetching relevant external knowledge [97]. Additionally, more generalized frameworks such as RichRAG refine RAG's retrieval and ranking mechanisms to address complex queries by exploring sub-aspects of a user's intent [114]. By breaking down a query into smaller intent-driven aspects, these systems optimize retrieval and ranking models to better handle fine-grained intent predictions . The multi-faceted retriever retrieves and ranks relevant documents based on the various sub-intents, ensuring a comprehensive response generation process.

These developments highlight the growing need for RAG systems to enhance intent prediction through fine-tuned models, which allows them to handle specialized or complex domains more effectively than generic LLMs. The use of retrieval systems alongside language models ensures that the most contextually appropriate and fine-grained information is retrieved and processed to better meet user needs.

Chapter 3

Modeling Search Trend for Air Pollution Detection

This chapter presents my previous research on improving data processing in health data and detecting elevated air pollution levels using multi-modal machine learning approaches. I explore two key topics, based on my work published in [72, 74].

First, I detail the Cross-modal Memory Fusion Network (CMFN), a model designed for multi-modal sequential learning with missing values, which I developed in collaboration with my co-authors. This research was published in the 43rd European Conference on Information Retrieval (ECIR 2021) titled "Cross-modal Memory Fusion Network for Multimodal Sequential Learning with Missing Values" [72]. In section 3.1.1, I discuss the model architecture, which leverages intra-modality and cross-modal dynamics to effectively impute missing values in health data. The remaining sections focus on experiments conducted on benchmark datasets, including air quality and sentiment analysis, demonstrating the model's superiority in handling missing data and improving prediction tasks.

Building upon the model I developed to handle missing values in multi-modal data, I investigated the multi-modal learning model to a specific air pollution de-

tection scenario. These models contribute significantly to time series forecasting and can be adapted to different problem settings or integrated together to further enhance performance. I introduce my research on leveraging web-based search data to detect elevated air pollution levels, published in JMIR Formative Research [74]. In section 3.2.1, I present the development of novel machine learning models that combine physical sensor data with Google Trends search volume data to predict pollution events in U.S. cities. This work, titled "Detecting Elevated Air Pollution Levels by Monitoring Web Search Queries," highlights the potential of search trends as an additional predictive signal for real-time environmental monitoring.

These research collectively showcases my contributions to advancing machine learning techniques for health and environmental data processing, offering robust solutions for missing data and pollution detection.

3.1 Cross-modal Memory Fusion Network for Multimodal Sequential Learning with Missing Values

In this sub-section, I introduce the Cross-modal Memory Fusion Network (CMFN), a recurrent neural network architecture designed to address the challenge of multimodal sequential learning with missing data. This work, published in ECIR 2021, focuses on handling health-related datasets that contain multiple modalities with varying rates of missing information [73]. CMFN leverages both intra-modality and cross-modal dynamics to robustly impute missing values, improving the model's ability to make accurate predictions despite incomplete data. By incorporating cross-modal interactions between different data streams, this model outperforms traditional methods that handle missing values independently. The following sections will de-

tail the problem setting, model design, and experimental results, highlighting the advantages of CMFN over baseline methods.

3.1.1 Problem Statement

In many real-world scenarios, information and data are multi-modal (e.g. heterogeneous features collected from multi-typed sensors for air quality surveillance [11, 135, 69]; and multi-modal perception for face-to-face communication [126, 127]). In these scenarios, features from different modalities are seamlessly used together for classification/regression purposes. However, multi-modal sequential data is often incomplete due to various reasons, such as broken sensors, failed data transmission or low sampling rate. For example, Figure 3.1a shows two time series of air quality data at Atlanta Fire Station #8, where two-thirds of fine particulate matter ($PM_{2.5}$) data is missing while relative humidity data is complete. Relative humidity data, as shown in Figure 3.1a, is promising for improving daily $PM_{2.5}$ surveillance because of its high correlation and low missing rate. Many previous studies [18, 106, 12, 122] have been developing models that could impute missing values in multivariate sequential data by either constructing local statistics or utilizing local and global recurrent dynamics. Although these methods have achieved remarkable success in multivariate sequential data of one modality, they can not be naturally adapted to multi-modal sequential data. Specifically, they are not designed to incorporate the information from modalities with lower missing rates for imputing the missing values of modalities with higher missing rates.

Previous studies [127, 128, 71] in multi-modal sequential learning have successfully explored intra-modality and inter-modality dynamics, leading to more robust and accurate predictions. Strategies for multi-modal sequential learning can be categorized into three main approaches. The first approach is early fusion, where multi-modal features are concatenated at the input level [85, 94]. This fusion strategy often fails

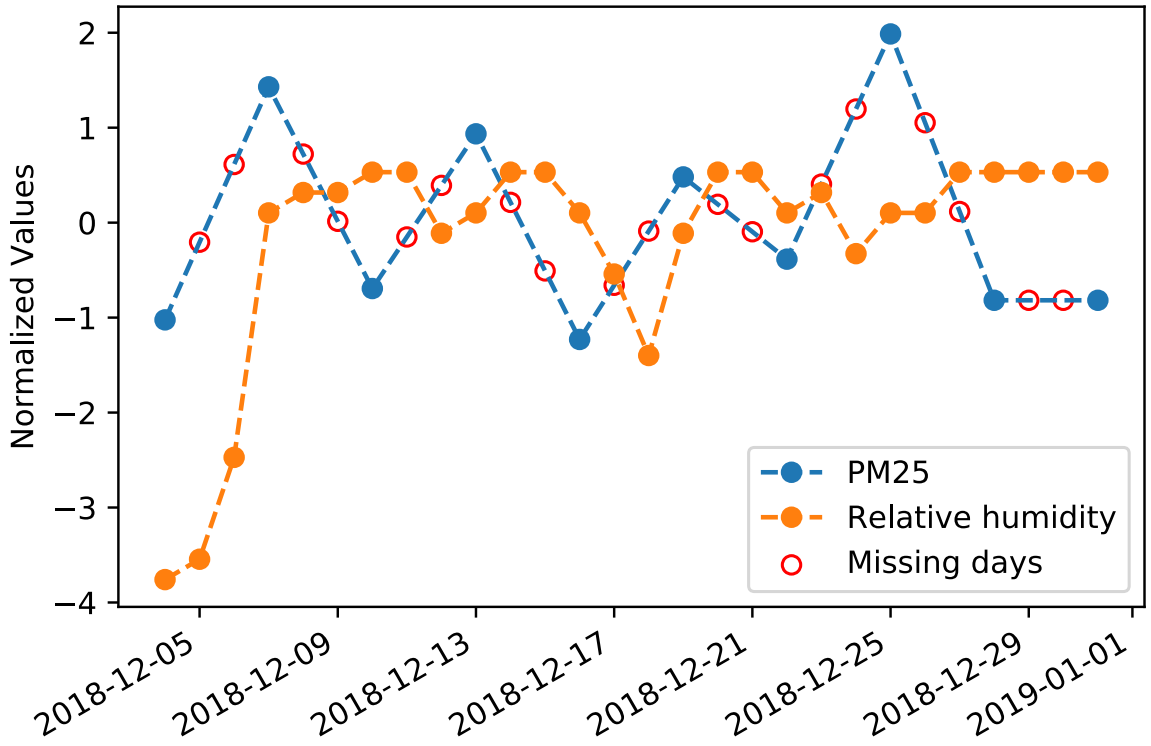


Figure 3.1: Two time series from $PM_{2.5}$ monitoring station at Atlanta Fire Station #8

to efficiently model intra-modality dynamics, as complex inter-modality dynamics can dominate the learning process or cause overfitting. The second approach is late fusion, which trains unimodal classifiers independently and performs decision voting [126, 111]. This strategy could lead to inefficient exploration of inter-modality dynamics by relying on the simple weighted averaging of multiple classifiers. The third approach involves designing models that can learn both intra-modality and inter-modality dynamics in an end-to-end manner [127, 128, 71]. It has been shown that by exploring the consistency and complementary properties of different modalities, the third strategy is a more effective and promising way of multi-modal sequential learning. However, there is few studies examining the condition when there are missing values in one or more modalities and how to leverage the intra-modality and inter-modality dynamics for missing value imputation remains an under-explored problem.

To address the aforementioned problems, I propose a novel cross-modal mem-

ory fusion network (CMFN) for multi-modal sequential learning with missing values. CMFN extends the memory fusion network [128], where recurrent neural networks (RNNs) are leveraged for learning intra-modality dynamics and attention-based modules are leveraged for learning inter-modality dynamics. Since the original RNN is unable to handle incomplete input, I introduced a novel variant of gated recurrent units (GRU) [22] called GRU-V to impute the missing values by leveraging modal-specific and cross-modal dynamics. The main contributions of this research are:

- I study a new problem of multi-modal sequential learning with missing values for online health data modeling.
- I propose a novel framework CMFN, with a GRU-V module to impute missing values in multi-modal sequential learning.
- I conduct experiments on both real-world datasets and synthetic datasets to validate the proposed approach.

In summary, multi-modal sequential learning can effectively model the health data with missing values by incorporating the information from different modalities, resulting in the exploration of intra-modal and inter-modal dynamics. All those challenges require practical and novel solutions.

3.1.2 Methodology

In this section, I first define the problem setting, and then I present the model architecture in detail.

Problem Definition

The input is multi-modal sequential data with $N \geq 2$ modalities. For those N modalities, I order them from high missing rate to low missing rate as modality

1, modality 2, ..., modality N. For each modality k, the input data is denoted as $X_k = \left[x_k^t : t \leq T, x_k^t \in R^{d_{x_k}}\right]$, where d_{x_k} is the input dimensionality of modality k. I also input the masking matrix $M_k = [m_k^t : t \leq T, m_k^t \in \{0, 1\}]$ to denote missing status ($m = 0$ means missing) and the time interval matrix $D_k = \left[d_k^t : t \leq T, d_k^t \in R^{d_{x_k}}\right]$ to denote the number of time steps since last observation.

Model Design and Training

The Cross-modal Memory Fusion Network (CMFN) is a recurrent model for multi-modal sequential learning with missing values, which consists of two main components: 1) A system of RNNs consisting of multiple RNNs for learning intra-modality dynamics. 2) DMAN and Multi-view Gated Memory [128] for learning inter-modality dynamics. As shown in Figure 3.1b, RNNs such as GRU and long short-term memory (LSTM) [51] are applied for modalities without missing values, GRU-V is applied for imputing the missing values with intra-modality and inter-modality dynamics for modalities with missing values.

GRU-V is inspired by the structure of GRU-D proposed by Che *et al.* [18]. To explain the procedure of missing value imputation, I assume that the input for modality 1 is feature matrix X_1, masking matrix M_1 and time interval matrix D_1. As shown in Figure 3.1b, at time step t, for the $N-1$ modalities with lower missing values, I concatenate their hidden outputs $\{h_2^t, h_3^t, \ldots, h_{N-1}^t\}$ as $h_{N\ldots2}^t$ to represent cross-modal dynamics. For modality 1, I have the hidden output h_1^{t-1} at last time step to represent modal-specific dynamics. I then concatenate the cross-modal and modal-specific dynamics, denoted as $c^{[h_1^{t-1}, h_{N\ldots2}^t]}$, and pass the concatenated tensor to a neural network $\mathcal{D}_v : R^{d_c} \mapsto R^{d_{x_1}}$ to infer the variance of the missing values from its empirical mean $\tilde{X_1}$ in modality 1 as:

$V_{X_1}^t$ are softmax activated scores, which is then used to infer the missing values as:

$\mathcal{X}_1^t$ are the inferred values, and I rescale $V_{X_1}^t$ from $[0,1]$ to $[-K,K]$ using rescale parameter K. Because all the input values are normalized, I set $K=3$ to represent the variance of input values. Following GRU-D, I then use a weight decay function $\Gamma_{D_1^t}$ to assign weights between the last observed value $X_1^{t'}$ and the inferred value $\mathcal{X}_1^t$ and get final imputed value $\hat{X}_1^t$ as:

$$V_{X_1}^t = \mathcal{D}_v\left(c^{[h_1^{t-1}, h_{N\ldots 2}^t]}\right) \tag{3.1}$$

$$\mathcal{X}_1^t = \tilde{X}_1 + 2K \cdot (V_{X_1}^t - 0.5) \tag{3.2}$$

$$\Gamma_{D_1^t} = \exp\left\{-\max\left(\tilde{\Gamma}, W_\Gamma D_1^t + b_\Gamma\right)\right\} \tag{3.3}$$

$$\hat{X}_1^t = \Gamma_{D_1^t} X_1^{t'} + (1 - \Gamma_{D_1^t}) \cdot \mathcal{X}_1^t \tag{3.4}$$

where W_Γ and b_Γ are model parameters that I train jointly with other parameters of the GRU. $\tilde{\Gamma}$ is the default weight decay, which is set as a hyper-parameter in range $[0,1]$.

3.1.3 Experimental Setting and Results

In this section, I describe experiments in four parts. First, I describe the datasets. Second, I present the baseline models. Then I describe the experimental setup. Last, I summarize experimental results comparing with state-of-the-art baselines.

Dataset

Air Quality Dataset Air Quality dataset is time series of daily measurement of $PM_{2.5}$ and meteorological data (i.e. relative humidity and temperature) in Atlanta Fire Station #8 monitoring site from Jan 1, 2011 to Dec 31, 2018. This dataset consists of two modalities and it facilitates a regression task of predicting $PM_{2.5}$

concentration based on data of the past 7 days.

CMU-MOSI Dataset Multimodal Opinion Sentiment Intensity (CMU-MOSI) dataset [126] is a collection of 93 opinion videos from online sharing websites with three modalities: language, vision, and acoustic. Each video consists of multiple opinion segments and each segment is annotated with sentiment in the range [-3, 3]. This benchmark dataset facilitates three prediction tasks: 1) Binary Sentiment classification 2) Seven-Class sentiment classification 3) Sentiment regression in range [-3, 3]. This dataset contains no missing values, so I synthetically introduce missing values by randomly masking 50% percent of the values in acoustic modality. I construct the synthetic datasets in two ways to test our model under different conditions. Synthetic Dataset #1: For 5 features in acoustic modality, I randomly mask values separately, which means this modality is partly masked when selected. Synthetic Dataset #2: I mask values for all 5 features randomly, which means this modality is masked totally when selected.

Experimental Setup

For the Air Quality dataset, I split the training (2011-2016), validation (2017) and testing (2018) sets chronologically. For the CMU-MOSI dataset, there are 1284, 229, and 686 samples in the training, validation, and testing sets respectively. I implement our models using Pytorch[1]. For all the experiments, the batch size is set to be 32 and all the parameters are tuned by the validation dataset.

Dataset	**Dates**	**Modalities**	**Prediction Task/Tasks**
Air Quality	2011/01/01 - 2018/12/31	2 ($PM_{2.5}$, meteorological data)	$PM_{2.5}$ (7-day regression)
CMU-MOSI	N/A	3 (language, vision, acoustic)	Binary classification, Seven-class classification, Sentiment regression

Table 3.1: Overview of datasets used in the CMFN experiments.

[1]https://pytorch.org

Baselines

Here, I use the following models for baselines and ablation studies.

- EFLSTM: LSTM model using early fusion strategy. The missing values are simply imputed by the last observed values and all modalities are concatenated into a single modality at the input level.

- MFN: State-of-the-art multi-modal learning model that learns the temporal representation for each modality using an RNN. The missing values are simply imputed by the last observed values.

- GRU-D: Baseline for multivariate sequential learning with missing values. All modalities are concatenated into a single modality using early fusion method at the input level.

- MFN-GRUD: This model is proposed for the ablation study and the RNNs in MFN are replaced with the GRU-D. Thus, it is a multi-modal learning architecture that imputes the missing values based only on intra-modality dynamics.

Results

Here is the revised version of your table split into two separate tables to better fit your column width:

Table 3.2: Comparison with state-of-the-art approaches for multi-modal sequential learning with missing values: Air Quality and CMU-MOSI Dataset #1.

Task	Air Quality		CMU-MOSI Dataset #1				
Metric	MAE	MSE	BA	F1	MA(7)	MAE	r
ELLSTM	3.19	15.5	0.726	0.725	0.325	1.051	0.584
MFN	3.17	15.35	0.739	0.735	0.322	1.012	0.618
GRUD	3.13	15.22	0.739	0.738	0.294	1.037	0.620
MFN-GRUD	3.07	14.8	0.736	0.729	0.321	**0.996**	**0.621**
CMFN	**3.04**	**14.21**	**0.755**	**0.751**	**0.354**	1.007	0.615

Table 3.3: Comparison with state-of-the-art approaches for multi-modal sequential learning with missing values: CMU-MOSI Dataset #2.

Task	CMU-MOSI Dataset #2				
Metric	BA	F1	MA(7)	MAE	r
ELLSTM	0.739	0.735	0.343	1.021	0.623
MFN	0.749	0.745	0.327	1.008	0.616
GRUD	0.755	0.750	0.331	**0.957**	0.652
MFN-GRUD	0.755	0.753	**0.354**	0.987	0.626
CMFN	**0.767**	**0.759**	0.353	0.958	**0.660**

Table 3.2 and Table 3.3 summarize the comparison between CMFN and proposed baselines for all the multi-modal sequential learning tasks. For the regression tasks, I report mean absolute error (MAE), mean squared error (MSE) and Pearson's correlation r. For binary classification, I report binary accuracy (BA) and binary F1 score. For multiclass classification, I report multiclass accuracy MA(k) where k denotes the number of classes. The results show that CMFN outperforms all the baseline methods in 8/12 tasks. For the CMU-MOSI dataset, when the features in acoustic modality are either partly missing (Dataset #1) or completely missing (Dataset #2), CMFN can robustly impute the missing values and outperform the compared methods. For the ablation study, the difference between CMFN and MFN-GRUD is that the latter only uses intra-modality dynamics for missing value imputation. The results show that CMFN outperforms MFN-GRUD in 9/12 tasks, which suggests that cross-modal dynamics can improve the missing value imputation performance.

3.1.4 Analysis & Discussion

In this study, I investigate a novel problem of exploring intra-modality and inter-modality dynamics for multi-modal sequential learning with missing values. I propose a new framework CMFN, which adopts modality-specific and cross-modal information for imputing missing values. To validate the framework, I instantiated a setup incorporating real-world data and synthetic data on benchmark multi-modal learning

data. Our result outperforms existing state-of-the-arts models, with ablation studies to show architectural advantages.

3.2 Detecting Elevated Air Pollution Using Web Search Queries

In this sub-section, I propose novel machine learning–based models to detect elevated air pollution levels at the US city level by using generally available meteorological data and aggregate web-based search volume data derived from Google Trends. Real-time air pollution monitoring is a valuable tool for public health and environmental surveillance. In recent years, there has been a dramatic increase in air pollution forecasting and monitoring research using artificial neural networks. Most prior works relied on modeling pollutant concentrations collected from ground-based monitors and meteorological data for long-term forecasting of outdoor ozone (O3), oxides of nitrogen, and fine particulate matter (PM2.5). Given that traditional, highly sophisticated air quality monitors are expensive and not universally available, these models cannot adequately serve those not living near pollutant monitoring sites. Furthermore, because prior models were built based on physical measurement data collected from sensors, they may not be suitable for predicting the public health effects of pollution exposure. Therefore, I propose to develop novel machine learning–based models using state-of-the-art deep learning methods to detect elevated air pollution levels.

3.2.1 Problem Statement

3.2.2 Methodology

In this section, I formalize the task of detecting elevated air pollution as a classification problem, and propose novel machine learning models for this task.

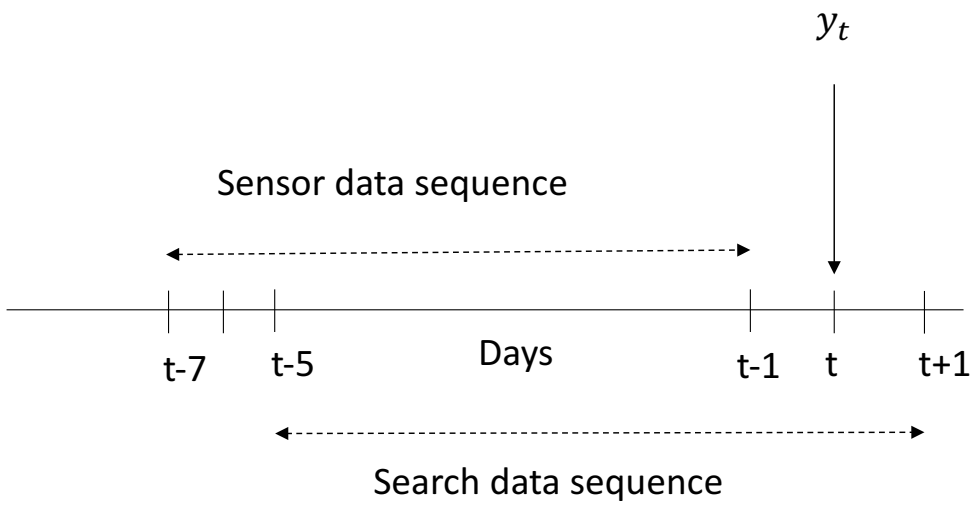


Figure 3.2: Input data sequences and prediction target illustrated on a timeline.

Problem Definition

Given sequences of physical sensor data $\boldsymbol{P} = [p_{t-T},\ ...\, p_{t-1}] \in R^{T \times d_p}$, and search interest data $\boldsymbol{S} = [s_{t-T+2},\ ...\, s_{t+1}] \in R^{T \times d_s}$, the task is to classify day t as "polluted" or not, where the positive class label indicates that the air pollution was above a pre-defined threshold. $T = 7$ denotes sequence length, and $d_p = 15$ and $d_s = 51$ are the number of physical sensor features and the number of search-related terms, respectively. Search interest data is provided to the model with a one-day lag. The prediction setup is illustrated in Figure 3.2. Following Zhao *et al.*, I set the sequence length T to 7 days, to base our predictions on one-week of data [136].

Model Design and Training

I propose two neural network models for learning a predictive representation of both $\boldsymbol{P}$ and $\boldsymbol{S}$ for pollution detection. Both models follow a two-branch composite architecture, as shown in Figure 3.3, where the left sub-network learns to extract features from physical sensor data, and the right sub-network does the same for search interest data. LSTM Composite Model (**LSTM**): I first describe an LSTM-based model where

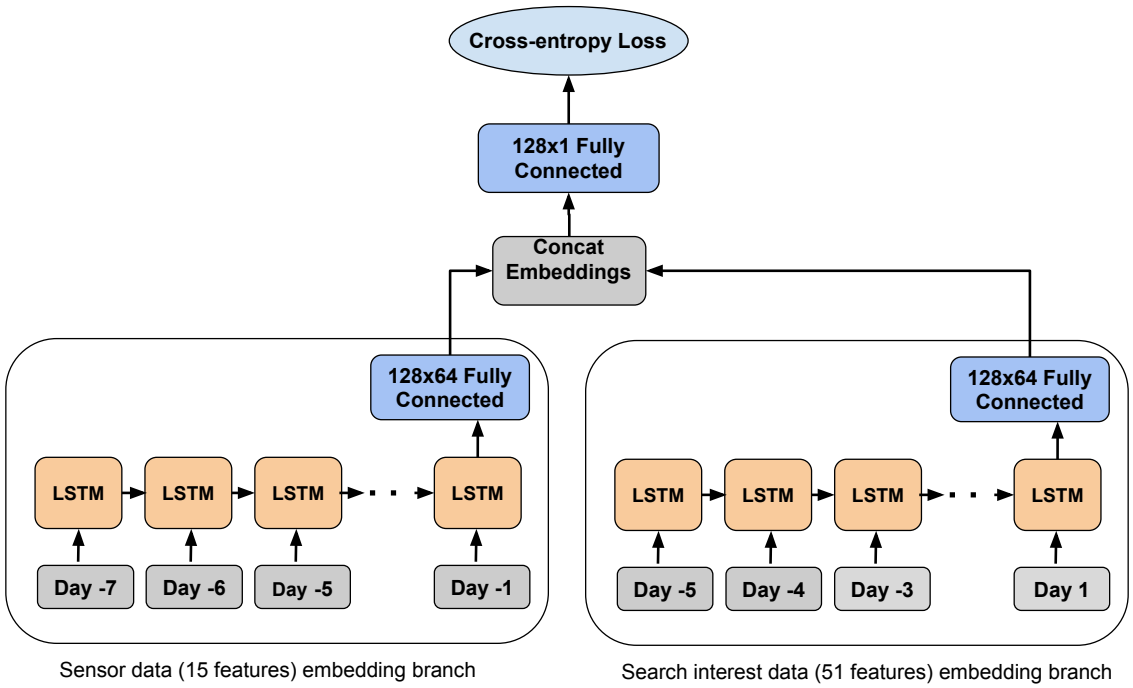


Figure 3.3: Architecture of the LSTM Composite Model

each sub-network consists of one sequence embedding layer and one fully-connected layer. See Fig. 3.3.

- Sequence Embedding Layers: To learn an embedding of the input time series, I use an LSTM cell (128 hidden units) with Rectified linear unit (ReLU) activation function and He initialization [47]. A 64-dimension fully-connected layer is then applied to the output of the LSTM cell.

- Fully-connected Interaction Layer: A fully-connected layer is employed to combine the features learned from the sensor data and search sequence data, in order to capture the non-linear interactions between these sequences. This layer uses a sigmoid activation function to produce a single probability of day t being polluted $\hat{y}_t$.

- Objective Function and Optimizer: The model is trained with a binary cross

entropy (CE) objective function and Adam optimizer [55].

$$CE(y, \hat{y}) = -(ylog(\hat{y}) + (1 - y)log(1 - \hat{y})) \tag{3.5}$$

where y are the binary class labels indicating whether each day is actually polluted, and $\hat{y}$ are the model's predicted probability of each day belonging to positive class (polluted day).

DL-LSTM Composite Model (**DL-LSTM**): This model differs from the previous model in the search interest embedding branch (See Figure 3.4). Instead of directly using search interest, this model attempts to learn the interaction patterns between search volume for each term, time lags, and the term-to-term correlations. Specifically, the the Dictionary Learning module transforms the original embedding vectors by back-propagating the error from the overall network prediction, to identify most predictive combinations of terms and time lags for each prediction task. To implement this idea, in addition to search interest data $\boldsymbol{S}$, I first initialize a search term embedding dictionary using their semantic embedding (GloVe 50-dimensional word vectors trained on tweets [93]). This dictionary is represented as a d_s by d_g matrix D_G, where $d_g = 50$ is the Glove embedding size. I transform D_G via a d_g by d_e ReLU-activated fully-connected layer, where $d_e = 150$ is the size of the new embedding, to create a new set of word embeddings that will be tuned for pollution detection. The resulting new dictionary is represented as a d_s by d_e matrix D_E, the sequence of search interest data $\boldsymbol{S}$ is multiplied by D_E and then feed into the LSTM composite model.

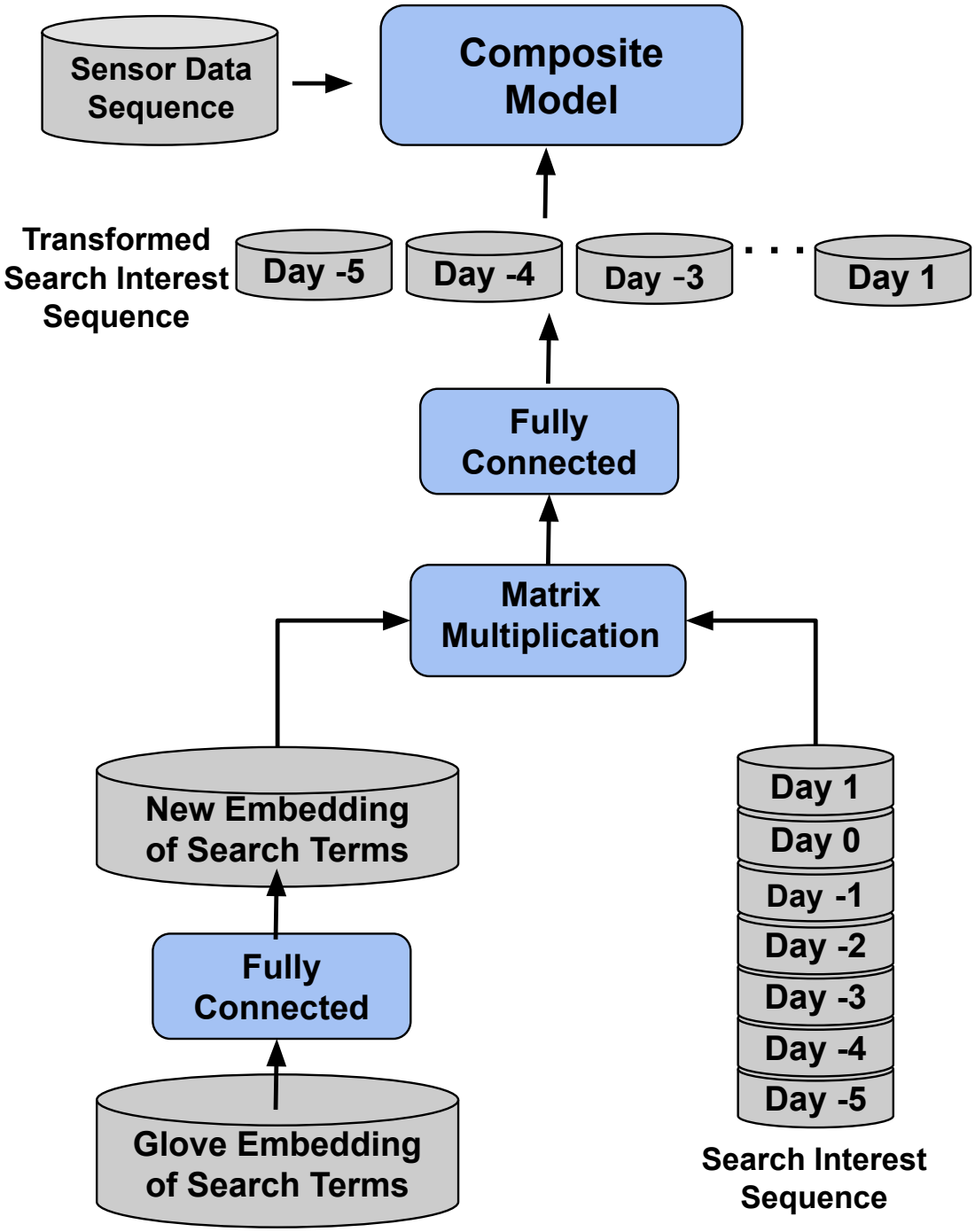


Figure 3.4: Architecture of the DL-LSTM Composite Model

3.2.3 Experimental Setting

Dataset

This section describes the source of the data used in this study, and details of the data collection method. Specifically, I first describe the ground truth data, collected through both general and specialized physical sensors. Then I describe and analyze the Web search interest data collected from Google Trends API. Finally, I present basic statistical properties of the data to provide intuition into the problem and to motivate our approach.

Data Collection

I collected daily air pollutant concentration data as well as temperature and relative humidity in ten largest U.S. metropolitan cities from Jan. 2007 to Dec. 2018. I focus on three air pollutants: ozone (O_3), nitrogen dioxide (NO_2) and particulate matter 2.5 ($PM_{2.5}$).

I collected the daily search frequency of pollution-related terms from Google Trends for the same 12-year period and cities. I created a curated list of 152 pollution-related terms based on our previous air pollution epidemiology studies and in reviewing the environmental health literature and downloaded the reports of trending results terms using PyTrends. For each PyTrends request, I downloaded the search history of pollution-related terms over a six-month window with one overlapping month for calibration. PyTrends provided us with search frequency scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Because of the PyTrends restriction, I downloaded the reports of trending results multiple times and the search frequencies are scaled, separately in each six-month window, which required us to calibrate the search frequency for the 12-year period. I calibrateed the search frequencies by joining the search logs on the overlapping periods (1 out of 6 months) for

Pollutant	Units	Value Range	EPA Standard	Description
O_3	ppm	[0.001, 0.106]	0.070	Daily max 8-hour average
NO_2	ppb	[1, 77]	100	Daily max 1-hour average
$PM_{2.5}$	ug/m_3	[1, 44]	35	Daily average

Table 3.4: Air pollutants description and notation

inter-calibration.

Ground-based Sensor Data The ground-based sensor data including the in-situ pollutant concentrations, maximum and mean temperature, relative humidity and dew-point temperature was retrieved from the EPA Air Quality System (AQS) and AirNow database. To come up with a single daily pollutant concentration value for each city, I used the median of all available monitoring sites within each city to avoid the impact of outliers. A description of pollutant levels in the collected dataset and the National Ambient Air Quality Standards (NAAQS) set by U.S. Environmental Protection Agency (EPA) for these pollutants are described in Table 3.4.

Web Search Interest Data By systematically reviewing the environmental health literature[35, 110, 101], environmental health experts suggested 51 pollution-related terms as seed search queries to identify known physiological responses to the selected pollutants, and common experiences and observations. These are primarily related to symptoms, observations and pollution source (e.g. cough, smoggy and wildfires). Daily search frequency of these search terms were retrieved from Google Trends. Google Trends API normalizes query frequency by geographic region and time span to represent relative popularity[2]. The resulting search volumes are scaled to the range of 0 to 100 to represent a topic's popularity relative to searches on all topics.

Search Term Expansion (STE) Since the exhaustive list of user queries is not available, reliance on the seed words may result in a poor prediction–due to the high mismatch rate between the user queries and our expected search words. Query

[2] `https://support.google.com/trends/answer/4365533?hl=en`

Expansion is a common approach to resolve this discrepancy. A recent study [145] showed that the initial set of seed words can be effectively expanded through semantic and temporal correlations. Thus, for each seed word I use Google Correlate[3] to retrieve the top 100 correlated query terms. Then I use the pre-trained word2vec model [83] to retrieve the vector representation of each query–phrases are mapped to the centroid of the constituent terms. A utility score is calculated for each candidate query by measuring the maximum cosine similarity between the query and the seed words. The queries with a high utility score are retained and the remaining queries are eliminated–we empirically set the utility cut-off to 0.55. This method expand the set of search terms for the total of 152 search terms to track.

Missing Data Imputation and Normalization Following the standard practices in environmental sensing, I fill in randomly missing data in historical pollutant concentration, temperature and humidity, with a rolling mean of window size 3. To fill in the missing data in infrequent search terms for which Google Trends does not return a count, I use random numbers close to zero ($e^{-10} \sim e^{-5}$). I normalize all the input features to standard scores by subtracting their mean values and then dividing it by the respective standard deviations.

Experimental Setup

Validation To tune model parameters and validate model performance, I split the available data into training (from Jan. 2007 to Dec. 2014), validation (from Jan. 2015 to Dec. 2016), and testing (from Jan. 2017 to Dec. 2018) sets. This eight-year training period provides a broad history to learn a relationship between input predictors and output variables, and the predictive models are evaluated based on their ability to make predictions for completely unseen periods. For evaluating our model, I make predictions for each day form Jan. 2017 to Dec. 2018 in the test

[3] `https://www.google.com/trends/correlate/`

dataset. The distribution of classes in train, validation, and test datasets is reported in (Table 3.5). Note that positive and negative classes are heavily imbalanced, with positive classes comprising, for instance, only 16% of training samples when $PM_{2.5}$ is the target pollutant.

Table 3.5: The distribution of classes in train, validation, and test sets.

Pollutant	Negative Samples			Positive Samples		
	train	validation	test	train	validation	test
O_3	24322	6269	6311	4896	1038	982
NO_2	23926	6119	6332	5292	1188	961
$PM_{2.5}$	24297	6745	6757	4921	562	536

Evaluation metrics

Because I defined this task as a classification problem, I used standard classification evaluation metrics. I report the accuracy and F1 score of the positive class (the harmonic mean of precision and recall) of predictions as evaluation metrics for all models. While accuracy measures the total fraction of correct predictions and could misrepresent model performance in presence of heavily imbalanced classes, the F1 score takes class imbalance into account and is, therefore, a more appropriate metric for our problem.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.6}$$

$$F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \tag{3.7}$$

where TP, TN, FP, and FN are the number of true positive samples, true negative samples, false positive samples, and false negative samples, respectively.

Baselines

I compare our model with several state-of-the-art methods as listed below:

- ***LR***: Logistic regression classifier with elastic net regularization.
- ***RF***: Random forest classifier with the number of trees and maximum depth tuned for prediction.
- ***LSTM***: Baseline LSTM model as shown in Figure 1, which combines physical sensor features, if available, with the search interest volume data directly, providing a direct adaptation of RNNs to this problem without any problem-specific extensions.
- ***LSTM-GloVe***: LSTM semantic model, which is a variant of LSTM model as described by Equation 1, I control the input of search interest data (*i.e.*, 51 seed search terms vs. 152 terms after STE) in this model. I refer to the variants as *LSTM-GloVe* and *LSTM-GloVe w/ STE* respectively.

Pollutant	L.A.	DC	PHILA.	DTX	ATL	BOS	NY	MIA	CHI	HOU
O_3(ppb)										
	55	54	53	53	53	48	49	45	49	49
NO_2(ppb)										
	43.7	38.1	36	25.2	27.8	30.7	45.3	25.5	43.7	27.7
$PM_{2.5}$(ug/m^3)										
	18.7	15.1	16.4	13.1	15.6	12.4	13.9	10.6	16.2	14.4

Table 3.6: Classification thresholds for three pollutants across 10 major MSAs in the U.S.

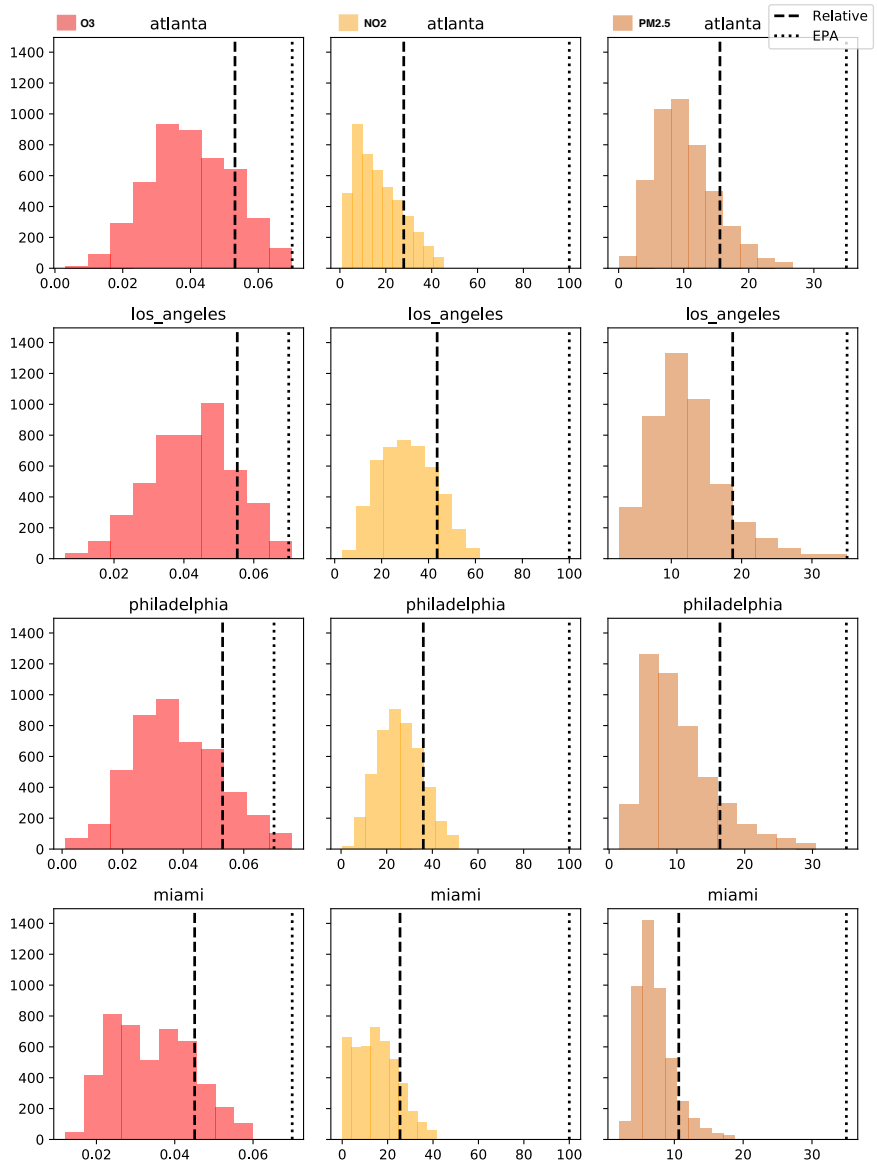


Figure 3.5: Distribution of pollution values for Atlanta, Los Angeles, Philadelphia, and Miami, with city-specific elevated pollution level (dashed line) and the general EPA-mandated standard (dotted line), for O_3 (left column), NO_2 (middle column), and $PM_{2.5}$(right column).

3.2.4 Results

In this section, we first present the findings of the data exploration. Then we present the principal findings of this study.

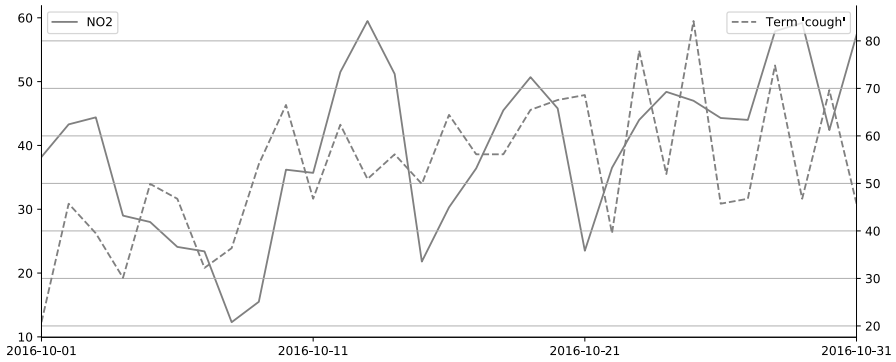


Figure 3.6: NO_2 levels and search interest for term "cough" in Atlanta, October 2016.

Insights from Collected Data

In this section, we describe the thresholds of abnormal air pollutant concentrations and then we present the lag between search anomalies and air pollution.

Thresholds of Abnormal Air pollutant Concentrations The major MSAs chosen for study in this work, have different distributions of pollutant concentration through time, almost always fall below the EPA standard 24-hour threshold (Figure 3.5). Despite this, multiple studies have shown that even at low concentrations, chronic exposure to air pollution negatively affects human health [41, 42]. Therefore, calibrating a meaningful threshold for each city, especially ones with generally lower levels of air pollution (e.g., Miami) may be critical for adequately protecting population health. A natural way to do this may be to set the threshold as one standard deviation above the mean daily pollutant concentration within each city, which is adopted in this study. The input predictors are also normalized within each city to reflect city-level dynamics. The resulting thresholds for the three pollutants and cities under investigation are reported in Table 3.6.

Lag between Search Anomalies and Air Pollution As shown in Figure 3.6, the normalized search frequency of the term "cough" is correlated with the concentration of NO_2 in Atlanta with a certain lag of time. To determine the lag between

elevated pollution levels and consequent pollution-related searches, the mean absolute Spearman's correlation between pollutant concentrations and search interest data was calculated, shifted forward in time for 0, 1, 2, and 3 days. As shown in Table 3.7, for O_3 and $PM_{2.5}$, the mean absolute Spearman's correlation increases with the increase of the shifted days. Considering that the task aims to detect elevated pollution levels as soon as possible, a lag of one day was applied to search data. In other words, search interest data from the current day was used to estimate whether air pollution was elevated on the previous day.

Evaluation Outcomes

In this section, we consider three conditions to evaluate the performance of using Web search data to detect elevated pollution, i.e., using only search data, using search data as an auxiliary data of meteorological data, and using search data as an auxiliary data of meteorological data and historical pollutant concentration.

Using Only Search Data: For areas where ambient pollution monitoring is unavailable, investigating whether Web search data could be used as the only signal for nowcasting elevated air pollution is a vital question. When relying on only search data for air pollution prediction, both the proposed DL-LSTM architecture and search term expansion contribute to the improvement of prediction accuracy. As shown in the "Search" section of Table 3.8, the LSTM-based models exhibit superior accuracy over the baseline LR and RF models for O_3 and NO_2. For $PM_{2.5}$, the proposed models do not perform better than the baseline LR or LSTM model because the validation and test dataset are heavily imbalanced (as shown in Table 3.8). In more detail, the proposed DL-LSTM w/ STE model achieves the highest F1 score (32.44% for O_3, 27.70% for NO_2) for detecting O_3 and NO_2 pollution.

Using Search Data and Meteorological Data: When meteorological data is available, we investigated the feasibility of using meteorological data with/without search activity data to nowcast air pollution under this condition. As shown in the "Met" and "Met +Search" sections of Table 3.8, the inclusion of Web search data improves the nowcasting accuracy for all three pollutants. In addition, the LSTM-GloVe w/ STE model achieves the highest F1 score (50.71% for O_3, 41.49% for NO_2) for detecting O_3 and NO_2 pollution. The LSTM-GloVe w/o STE model achieves the highest F1 score (26.99%) for detecting $PM_{2.5}$ pollution.

Using Search Data, Meteorological Data and Historical Pollutant Concentration: When historical pollution concentration is available, search activity data is added as auxiliary data to both meteorological data and historical pollution data. As shown in the "Met+Pol" and "Met+Pol+Search" sections of Table 3.8, the inclusion of Web search data improves the nowcasting accuracy for O_3 and $PM_{2.5}$. However, for NO_2, the inclusion of Web search data does not improve the nowcasting accuracy, which indicates increases in NO_2 concentrations may not be directly noticeable by people sufficiently to increase their search interest. This difference in performance for different pollutants and locales merits further investigation.

City-level Analysis of Ozone Pollution Prediction

We investigated the potential of using search interest and meteorological data to replace ground-based ozone sensor data for ozone pollution prediction in individual cities. As shown in Table 3.9, including search interest data (Met+Search) to augment purely meteorological data (Met) increases both accuracy and F1 metrics for most cities. While these metrics are not reaching the performance when the ground-level pollution sensors are available (Met+Pol), at least for two of the major MSAs (Philadelphia and Houston), search volume data indeed provides a useable alternative

to pollution monitors, with only 1.6% and 0.14% degradation in accuracy, respectively. Besides, the differences in model performance across different cities indicate that the online search pattern could vary from city to city. As shown in Table 3.10, the top five correlated terms differ across US cities in 10 years. The variation of search patterns could lead to a degraded prediction performance for certain areas, leaving promising directions for improvements.

Pollutant	Lag = 0 Search Term (Spearman's correlation)	P^a	Lag=1 Search Term (Spearman's correlation)	P^a	Lag=2 Search Term (Spearman's correlation)	P^a	Lag=3 Search Term (Spearman's correlation)	P^a
O_3								
	cough(-0.34)	¡.001	cough(-0.38)	¡.001	cough(-0.41)	¡.001	cough(-0.41)	¡.001
	bronchitis(-0.31)	¡.001	bronchitis(-0.32)	¡.001	bronchitis(-0.33)	¡.001	bronchitis(-0.35)	¡.001
	traffic(0.26)	¡.001	traffic(0.27)	¡.001	traffic(0.26)	¡.001	smoke(0.24)	¡.001
	smoke(0.23)	¡.001	chest pain(-0.23)	¡.001	chest pain(-0.23)	¡.001	traffic(0.23)	¡.001
	snoring(0.22)	¡.001	snoring(0.22)	¡.001	smoke(0.22)	¡.001	chest pain(-0.22)	¡.001
NO_2								
	asthma(0.20)	¡.001	sulfate(0.20)	¡.001	sulfate(0.16)	.002	cough(0.16)	.002
	sulfate(0.19)	¡.001	bronchitis(0.16)	.002	bronchitis(0.15)	.005	copd(-0.16)	.003
	cough(0.17)	¡.001	inhaler(0.15)	.005	cough(0.14)	.008	bronchitis(0.14)	.008
	bronchitis(0.17)	.001	cough(0.14)	.006	inhaler(0.11)	.03	wheezing(-0.12)	.02
	inhaler(0.16)	.002	difficulty breathing (-0.12)	.02	headache(-0.11)	.03	headache(-0.10)	.04
$PM_{2.5}$								
	wildfires(0.14)	.009	copd(-0.15)	.005	air pollution(0.19)	¡.001	air pollution(0.18)	¡.001
	copd(-0.11)	.03	wildfires(0.14)	.007	copd(-0.17)	.001	copd(-0.18)	¡.001
	snoring(0.11)	.03	air pollution(0.14)	.008	wildfires(0.14)	.009	wildfires(0.15)	.004
	inhaler(0.10)	.06	asthma attack(0.11)	.04	respiratory illness(0.10)	.05	sulfate(-0.11)	.03
	difficulty breathing (-0.09)	.08	respiratory illness (0.10)	.05	traffic(0.10)	.06	traffic(0.11)	.04

P^a value, with n = 366

Table 3.7: Cross correlation of top five search terms with different lags for three pollutants in the Atlanta metropolitan area in 2016.

3.2.5 Analysis & Discussion

Sensitivity Analysis of Air Pollution Thresholds Classification thresholds play an essential role in our model. In this study, a standard deviation threshold

from the mean of corresponding pollutants was used as a "probability threshold" to detect air pollution on a spatial-temporal resolution. However, the proposed method is sensitive to the threshold. We further investigate the performance of the proposed method on a variety of fixed classification thresholds. As shown in Figure 3.7, Figure 3.8 and Figure 3.9, we fixed the classification thresholds for all ten cities for detecting Ozone, NO_2 and $PM_{2.5}$ pollutions. The result shows that the meteorological and search data are complementary and combining search and meteorological data leads to better prediction performance for all classification thresholds.

In summary, while Web search data cannot yet fully replace traditional ground-based pollution monitors, it serves as a valuable supplementary signal, enhancing the sensitivity and responsiveness of pollution detection models, particularly for capturing unusual spikes in air quality events. The fluctuating correlation between search terms and pollution concentration across different cities suggests a need for city-specific fine-tuning to ensure the model's effectiveness. My findings indicate that no single search term works optimally for predicting all pollutants; thus, search term selection and model adaptation should be tailored to each pollutant type and location.

While metrics such as F1 scores show improvement in model performance, these metrics alone may not capture the full utility of integrating search data into health forecasting. For example, search trends offer unique insights into public awareness and behavioral responses to pollution events, especially when evaluated under scenarios involving media coverage and shifts in public search interest. This capability makes search data uniquely suited as a surveillance tool, as it reflects real-time public engagement and awareness, which meteorological data alone cannot capture.

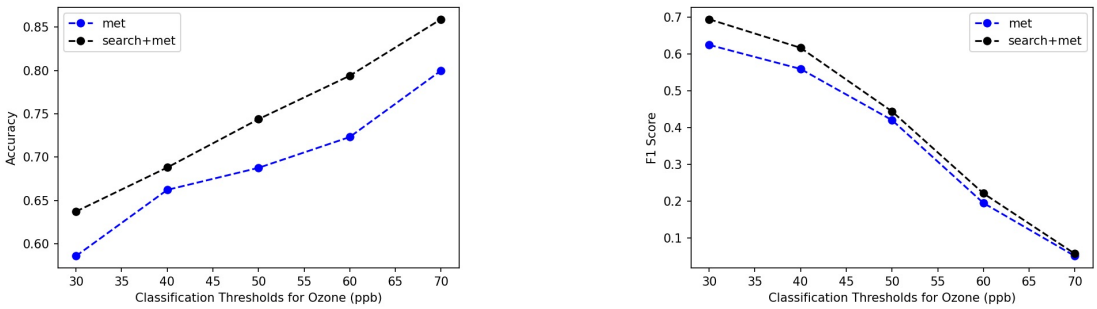


Figure 3.7: Accuracy (left figure) and F1 score (right figure) for detecting Ozone pollution on various classification thresholds, with Met (LSTM model) and Met+Search (DL-LSTM w/ STE) as features.

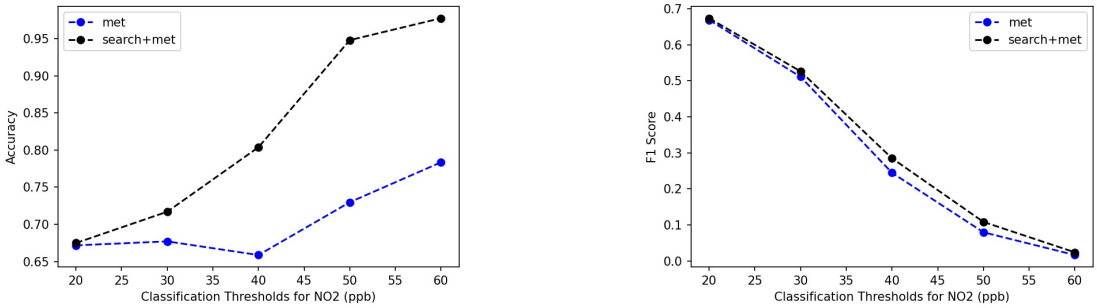


Figure 3.8: Accuracy (left figure) and F1 score (right figure) for detecting NO_2 pollution on various classification thresholds, with Met (LSTM model) and Met+Search (DL-LSTM w/ STE) as features.

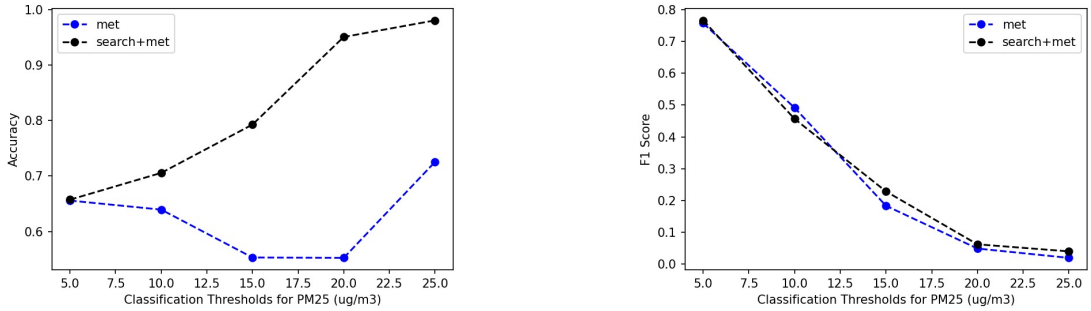


Figure 3.9: Accuracy (left figure) and F1 score (right figure) for detecting $PM_{2.5}$ pollution on various classification thresholds, with Met (LSTM model) and Met+Search (DL-LSTM w/ STE) as features.

Features	Model	O_3 Accuracy% (F1%)	NO_2 Accuracy% (F1%)	$PM_{2.5}$ Accuracy% (F1%)
No Prior Knowledge				
	All Positives	13.46 (23.73)	13.18 (23.28)	7.35 (13.69)
	All Negatives	86.54 (0.0)	86.82 (0.0)	92.65 (0.0)
	Random (Prob=0.5)	50.29 (20.63)	50.56 (20.68)	50.65 (12.67)
Search				
	LR	36.93 (17.77)	53.97 (24.17)	78.29 (10.72)
	RF	33.53 (23.36)	55.22 (18.1)	*92.65[1] (0.0)*
	LSTM	46.73 (23.63)	69.68 (21.62)	89.96 (7.58)
	LSTM-GloVe	53.23 (28.45)	63.44 (27.4)	90.09 (3.73)
	LSTM-GloVe w/ STE	69.17 (28.04)	46.85 (26.51)	91.73 (1.31)
	DL-LSTM	62.46 (30.4)	65.99 (26.19)	88.61 (7.97)
	DL-LSTM w/ STE	69.61 (32.44)	56.84 (27.7)	87.59 (6.99)
Met				
	LR	62.57 (39.81)	63.64 (37.25)	58.58 (22)
	RF	78.76 (50.59)	71.77 (39.88)	73.78 (24.67)
	LSTM	76.54 (48.29)	72.52 (41.27)	67.89 (24.69)
Met+Search				
	LR	55.99 (36.56)	62 (36.25)	61.25 (21.5)
	RF	81.39 (45.35)	73.77 (38.71)	87.96 (23.78)
	LSTM	78.18 (47.65)	77.75 (40.31)	88.14 (21.29)
	LSTM-GloVe	80.04 (49.37)	72.75 (40.35)	85.38 (26.99)
	LSTM-GloVe w/ STE	81.85 (50.71)	74.21 (41.49)	85.42 (26.13)
	DL-LSTM	77.97 (48.94)	74.81 (40.53)	84.94 (24.07)
	DL-LSTM w/ STE	80.16 (49.32)	72.99 (40.34)	87.04 (21.32)
Met +Pol				
	LR	67.38 (44.61)	70.05 (44.09)	74.45 (32.82)
	RF	82.81 (57.23)	80.35 (51.24)	86.45 (40.63)
	LSTM	86.97 (63.01)	84.64 (55.59)	85.25 (43.19)
Met+Pol+Search				
	LR	66.91 (43.71)	69.13 (43.6)	74.45 (32.82)
	RF	82.76 (55.91)	78.91 (47.72)	89.43 (37.57)
	LSTM	87.11 (61.54)	84.71 (54.02)	90.74 (44.81)
	LSTM-GloVe	87.94 (63.81)	82.98 (53.78)	88.19 (46.55)
	LSTM-GloVe w/ STE	87.63 (63.83)	83.81 (54.59)	88.24 (46.51)
	DL-LSTM	87.30 (63.02)	82.65 (53.65)	89.66 (47.35)
	DL-LSTM w/ STE	87.60 (63.61)	83.40 (53.58)	89.25 (46.59)

Table 3.8: Accuracy and F1 score of the LR, RF, and LSTM models for detecting elevated pollution across 10 major U.S. cities, for varying input feature combinations: no prior knowledge, search data only (Search), meteorological data only (Met); meteorological data and search data (Met +Search), meteorological data and historical pollutant concentration (Met +Pol) and all input features (Met +Pol+Search).

Features		L.A.	DC	PHILA	DTX	ATL	BOS	NY	MIA	CHI	HOU
Accuracy %											
	Met	72.6	77.4	83.29	83.42	83.56	75.62	68.36	58.09	76.71	85.89
	Met +Search	76.71	80.68	87.4	79.86	83.84	78.63	74.93	69.29	80	90.14
	Met +Pol	85.89	86.99	89.04	89.04	88.22	84.66	86.85	82.02	86.85	90
F1 %											
	Met	51.69	48.28	53.79	53.28	48.72	46.06	44.07	32.52	56.19	57.26
	Met +Search	54.3	50.53	58.56	41.9	42.72	48	47.86	35.84	57.56	59.09
	Met +Pol	68.11	60.58	64.29	64.6	56.12	55.56	63.64	55.48	70.73	67.26

Table 3.9: City-level accuracy and F1 Score for detecting elevated O_3 pollution in 10 U.S. cities, with Met (LSTM model), Met+Search (DL-LSTM w/ STE) and Met+Pol (LSTM model) as features.

Search Term (Spearman's correlation, lag = 1)									
L.A.	DC	PHILA	DTX	ATL	BOS	NY	MIA	CHI	HOU
cough (-0.40)	bronchitis (-0.25)	cough (-0.33)	cough (-0.25)	bronchitis (-0.14)	smoke (-0.11)	bronchitis (-0.31)	bronchitis (0.14)	wildfires (0.18)	ozone (0.12)
bronchitis (-0.33)	cough (-0.25)	traffic (0.27)	bronchitis (-0.24)	cough (-0.11)	haze (-0.07)	traffic (0.29)	air pollution (0.13)	smoke (0.08)	air pollution (0.12)
wildfires (0.24)	coughing (-0.19)	bronchitis (-0.20)	ozone (0.17)	chest pain (-0.10)	code red (-0.06)	cough (-0.25)	cough (0.13)	shortness of breath (0.04)	asthma (0.06)
traffic (0.14)	headache (-0.14)	organic carbon (-0.10)	wildfires (0.15)	respiratory infection (-0.09)	coughing (0.06)	wildfires (0.19)	power plants (0.09)	heart murmur (0.04)	organic carbon (0.05)
respiratory infection (-0.12)	wildfires (0.13)	respiratory infection (-0.09)	coughing (-0.14)	wheezing (-0.07)	smog (0.05)	wheezing (-0.15)	nitrogen dioxide (0.08)	tailpipe (0.04)	wildfires (0.05)

Table 3.10: Top five correlated search terms for O_3 pollution in 10 U.S. cities: Jan. 1, 2010 to Dec 31, 2019.

Chapter 4

Modeling Search Trend for Infectious Disease Forecasting

In this section, I propose a novel self-supervised message-passing neural network (SMPNN) framework for modeling local and cross-location dynamics in pandemic forecasting, building on the foundation of recent advancements in graph neural network (GNN) modeling of web search data. Our work and the methodology are published in the paper titled "Graph Neural Network Modeling of Web Search Activity for Real-time Pandemic Forecasting" presented at the 2023 IEEE International Conference on Healthcare Informatics (ICHI) [75].

The utilization of web search activity for pandemic forecasting has significant implications for managing disease spread and informing policy decisions. However, web search records tend to be noisy and influenced by geographical location, making it difficult to develop large-scale models. While regularized linear models have been effective in predicting the spread of respiratory illnesses like COVID-19, they are limited to specific locations. The lack of incorporation of neighboring areas' data and the inability to transfer models to new locations with limited data has impeded further progress. To address these limitations, I propose the SMPNN framework utilizes an

MPNN module to learn cross-location dependencies through self-supervised learning and improve local predictions with graph-generated features.

4.1 Problem Statement

Over the past decade, there has been an increasing interest in using signals generated from online search activity to predict infectious diseases, such as seasonal influenza and the H1N1 pandemic [43, 137, 21, 108]. Similarly, since the outbreak of COVID-19, several studies have investigated using online search activity to predict the increase in COVID-19 cases based on the intuition that people with relevant symptoms will search the Web for help [64, 91, 123]. For example, Fig. 4.1 shows two time series of COVID-19 related symptom "Rhinitis" search activity and daily confirmed COVID-19 cases in Norfolk, UK during March to May 2020. The peaks of these two curves are highly synchronized and have a strong correlation. In addition, by analyzing the Google search trends [80] and Twitter data, Panuganti *et al.* [91] calculated the relative correlation of online activity concerning different COVID-19 relevant symptoms with the disease incidence and concluded that Google search and tweet frequency regarding "fever" and "shortness of breath" are more robust indicators than "smell loss" for COVID-19 incidence. Meanwhile, Yom-Tov *et al.* [123] analyzed searches for COVID-19 relevant symptoms on Bing search queries from users in England and found that queries for "fever" and "cough" symptoms were the most correlated queries with future COVID-19 cases during the early stages of the pandemic. These studies indicate the feasibility to build COVID-19 forecasting models based on the search activity for COVID-19 relevant symptoms.

Location-specific regression models are the most widely-used method for pandemic forecasting using web search activity. The well-known Google Flu Trends method (the GFT method) applied a linear logit regression model on the aggregated search

volume of influenza-relevant queries [43]. Although the GFT method is effective in selecting disease relevant queries, [28] reported that the GFT predictions could be very inaccurate in practice. To overcome this limitation, several studies propose to use linear autoregressive (AR) models with the Elastic Net regularization to learn a sparse model directly on the time series of disease-relevant search queries [108, 144, 64]. For example, Lampos *et al.* [64] have built supervised AR models on COVID-19 relevant search time series and show that they could make predictions preceding the reported confirmed cases and deaths several days ahead. They also show that linear AR models could minimize the concerns that no sufficient data exists at the initial stage of disease outbreaks. Although linear AR models have been built for several respiratory diseases, they have been questioned for lacking the ability of making stable and accurate predictions, mainly because location-specific models tend to be impaired by the irregular change of search activity caused by short-term change in news or media exposure [108].

Furthermore, it is imperative to note that linear logit regression-based methodologies pose challenges in detecting search novelty [140], user interactions [138, 141], and broader geographical connections [142]. As a result, there is a pressing need for research that surpasses location-specific limitations, provides more abundant structure, and possesses the ability to predict and analyze disease tracking effectively.

Graph neural network (GNN) models have been proposed for exploring cross-location dependencies to make more robust prediction for several infectious diseases [32, 90, 139]. Deng *et al.* [32] propose a graph message neural network with cross-location attention for long-term seasonal influenza prediction with historical disease incidence time series as input and show that the cross-location dependencies in the data improves the model performance. Furthermore, Panagopoulos *et al.* [90] consider the mass mobility data between multiple regions and propose a message passing neural network (MPNN) model to predict the development of COVID-19 based on

past disease incidence. In their study, mobility is used as an indicator of spatial connectedness between locations. With MPNN, they update each vertex (region) based on messages received from neighboring regions. According to their results, MPNN has a superior ability to predict the development of diseases compared to multiple baseline models. Although cross-location dependencies in past disease incidence have been explored by prior studies for pandemic forecasting, there is limited work on combining the web search data with geographical graphs for pandemic forecasting, and it remains an open question whether GNNs could outperform location-specific regression models on the web search data.

Additionally, current models based on past disease incidence and mobility data are limited in exploring cross-location dependencies to make more robust predictions for infectious diseases [39]. While Graph Neural Network (GNN) models have been proposed to address this challenge and have shown promising results, there is limited work on combining web search data with geographical graphs for pandemic forecasting [113]. Furthermore, current models have limitations in accurately predicting the development of diseases during the early stages of outbreaks, which is a critical time for taking preventive measures. These limitations have hindered the advancements in pandemic forecasting and surveillance, especially in the context of emerging infectious diseases where timely and accurate predictions are crucial for controlling the spread of the disease.

4.2 Methodology

In this section, I first formulate the problem. Then I present the proposed neural network architecture and how it aggregates features for predicting the development of COVID-19.

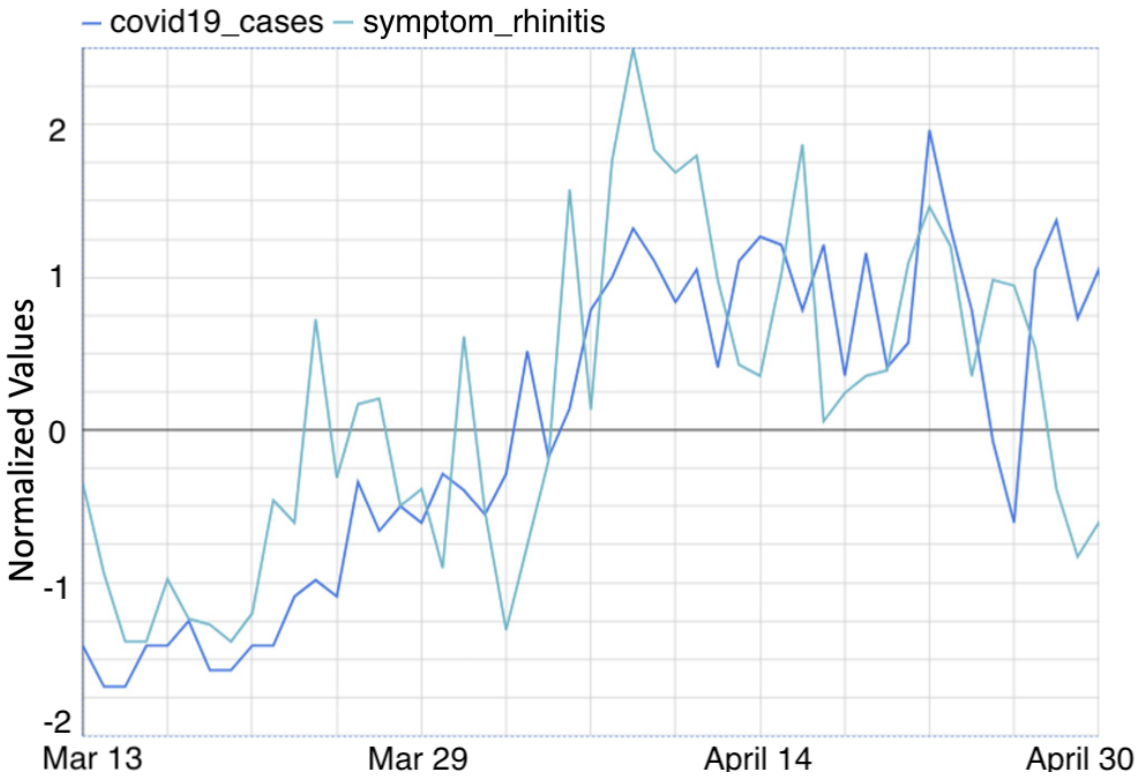


Figure 4.1: Two time series from normalized google search volumes of "Rhinitis" and normalized daily confirmed cases in Norfolk, UK for March to May 2020.

4.2.1 Problem Formulation

Input. I construct the daily snapshot of the search activity network as a (geographical) Graph $G = (V, E)$, where $n = |V|$ denotes the number of nodes and the weight $w(u, v)$ of the edge (u, v) represents spatial connectedness index between vertex u and vertex v. Specifically, for a given country, the nodes represent its subregions, and the edge weights are calculated by the mobility and social connectedness between the nearby sub-regions.

Spatial Aggregation. Given that people in nearby regions could move and contact with each other, the search activity in one region could be influenced by nearby regions. Therefore, the spatial connectedness index between the regions u and v at time t could be multiplied by the search activity $s_u^{(t)}$ of region u at time t to generate a relative value which represents the extent to which search activity in region v is in-

fluenced by region u at time t. Specifically, let $\mathbf{x}_u^{(t)} = \left(s_u^{(t-d)}, \ldots, s_u^{(t)}\right)^\top \in \mathbb{R}^{d*l}$, where $s_u^{(t)} \in \mathbb{R}^l$ is a vector of node features, which consists of the normalized search volume of l search terms of the past d days in region u. I use the search volumes of multiple days rather than considering only the previous day for prediction because search volumes vary greatly between days. In summary, the spatial aggregation process could compute a feature vector for each region with the following formula:

$$\mathbf{AX}^{(t)} = \begin{bmatrix} w_{1,1}^{(t)} & w_{2,1}^{(t)} & \ldots & w_{n,1}^{(t)} \\ w_{1,2}^{(t)} & w_{2,2}^{(t)} & \ldots & w_{n,2}^{(t)} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1,n}^{(t)} & w_{2,n}^{(t)} & \ldots & w_{n,n}^{(t)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1^{(t)} \\ \mathbf{x}_2^{(t)} \\ \vdots \\ \mathbf{x}_n^{(t)} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_n \end{bmatrix} \tag{4.1}$$

where $\mathbf{A}$ is the spatial connectedness matrix of $G^{(t)}$ and $\mathbf{X}^{(t)}$ is a matrix whose rows consists of the node features of each region. After spatial aggregation, $\mathbf{z}_u \in \mathbb{R}^{d*l}$ is a vector that aggregates the search activity within and towards region u.

Output. The goal of our work is to predict y_u^{t+k}, which is the reported number of COVID-19 cases for region u at k days after day t.

4.2.2 Model Designs

The main aim of our work is to model people's web search activity in graph G from real-time data, and measure the deviations from their search behavior to facilitate disease surveillance. To meet this goal, I design a two-stage framework: (1) Self-supervised MPNN module to generate cross-location features and (2) Location-specific regression module for disease prediction based on past search volumes and graph-generated search features.

Self-supervised MPNN module

MPNN framework represents a family of graph neural network models which use the message, update and readout functions to learn representation from the nodes in the graph [42]. I apply two neighborhood aggregation layers in the network and each layer learns from the graph structure and the node representation from the previous layer. I calculate the node representation for each layer using the following formula [56]:

$$\mathbf{H}^{i+1} = f\left(\tilde{\mathbf{A}}\mathbf{H}^{i}\mathbf{W}^{i+1}\right), \tag{4.2}$$

where $\mathbf{H}^{i}$ denotes the node representation matrix of the previous layer. $\mathbf{H}^{0} = \mathbf{X}$, represents the initial feature matrix. $\mathbf{W}^{i}$ denotes the parameter matrix of layer i and f is ReLU activation function. I train the parameter matrix using following loss function:

$$\mathcal{L} = \frac{1}{n}\sum_{u \in V}\left(s_u^{(t+k)} - \hat{s}_u^{(t+k)}\right)^2, \tag{4.3}$$

where s_u^{t+k} denotes the search volume of the search terms for region u at day $t+k$ and $\hat{s}_u^{(t+k)}$ denotes the predicted search volume of the search terms at day $t+k$.

Location-specific regression module

At the second stage, I apply location-specific regression models $f(\cdot)$ to predict disease incidence based on past search volumes and graph-generated search features for L symptoms. I predict the disease incidence according to the formula as shown below:

$$\hat{\mathbf{y}}_{\mathbf{u}} = f(\mathbf{S}_{\mathbf{u}}, \beta_u) + \epsilon_u. \tag{4.4}$$

When I use linear autoregressive model as the regression module, I optimize the model according to the following formula:

$$\arg\min_{\mathbf{w_u}, b_u} \left(\|\mathbf{y_u} - \mathbf{S_u}\mathbf{w_u} - b_u\|_2^2 \right), \tag{4.5}$$

where $\mathbf{y_u} \in \mathbb{R}$ is the reported cases in region u, and $\mathbf{w_u} \in \mathbb{R}^{2*l*d}, b_u \in \mathbb{R}$ denote the feature weights and regression intercept, respectively. Note that the time index of $\mathbf{S_u}$ is omitted for the simplicity of notation. In fact, for a specific search term out of l search terms, I use the search volumes of past d days and the graph-generated features from MPNN module of past d days.

End-to-end training pipeline

As mentioned above, I have two options for training the SMPNN model. The first way is to train SMPNN algorithm in an end-to-end way, where I apply a regression layer on top of the MPNN module. The pseudocode of the end-to-end SMPNN algorithm is described in Algorithm 1. The second way is to train the self-supervised MPNN module and location-specific regression module separately. Compared to end-to-end training of SMPNN, the second option preserves more location-specific information and has the flexibility to choose different regression models for location-specific regression.

4.3 Experimental Setting

In this section, I describe the dataset used in the experiment. Then, I describe the experimental setup and baselines in more detail.

4.3.1 Dataset

In this subsection, I introduce how I build our datasets from England and the US. Specifically, I collect England data from an open benchmark dataset provided by Panagopoulos *et al.* [90] and collect the US data from the Google COVID-19 open

Algorithm 1: SMPNN algorithm

[1] Time series data $\{X, y\}$ from multiple regions, spatial connectedness matrix A Model parameters Θ, prediction result y

for *each epoch* **do**

end

Randomly sample a mini-batch

for *each region* i **do**

end

Self-supervised process $\mathbf{h}_i \leftarrow \text{Graph Message Passing}\,(\mathbf{x}_{i:}, A)$
$\hat{s}_i \leftarrow \text{Output}\,([\mathbf{h}_i; \mathbf{x}_{i:}])$

for *each region* i **do**

end

Location-specific regression $\hat{y}_i \leftarrow \text{Linear Regression}\ (\mathbf{x}_{i:}; \hat{s}_i)$

$\Delta\mathcal{L}(\Theta) \leftarrow \text{BackProp}(\mathcal{L}(\Theta), \mathbf{y}, \hat{\mathbf{y}}, \Theta)$

$\Theta \leftarrow \Theta - \eta\Delta\mathcal{L}(\Theta)$

Table 4.1: Dataset statistics for England and USA.

COUNTRY	TIME	AVG CASE	MAX CASE	SD
ENGLAND	3/20-5/20	25.04	152.58	20.17
USA	9/20-12/21	279.56	10682.70	477.91

dataset [80]. Their data statistics are shown in Table 4.1. All disease cases are normalized as cases per million people.

- **England** This dataset contains daily COVID-19 confirmed cases from 48 regions in England, ranging from March 13, 2020 to May 12, 2020. I consider this dataset as COVID-19 forecasting at very early stage. Locations are represented as the NUTS3 regions. The spatial connectedness matrix is calculated based on the mobility between regions, which is collected from the movement data of meta Data For Good disease prevention maps [48].

- **USA** This dataset contains daily COVID-19 confirmed cases from 60 counties in the US, ranging from September 1, 2020 to December 31, 2021. This dataset contains the three most populated counties in the US (i.e. Kings county in New York, Cook county in Chicago and Los Angeles county in Los Angeles)

and their nearby counties, ranging from September 2020 to December 2021. I consider this dataset for COVID-19 forecasting in a longer period. Locations are represented as the GADM level 2 regions. The spatial connected matrix is calculated from the social connectedness dataset of meta data for good project [5].

I collected county-level search data from Google COVID-19 search trends symptoms dataset [80]. Specifically, according to the existing publications, I consider several symptoms, i.e. 'fever', 'cough', 'hay fever', 'fatigue', 'diarrhea', 'rhinitis' and 'shortness of breath'. For England, I track five symptoms including 'fever', 'cough', 'hay fever', 'rhinitis' and 'shortness of breath'. For the US, I also track 'fever', 'cough', 'hay fever', 'fatigue' and 'diarrhea' because 'rhinitis' and 'shortness of breath' volumes are missing for US counties. Search volumes of each symptom is normalized to 0-100 as the normalized popularity of a symptom.

4.3.2 Experimental Setup

I train the models using the data from day 1 to day T to predict disease incidence at day $T + k$. [124] reports that certain search symptoms (e.g. 'fever') could reach the highest prediction performance when k is equal or larger than 5. Therefore, I set k from 1 to 7 days in this study. For England, I increase T one day at a time with T initially set to 30 days and a validation set of last 10 days. For the US, I increase T one month at a time with T initially set to 2 months and validation set of last one month. With this experiment setup, I can predict disease incidence as early as possible.

I evaluate the performance of the models using the mean absolute error (MAE) since absolute changes in the disease cases are most widely-used metrics in pandemic

forecasting task:

$$MAE = \frac{1}{n}\sum_{u \in V} \left| \hat{y}_u^{(t)} - y_u^{(t)} \right| \tag{4.6}$$

Note that all reported cases are normalized with the population of that region throughout the experiments (i.e. cases per million people).

4.3.3 Baselines

I compare our model with several state-of-the-art methods as listed below:

- **Autoregressive Moving Average (ARMA)** [57] represents the linear autoregressive model. ARMA contains the autoregressive terms and moving-average terms together. The order of the moving average is set to 2 in implementation.

- **Random Forest (RF)** [7] is a non-linear regression model, which is a meta estimator that fits a number of regression decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

- **Support Vector Regression (SVR)** [17] is a non-linear regression model, which is a nonparametric technique which relies on kernel functions to make predictions.

- **MPNN** [90] by design, could serve as an end-to-end model to predict the disease incidence from the search activity graph. Comparing to the location-specific regression models, I follow a similar design as described in section 4.2.2 while replacing the loss function as below:

$$\mathcal{L} = \frac{1}{n}\sum_{u \in V} \left(y_u^{(t+k)} - \hat{y}_u^{(t+k)} \right)^2 \tag{4.7}$$

where y_u^{t+k} denotes the reported number of cases for region u at day $t+k$ and $\hat{y}_u^{(t+k)}$ denotes the predicted number of cases.

Hyper-parameter Setting For all the models, I use the same validation set to select the best model as decribed in section 4.3.2. Specifically, for RF model, I explore the tree depth from 3 to 9 to control model complexity. For SVR model, I use polynomial kernel and explore the regularization term C from 0.1 to 2. For all the neural network models, I use two neighborhood aggregation layers with the number of hidden units equals to 64 and store the model that achieves highest validation accuracy. To control model complexity, I apply batch normalization and dropout with ratio set to 0.5 to every neighborhood aggregation layer.

4.4 Results

Table 4.2: Mean absolute error for COVID-19 forecasting in number of cases per million people per region.

Model		**England**							**USA**						
Days ahead (k)		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ARMA		17.45	16.81	17.17	17.46	17.60	16.55	15.77	233.1	235.2	237.2	237.5	235.2	228.8	226.8
RF		13.07	14.03	14.79	14.35	13.99	13.34	13.03	245.5	243.9	247.0	242.8	245.2	239.4	234.9
SVR		15.11	14.21	**14.07**	14.00	14.23	13.95	13.60	215.0	**219.7**	223.0	220.3	217.1	209.1	207.4
MPNN		18.19	17.52	18.07	18.71	18.35	18.03	18.96	221.7	221.3	221.6	215.8	226.0	217.8	222.2
end-to-end SMPNN		19.20	19.00	19.28	19.95	18.71	19.28	20.33	221.4	226.9	226.6	227.3	213.9	221.7	219.7
SMPNN	w/ ARMA	16.26	16.50	16.53	17.18	16.78	15.76	16.34	228.7	239.5	232.6	227.6	224.7	216.7	216.7
	w/ RF	**12.83**	14.02	14.62	**13.99**	**13.36***	**12.67***	**12.81**	250.6	244.5	241.4	238.1	237.6	233.4	233.6
	w/ SVR	14.40	**13.85**	14.25	14.13	14.07	13.57	13.45	**212.1**	228.7	**216.4**	**207.0***	**202.2***	**198.4***	**197.2***
Relative Improvement		↑1.8%	↑1.3%	↓1.3%	↑0.1%	↑4.5%	↑5.0%	↑1.7%	↑1.4%	↓4.1%	↑2.4%	↑4.1%	↑6.9%	↑5.1%	↑4.9%

Notes: The numbers are computed as the average of 21 runs/days for the UK and 11 runs/months for the USA, where $^*p < .05$

Table 4.2 summarizes the comparison between SMPNN and baselines for the pandemic forecasting tasks on England and US datasets. We investigate the different settings of predicting disease incidence one to seven days ahead ($k = 1, 2, \ldots, 7$). For the regression tasks, we report the mean absolute error (MAE) of disease cases for two countries. For England, the models are trained and predict daily in a two-month window. SMPNN outperforms all the baseline methods in 6/7 tasks, with MAE reduction up to 5.0% when k equals to 6. For the US, the models are trained the predict

daily in a sixteen-month window. SMPNN outperforms all the baseline methods in 6/7 tasks, with MAE reduction up to 6.9% when k equals to 5. For England and the US, the lowest MAE is achieved when k equals 6 and 7 respectively, which is consistent with previous studies [64, 124].

4.5 Analysis & Discussion

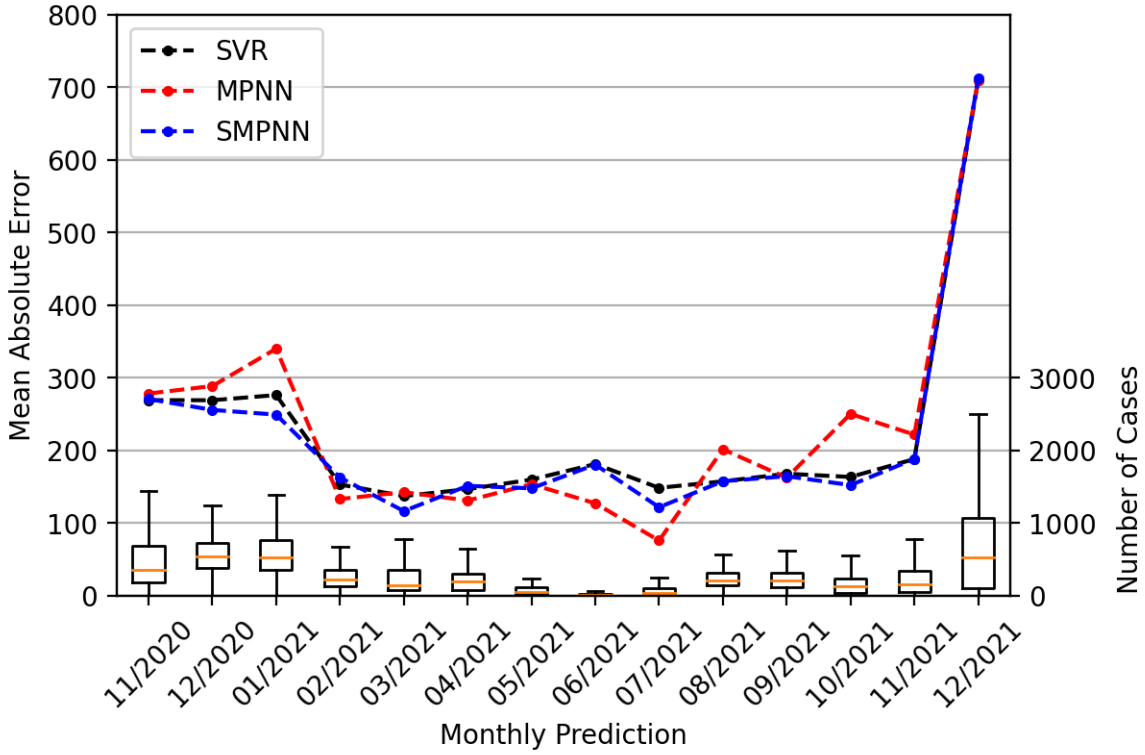


Figure 4.2: Monthly predictions for the US.

The baseline ARMA, RF and SVR models rely on location-specific dynamics for training, while end-to-end MPNN replies on cross-location dynamics with graph as the input. By design, SMPNN learns from both location-specific and cross-location dynamics, thus achieving lowest prediction errors as shown in in Table 4.2. We further investigate how different models perform at different stages after disease outbreaks. As shown in Fig. 4.2, the box plots show the distribution of monthly new

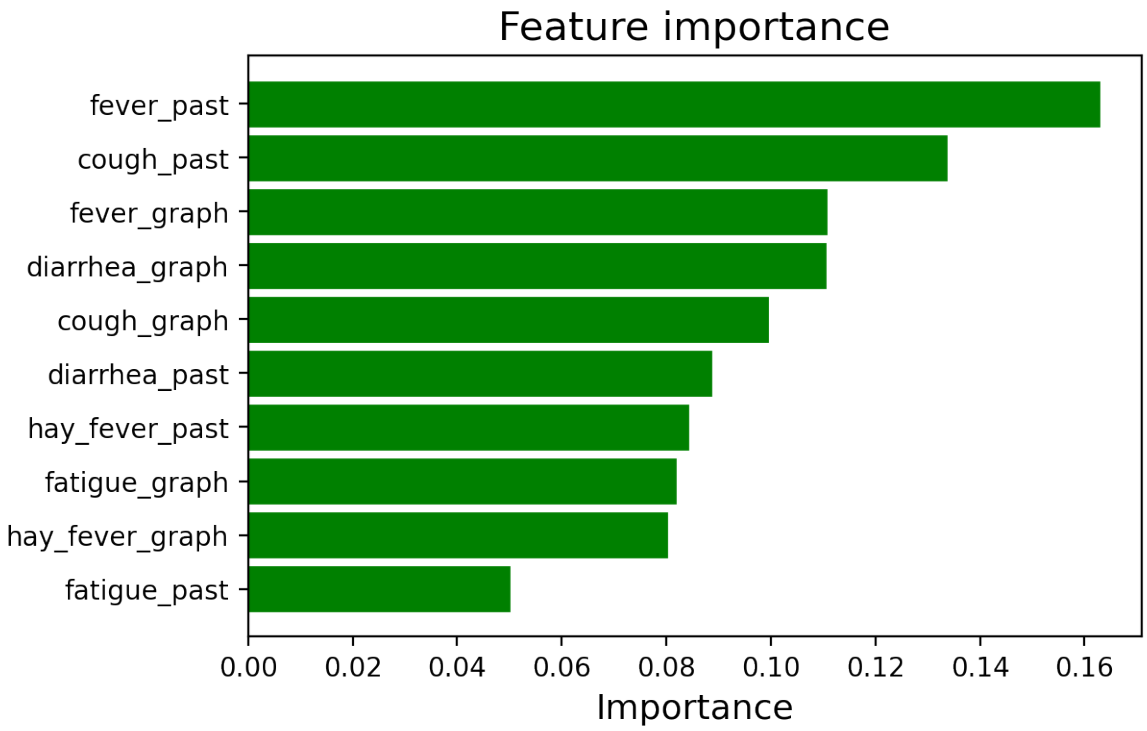


Figure 4.3: Feature importance for SMPNN (k=7).

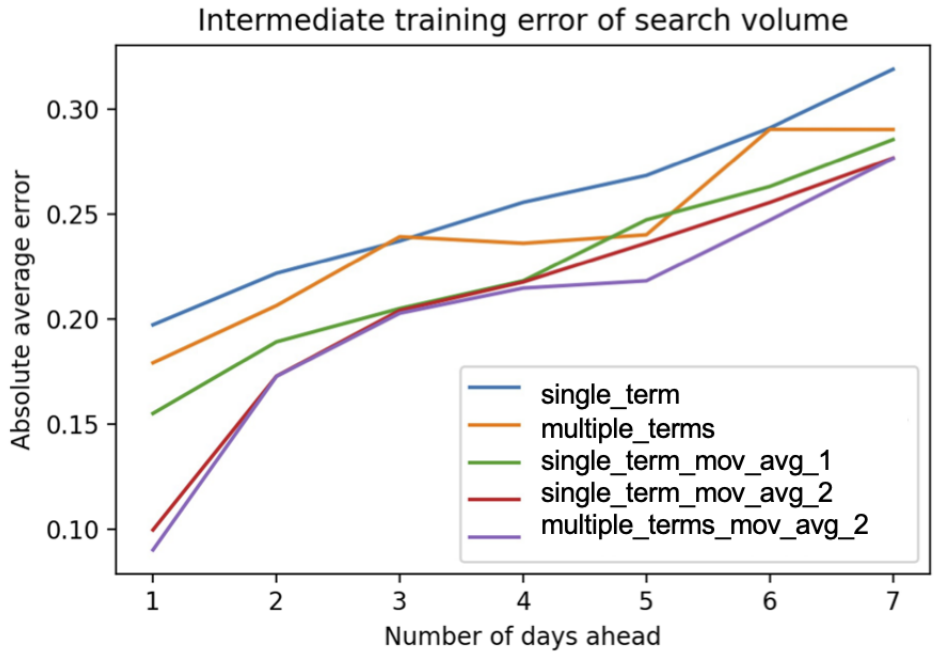


Figure 4.4: Intermediate training errors when training SMPNN model.

COVID-19 cases and the line plots represent the mean absolute error for SVR, MPNN and SMPNN models. At the early stage of prediction (i.e. the earliest predictions

Table 4.3: Pearson correlation of top ten search terms for UK and USA across all regions.

Search Term in UK	Correlation	Search Term in USA	Correlation
Rhinitis	0.446**	Ageusia	0.640***
Hay fever	0.442**	Anosmia	0.604***
Hair loss	0.390*	Low grade fever	0.548***
Allergy	0.366*	Fever	0.527***
Abdominal obesity	0.362*	Pneumonia	0.501***
Dermatitis	0.359*	Hypoxemia	0.468***
Itch	0.358*	Chills	0.466***
Sleep disorder	0.347*	Common cold	0.459***
Rosacea	0.305*	Shivering	0.416**
Insomnia	0.297	Dysgeusia	0.409**

Notes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 4.4: Pearson correlation of bottom ten search terms for UK and USA across all regions.

Search Term in UK	Correlation	Search Term in USA	Correlation
Pericarditis	0.003	Myalgia	0.172
Tumor	0.003	Xerostomia	0.166
Rheum	0.002	Infection	0.164
Bunion	0.002	Erectile dysfunction	0.151
Ataxia	0.002	Hypochondriasis	0.151
Anemia	0.002	Grandiosity	0.137
Petechia	0.001	Bradycardia	0.136
Blushing	0.0003	Periorbital puffiness	0.136
Varicose veins	0.0001	Burning chest pain	0.130
Hypoglycemia	0.00006	Palpitations	0.127

Notes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

in 11/2020, 12/2020 and 01/2021), SMPNN model outperforms all other models. SMPNN and SVR achieve the lowest MAE in 03/2021 while MPNN achieves the lowest MAE four months later in 07/2021. We also investigate how the location-specific features and the graph-generated features contribute to SMPNN model by calculating their average weights. As shown in Fig. 4.3, "fever_past" and "cough_past" (location-specific features) contribute most to SMPNN model, which is consistent with previous studies where search relevant to "fever" and "cough" contributes most to COVID-19 prediction [91, 124]. "fever_graph", "diarrhea_graph" and "cough_graph" (graph-generated features) are ranked third to fifth out of ten features, which shows cross-location dynamics is also important for COVID-19 forecasting.

Aside from the predictive capability of the model, we also explored the impact of the search terms on the model's performance during training stage. Two perspectives have been taken into consideration when evaluating these terms. As a first step, we proposed several term combinations which would be fed into the model and training errors would then be measured. The dataset from the UK is used to measure training errors in our analysis. For single term, we used the "rhinitis", and "fever", "cough",

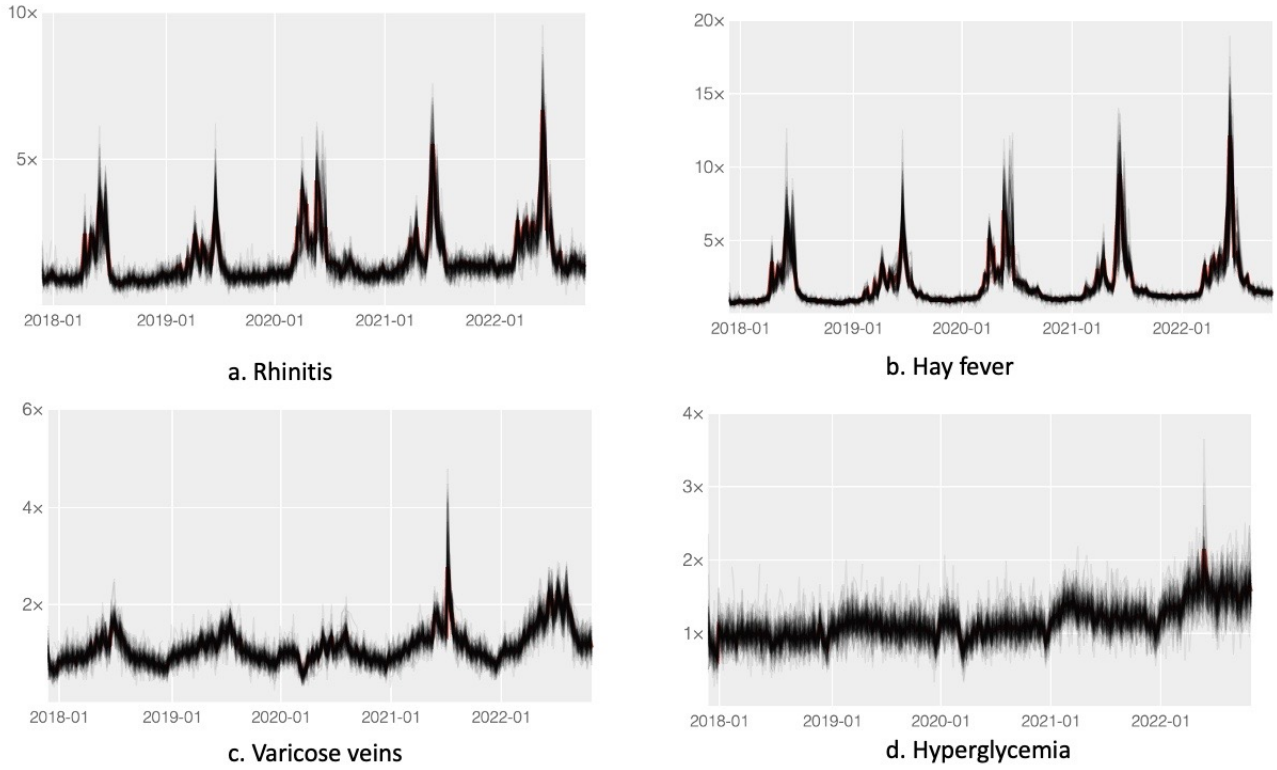


Figure 4.5: Normalized search volume ratios for search terms across all sub-regions in the UK (COVID-19 Search Trends symptom dataset [81]).

"hay fever", "rhinitis" and "shortness of breath" five terms as multiple terms. Furthermore, we also compared these different data representations (i.e., utilizing the moving average in our case) at the same time. As shown in Fig. 4.4, it's clear that using only one of the terms (i.e., "rhinitis") could introduce the most training error during the training phase. The training error decreases as the number of terms increases, which could be interpreted as evidence that COVID-19 involves a wide variety of symptoms. A moving average is more likely to provide a lower training error when compared to a time series representation.

In addition, we compared the Pearson correlation between the top ten and lowest ten search terms in the UK and the USA. From Table 4.3 and 4.4, it can be seen that search terms vary between countries, which highlights the importance of location-specific regression. It has been found that search terms with a high Pearson correlation

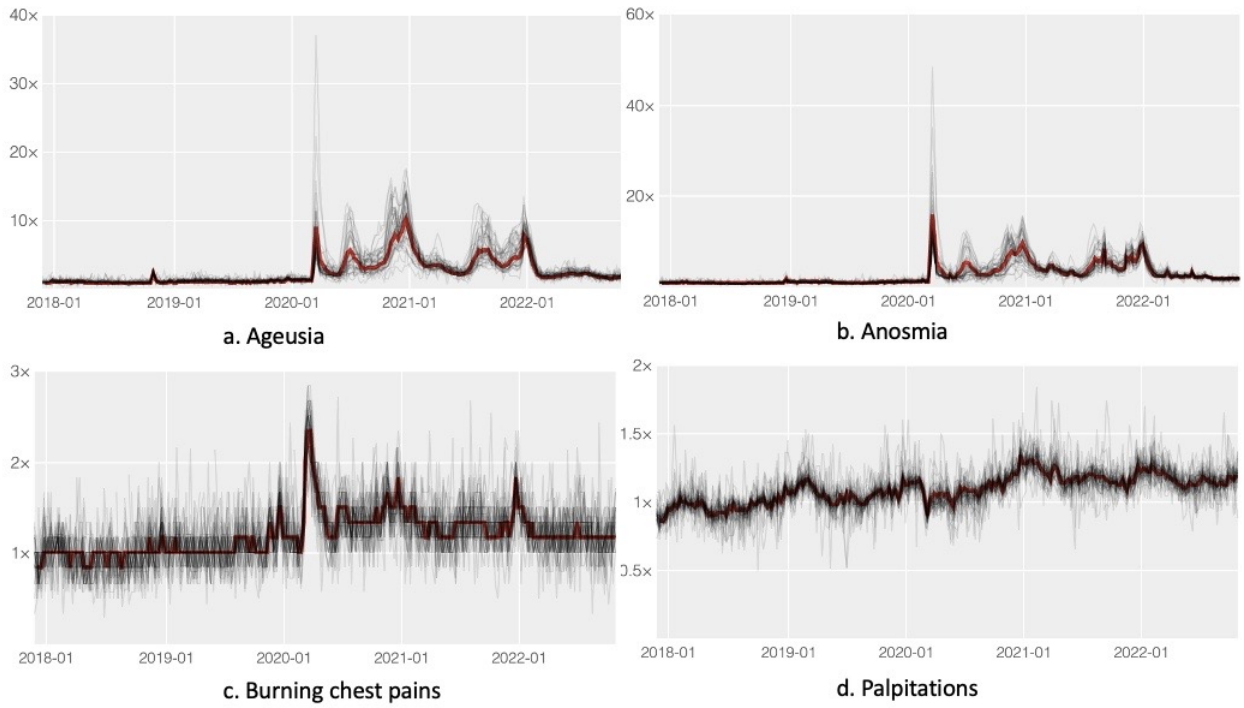


Figure 4.6: Normalized search volume ratios for search terms across all sub-regions in the USA (COVID-19 Search Trends symptom dataset [81]).

are relevant to the symptoms of COVID-19 within a country. In the future, we will also explore the potential of these terms and leverage them as part of our work.

Furthermore, to validate our model, we also visualize the search term trends within UK and USA in Fig. 4.5 and Fig. 4.6. This COVID-19 Search Trends Symptoms dataset [81] provides aggregated, anonymous trends in the Google searches for over 400 health symptoms, signs, and conditions, such as cough, fever, difficulty breathing, and other health conditions that are commonly searched for online. For each region, the dataset gives a time series of the number of searches that have been conducted for each of the symptoms over time. These charts about symptom searches in the United Kingdom could display various types of data related to the frequency and distribution of online searches for COVID-19 symptoms across different regions in the country. We can observe that the conditions 'Rhinitis', 'Hay fever' for UK and 'Ageusia', 'Anosmia' for USA contribute a higher frequency with time, which aligns

with our model's use. We also observe 'Varicose veins', 'Hyperglycemia' for UK and 'Burning chest pains', 'Palpitations' for USA contribute lower frequency with time, which aligns with our observations with their low Pearson correlation with COVID-19 cases. Fig. 4.5 and Fig. 4.6 show the search terms trend along with time. On these charts, the peak indicates that there have been more searches related to the search term. According to the search trends across all sub-regions, we observe that they share a similar trend during the progress of COVID-19, which validates our design to include geographical proximity information in our model design.

In conclusion, I introduced a novel approach to pandemic forecasting combining web search activity data and location relationships in a graph. The proposed framework, SMPNN, merges the best of existing message passing networks and location-based regression models. The method was validated using two real-world COVID-19 datasets and was shown to outperform prior state-of-the-art models, particularly in the early stages of outbreaks, by incorporating spatial graph features. This work makes significant advancements in the field of disease surveillance and forecasting, offering a new approach, methodology, datasets, and insights that integrate web search data and spatial information.

While foundation models like TimesFM[29] offer substantial zero-shot performance due to extensive pretraining on general datasets, their adaptability to highly specialized domains, such as health search trends and infectious disease forecasting, may be limited. The models developed in this research focus on leveraging health-specific data sources, allowing for a more granular and tailored approach to capturing disease transmission patterns and environmental impact on health. This domain-specific training addresses the unique temporal dependencies and seasonality inherent in health data, providing a more effective solution for pandemic forecasting and public health monitoring.

Chapter 5

Search Intent Understanding

In this section, I propose a weakly-supervised method of user search query based on user behavior data. User behavior data contains crucial information to understand users' search intents. Learning user behavior data can improve the semantic understanding of user search queries and enable the understanding of user search queries in a contextual setting. However, current weakly-supervision methods model user click behavior by extracting co-click information based on the URL level, which leads to higher randomness and noise given that the randomness exists in user click behavior. Analyzing the annotations of user-clicked URLs, such as the document title, document labels, and URL text provides a different view of user behavior but often introduces new challenges of how to apply the weakly-supervised method to a different level of user behavior annotations, i.e. weakly supervision of clicked document types and clicked document topics. This research builds on my joint work with Dr. Harshita Sahijwani, focusing on enhancing healthcare search intent recognition under the title "Enhancing Healthcare Search Intent Recognition with Query Representation Learning and Session Context", where Harshita contributed to the multi-label classification part of this research.

In more detail, I propose to apply pre-trained LLMs and semi-supervised ap-

proaches to achieve the learning of user search intents given different levels of user behavior annotations. Specifically, I propose a methodology to better grasp users' intent by weakly supervised learning from clicked documents topics/types and fully exploring the query logs by modeling the user feedback types/topics of clicked URLs/documents. Section 5.2 describes the methodology of weakly-supervised training at the user click URL level and clicked document-type level and document-topic level. The user type a query to the search engine and clicked on the given results as URL links. The search engine and database store the URL links, the document titles and abstracts of the URL, and the document labels. The click annotation information can help us understand user search behavior at different levels.

5.1 Problem Statement

Search engines in the health domain are particularly reliant on their ability to discern user intent to provide relevant results [53, 115]. Users often input a variety of search queries such as symptoms, drugs, specific doctors or health insurance information, expecting the search engine to comprehend these queries and accurately providing user with documents or web pages matching their needs. The categorization of user search intent (i.e., search intent classification) enables the search engines to provide organized and relevant results, which further leads to higher user satisfaction.

Table 5.1: Examples of medical search queries with corresponding search intents.

Medical Search Query	**Search Intent**
lice treatment	Seeking drug and wellness info
bd nano 2nd gen pen needle 32 gauge x 5/32	Seeking drug information
hawaii advance directive form	Managing health accounts; seeking wellness information

Detecting user search intent in general, and in particular in the health domain, is a challenging problem and has been an active area of research for many years. The

main reason for the difficulty in predicting user intent is that medical search queries are inherently ambiguous due to conflation of specialized and colloquial terms, and can be difficult to interpret without context even by human annotators [116, 112], as shown in Table 5.1.

To address this problem, I build upon and expand on the general approach of representation learning. Specifically, I observe that the search logs in the search engine contain substantial information in user search intents [1, 117, 6]. For example, the co-click queries could be used as weak supervision for queries sharing similar search intent and the clicked document or web page annotations could also indicate users' search needs.

Hence, a well established strategy for query intent modeling using search logs is to harness implicit user feedback based on user click behavior for learning query representations [132]. This approach usually involves constructing the query pairs that lead to the same click (termed "co-click" queries) as indicative of similar user intent. Prior studies applied contrastive learning for utilizing the co-click query pairs as weak supervision, where they use a pairwise loss function that ensures that positive (co-click) query pairs are closer to each other, while negative query pairs are farther in their representation space. I review prior work in this area in Section 2.3.2.

However, the effectiveness of this overall approach can degrade in the presence of ambiguous queries (i.e., those for which multiple intents are possible), and for situations like Health where a specialized health search engine may receive only a fraction of the click volume of general-purpose search engines.

Furthermore, the approach of learning single query embeddings for single-label intent recognition may not suffice to represent the multifaceted nature of user queries [125]. As shown in Table 5.1, in the case of a user searching "lice treatment", the intent might span multiple categories, such as seeking drug information ("drug info") and treatment methods ("health wellness"), underscoring the need for a more com-

prehensive recognition strategy. Recognizing the limitations of single-label models in capturing the full spectrum of user intents, I propose a shift to multi-label intent recognition.

Therefore, a natural question is, can I make the representation learning more robust and effective for the multi-faceted nature of Health queries, and the inherent noise in click logs from small-scale search engines? Stated differently, I aim to investigate: **RQ 1: How to improve multi-faceted query representation for Health search, by learning from moderately sized click logs?**

To address RQ1, I introduce a novel multiset loss function, specifically designed to address the inherent ambiguity in user click logs for more accurate user intent learning. Unlike traditional methods, the approach significantly enhances the learning process by utilizing clicked document annotations as the weak supervision. By implementing a clustering-based learning approach, I can effectively harness the rich information embedded in user click logs, thereby allowing for a clustering-driven representation of search queries. This approach not only addresses the previously mentioned challenges but also directly facilitates the downstream search intent classification tasks.

While it is possible to effectively learn a *global* query intent representation, the *individual* users may have a different intent for a query within a specific search session. As shown in Table 5.2, there is a general disagreement between global intents from search logs and session-specific individual intents for identical search queries. This disparity underscores the complexity inherent in accurately understanding user intent based on isolated queries as opposed to considering the entire session context [45, 89]. Such observations motivate the research into the refinement of search query representation in the context of user's search sessions. Therefore, I propose the second research question: **RQ2: How to enhance search query representation from search context for session-based search intent?**

Table 5.2: Disparity between global and session-level search intents for the same search queries

Query (Step 1)	**Query (Step 2)**	**Query (Step 3)**	**Global Intent (Step 3)**	**Session Intent (Step 3)**
healing	acupuncture	chiropractor	Scheduling medical appointments; dealing with billing and coverage	Seeking wellness information
online visit	e visit	zofran	Seeking drug information	Seeking communication information

5.2 Methodology

In this section, I first define the problem setting, and then describe the fine-tuning and search intent prediction methodologies.

5.2.1 Problem Formulation

In this study, I introduce an advanced approach for recognizing healthcare search query intent. The approach aims to fine-tune the search query embeddings using a novel multi-set loss function to align the query embeddings with the clicked document sets, cluster and classify the queries based on the learned embeddings, and further classify the queries in the context of search sessions. Table 5.3 lists all the symbols used in this study. This section is methodically structured into three distinct parts, with two questions focusing on improving search query representation and one downstream task of search intent classification:

Query Representation Learning Let $\mathcal{Q}$ be the set of all user queries and $\phi : \mathcal{Q} \to \mathbb{R}^d$ be the function mapping queries to their d-dimensional vector representations. The goal is to learn the function ϕ such that for any pair of queries $(q, q') \in \mathcal{Q}$, the similarity in the embedding space reflects their semantic similarity.

Table 5.3: List of symbols used in this study.

Symbol	Description
q	A healthcare search query.
$E(q)$	The embedding of the search query q.
(q, q^*)	Pair of queries with same clicked documents (Co-click queries).
(q, q^-)	Pair of queries with different clicked documents.
$\mathcal{C}_i$	Search queries in document set i with same intent.
$\mathcal{S}$	Set of session context vectors.
$\mathbf{s}_i$, $\mathbf{a}_i$, $\mathbf{p}_i$	Search query, clicked document type, and search page context at step i.
K	Number of unique clicked document sets for representation learning.
I	Number of total downstream search intents.

This can be expressed as an optimization problem:

$$\min_{\phi} \sum_{(q,q')\in\mathcal{Q}} \mathcal{L}\left(\phi(q), \phi(q')\right), \tag{5.1}$$

where $\mathcal{L}$ is a loss function measuring the discrepancy between the embeddings of queries q and q', which should be small for semantically similar queries and large for dissimilar ones. The choice of $\mathcal{L}$ depends on the application's specific requirements.

Multi-label Search Intent Classification Let $\mathcal{Q}$ be the set of all user queries and $\mathcal{Y}$ be the set of all potential intent labels, with I being the total number of distinct intents, i.e., $I = |\mathcal{Y}|$. For each query $q \in \mathcal{Q}$, the task is to predict a subset of labels $\mathcal{Y}_q \subseteq \mathcal{Y}$ that accurately reflects the intents of q. This is formulated as a mapping function $f : \mathcal{Q} \to 2^{\mathcal{Y}}$, where $2^{\mathcal{Y}}$ is the power set of $\mathcal{Y}$. The query q is represented by its embedding $\mathbf{E}(q)$, from which a probability distribution over the intent labels in $\mathcal{Y}$ is derived. Labels are then selected based on a thresholding mechanism to form the set $\mathcal{Y}_q$. The objective in training the model is to optimize a loss function that evaluates the accuracy of predicted labels for each query. This process typically involves techniques like Binary Cross-Entropy (BCE) applied across

all I labels. Labels are then selected based on a thresholding mechanism to form the set $\mathcal{Y}_q$.

5.2.2 Proposed Approach

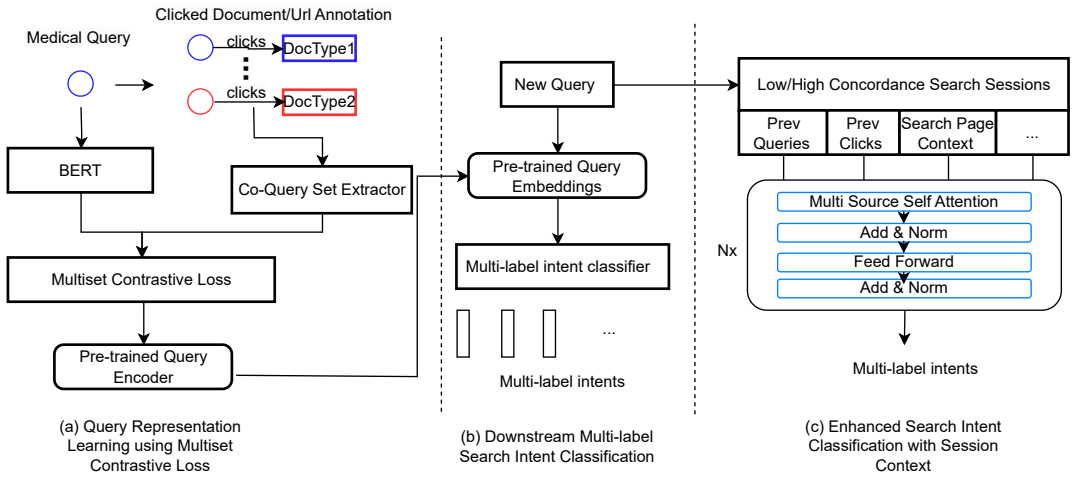


Figure 5.1: A comprehensive approach for query representation learning and intent classification: (a) illustrates the process of leveraging Multiset Contrastive Loss for encoding medical queries using BERT, (b) depicts the application of the resulting embeddings for multi-label intent classification, and (c) shows the enhancement of intent classification by integrating multi-source session context into the model.

Enhancing Query Representation Learning As shown in Figure 5.1a, the methodology utilizes the transformer-based query encoder (i.e. BERT [34]) as the initial query encoder. I enhance query representations through contrastive loss functions $\mathcal{L}_{contrast}$, tailoring the encoder to the intents of health search queries. The contrastive loss functions distinguish between pairs of queries clicked on the same document denoted as (q, q^*) and those clicked on different documents denoted as (q, q^-), ensuring the encoder captures this distinction. The representation of each query, E_{i_j} for q_{i_j} in the document set $\mathcal{C}_i$, is optimized to align closely with other queries in the document set $\mathcal{C}_i$ and diverge from dissimilar queries in the document set $\mathcal{C}_j$, with the total number of document sets denoted as K. This process refines the model to discern the semantic difference of search queries, informed by the user-clicked document

annotations.

Enhancing Query Representation from Search Context In the previous step, the query encoder is learned from the user behavior in aggregated search logs, and the search intents are inferred from the global statistics. However, the search intents could be different in different session contexts. To address this issue, I incorporate session context information during classification-stage fine-tuning to enhance the accuracy of search intent classification. Let $\mathcal{S} = \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ denote the set of session context vectors, each vector $\mathbf{s}_i$ capturing the temporal and relational dynamics of user interactions within a session. The session context includes the search queries $\mathbf{s}_n$, user clicked document annotation $\mathbf{a}_n$ and search page context $\mathbf{p}_n$ at search step n. The process involves analyzing these vectors to predict the categorization of queries at step n. The representation of each query q_i is optimized to align with the corresponding session context $\mathbf{s}_i$, thus enhancing the model's ability to tailor search results to individual user sessions. This methodology refines the query representation to capture user intent more accurately in session context.

Downstream task: Multi-label Search Intent Classification. In the task of multi-label search intent classification, I refine query embeddings and map them to the predefined categories using the classification loss funtion $\mathcal{L}_{class}$. Given the pre-trained query encoder from the previous step, the aim is to predict the probability distribution of each query q across all predefined healthcare intent categories in Y. To achieve this, I transform the encoder's output into a probability vector, where each element corresponds to the likelihood of q belonging to a specific category in Y. This probabilistic approach allows for a comprehensive understanding of q's alignment with each potential intent category. The classification model, denoted as $\mathbf{M}$, takes the pre-trained query encoder and outputs a probability distribution over $\mathbf{Y}$, predicting the most likely category. This process is formalized as $\mathbf{y} = \mathbf{M}(\mathbf{E}(q))$, optimizing the mapping from query embeddings to category predictions.

5.2.3 Pairwise Loss Function

In this study, I focus on fine-tuning the embeddings of search queries to gain deeper insights into user intents. Following previous studies [115, 132], I first employ a pairwise loss function that optimizes the embeddings of the search queries based on their click patterns in a contrastive learning manner.

Formally, let (q, q^*) represent a pair of co-click queries on the same document type, and (q, q^-) represent a pair of queries on different document types. The pairwise loss function, denoted as l_{pairwise}, is defined as follows:

$$l_{\text{pairwise}} = \sum_{q} \left(\frac{1}{1 + \exp\left(\cos\left(E(q), E(q^*)\right)\right)} - \frac{1}{1 + \exp\left(\cos\left(E(q), E(q^-)\right)\right)} \right) \tag{5.2}$$

5.2.4 Multiset Loss Function

The multiset loss function is proposed to mitigate the impact of noise present in user co-click log data. This noise primarily arises from randomly co-clicked query pairs, which can obscure the true intent behind user searches. To address this, the multiset loss function employs a clustering-based approach, contrasting batches of search queries within the same cluster against those in different clusters. This method allows for a more meaningful grouping of queries based on user click patterns by utilizing the clicked document annotations, enhancing the accuracy of the representation learning process.

Central to the multiset loss function is the cosine similarity measure, which is computed between individual query embeddings and the centroid of embeddings corresponding to queries that share the same clicked document type. This approach combines intra-set and inter-set loss components to effectively capture the similarities and differences in query embeddings within and across different document sets.

Intra-Set Loss Function: This component measures the average similarity of query embeddings within the same set. It is formulated as:

$$l_{\text{intra}} = \sum_{i=1}^{K} \frac{1}{N_i} \sum_{j=1}^{N_i} w_{i_j} \frac{1}{1 - \exp\left(\frac{\mathbf{E}_{i_j} \cdot \mathbf{C}_i}{\|\mathbf{E}_{i_j}\| \|\mathbf{C}_i\|}\right) / e + \epsilon} \tag{5.3}$$

where w_{i_j} weights the contribution of each query embedding $\mathbf{E}_{i_j}$ in cluster i, and ϵ is a small value added to prevent division by zero.

Inter-Set Loss Function: Conversely, the inter-set loss function is introduced to maximize the distance between embeddings of queries from different sets. This aspect is crucial for ensuring that queries from distinct sets are not mistakenly grouped together during representation learning. The inter-set loss is defined as:

$$l_{\text{inter}} = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} \frac{1}{N_i} \sum_{k=1}^{N_i} w_{i_k} \frac{1}{1 - \exp\left(\frac{\mathbf{E}_{i_k} \cdot \mathbf{C}_j}{\|\mathbf{E}_{i_k}\| \|\mathbf{C}_j\|}\right) / e + \epsilon} \tag{5.4}$$

where w_{i_k} weights the contribution of each query embedding $\mathbf{E}_{i_k}$ in cluster i, and ϵ is a small value added to prevent division by zero.

Multiset Loss Function: The multiset loss function l_{multiset} elegantly combines these two components, encapsulating the dual objectives of enhancing intra-cluster cohesion and inter-cluster separation:

$$l_{\text{multiset}} = -\log\left(\frac{l_{\text{intra}}}{l_{\text{inter}}}\right) \tag{5.5}$$

The mathematical formulation of the multiset loss, denoted as l_{multiset}, is presented in Equation 5.5. Here, K denotes the number of unique clicked document sets. For each label i, N_i is the number of queries associated with that label, $\mathbf{E}_{i_j}$ represents the embedding of the j^{th} query in document set i, and $\mathbf{C}_i$ is the centroid of embeddings for set i. The loss function computes the ratio of the average intra-cluster cosine

similarity to the average inter-cluster cosine similarity, thus encouraging the model to form tightly knit clusters of queries with shared intents while distancing those with differing intents.

The "sets" used in the multiset cosine similarity loss could be document types or document topics. The loss is calculated as the negative log of the ratio of the average cosine similarity of each set to the average cosine similarity of all sets. With this design, when N is the average number of queries in each document set in the dataset, the theoretical computational cost of multiset loss is $O(K^2 * N)$. As opposed to pairwise loss, which has a theoretical complexity of $O(K * N^2)$, multiset loss is computationally more efficient, given that $K \ll N$, in most conditions.

An essential factor in improving session-based intent recognition is preserving the ordering information within sessions. When full session data is available during pretraining, maintaining session order is crucial, as the sequence of user interactions provides valuable context. In our experiments, the full session data is limited (due to strict data processing and supervision by our collaborators). However, we emphasize that pretraining the model with session-specific ordering, while preserving location and context, would enable more accurate interpretation of session-based intent by closely aligning with real-time user behavior patterns.

5.2.5 Multi-Label Search Intent Classification

In this study, I treat multi-label search intent classification as the major downstream task after query representation learning. The objective of multi-label intent classification is to assign a set of intent labels to a given input text.

Model Architecture The model architecture is based on the Transformer-based BERT model. Given the pre-trained BERT model, I further fine-tune the model on the downstream intent classification dataset using a custom loss function designed for multi-label classification. I employ the pre-trained MSet-BERT to obtain dense

vector representations for each search query. For the sequence of search query terms, I get the final vector corresponding to the [CLS] token, which is designed to hold the aggregate sequence representation, and is then passed through a series of fully connected layers for each search intent.

Multi-Label Loss Function For training the model, I use the Binary Cross-Entropy (BCE) loss, treating each label as an independent binary classification. The BCE loss for a single instance (x, Y) with the predicted output $\hat{Y}$ is defined as:

$$\text{BCE}(Y, \hat{Y}) = -\frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{5.6}$$

where y_i is the binary indicator (0 or 1) if label i is the correct classification for x, and $\hat{y}_i$ is the predicted probability of x having label i.

5.3 Experimental Setting

In this part, I have two major tasks. First, I want to validate the approach by clustering the embeddings of search queries and identifying the search queries within the same cluster as from the same search intent. Second, I want to demonstrate the approach can be applied to real-world search scenarios by giving a search query, and correctly classifying it to a search intent.

The metric of the first task is adjusted rand index (ADI) and normalized mutual information (NMI) and the metric of the second task is accuracy and F1 score. ADI and NMI can measure how the clusters of search queries match the distribution of their intents. Accuracy and F1 score can measure the performance of classifying intents of given search queries. I aim to evaluate the methods on two datasets, i.e. KP dataset and the TripClick dataset. The detail of the dataset is discussed in the following section.

5.3.1 Datasets

We experiment with two datasets: a private Health Search (HS) dataset and a public TripClick dataset.

Health Search (HS) Dataset The HS dataset, collected from a health website internal search engine, spans Jan 2022 to Sept 2023. It includes queries, clicked documents, query-document pairs, and document attributes like titles, document URLs, and document types. When I generate user-clicked data, I filter the click logs by the click count greater than 2. For session-based search intent classification, I specifically set aside the data from January 2022, which is not included in the representation learning dataset. This exclusion is methodologically significant as it allows for a focused comparison between global and session-level search intents using a distinct and controlled dataset. January 2022 was chosen for this analysis to provide a clear baseline for assessing how individual search intents align or diverge over global search intent, thus offering insights into the user search behavior in healthcare contexts.

TripClick Dataset TripClick is a large-scale dataset of click logs in the health domain from the Trip Database health web search engine[96]. The TripClick dataset contains about 5.2 million user interactions collected between 2013 and 2020 and contains the following information: the queries, the clicked documents, and document attributes. The document attributes include the document title, the document URL, and the document type. For this dataset, I filter the click logs by the click count greater than 5, which leads to all unique queries for the query representation learning. Similar to HS dataset, I specifically set aside the data from January 2018 and February 2018, which is not included in the representation learning dataset for session-based search intent classification.

Co-Query Set Extractor

The Co-Query Set Extractor groups queries into different sets based on their interactions with clicked documents. It analyzes clicked document annotations to identify patterns and similarities among queries. This approach clusters queries that exhibit similar user engagement or seek related information. For instance, if multiple queries lead users to click on the same set of documents (i.e. same document type or document URL pattern), these queries are considered related and grouped together.

Intent Annotation

Understanding user intent is key to personalized healthcare experiences. It involves a detailed methodology to accurately annotate healthcare queries with specific user intents. I employ heuristic methods to systematically identify and label these intents. For example, a query like 'symptoms of flu' would be annotated with the intent of seeking medical information based on the analysis of the web pages associated with that query.

HS query intent annotation In the HS dataset, I aim to leverage intent classification to categorize queries into some navigational intents (e.g., "Book Appointment", "Access Records") and topical intents like "Drug Information". Identifying these intents allows us to proactively surface relevant components ("Find Doctors", "Locations") directly addressing their needs. A semi-automated approach based on search logs is employed to generate the intent labels for such a classification. This method utilizes specific patterns and keywords within clicked URLs to infer user intent. A set of URL patterns is first curated for each query intent, guided by the site's organized document structure and expert knowledge. The query intent is identified based on the user's query and the matching pattern from the most frequently clicked URL. This methodology considers all top-clicked URLs per query, assigning one or more

labels based on their corresponding classes.

TripClick query intent annotation In the TripClick dataset, detailed document annotations provide insights into search intent. I analyze the click patterns of queries on different document types and use these interaction data to infer search intent. This involves collecting statistical data on query clicks for each document type. These clicks are then weighted to reflect their importance and associated with relevant document types as intents in the test dataset.

5.3.2 Experimental Setup for Intent Classification

I create the train, validation, and test sets for intent classification using a variation of stratified sampling in approximately 60:20:20 ratio. Since I perform multi-label classification, I first group the queries by label combinations. I then divide the samples from each group among the three sets.

For the HS dataset, all sets have the same proportion of class labels/class label combinations. Their sizes are in the ratio 60:20:20. All three sets have similar distributions over class labels.

The TripClick dataset has a larger number of classes (document types) and some combinations of labels are extremely infrequent. For simplicity, I first create the train-validation-test split with labels that have at least 3 samples. I then add all the queries with unique label combinations to the test set.

5.3.3 Experimental Setup for Session-based Intent Classification

For session-based intent classification, I carefully curate search sessions based on their sequence lengths, specifically ranging from 2 to 6 and session length of 4, as per [89] suggested, was used for representing contextual search. This range is chosen to

effectively separate the datasets for training, validation, and testing, ensuring each set represents varying complexities of user interactions. For the training set, I utilize sessions from search step 1 to search step n-1, focusing on building the training model. The validation and test sets comprise sessions from search step 2 to search step n, aiming to determine the classification threshold and evaluate the model's performance. The specific criteria for session selection, rooted in sequence length, are designed to represent real-world user search behavior, thus enhancing the relevance and implication of the classification task.

5.3.4 Baseline Models

BERT BERT model [34] is a pre-trained language model trained on the large-scale corpus. It can be fine-tuned on the downstream tasks via either the contrastive loss functions or classification loss functions. The BERT base model (uncased) is used as one of the baseline models and tailored to the study's objectives.

PairWise-BERT The PairLoss-BERT model represents an advanced baseline [132, 115]. It undergoes pre-training with a contrastive learning approach using the pairwise loss function. This model sets a state-of-the-art baseline for the approach, demonstrating the effectiveness of contrastive learning in query representation.

MSet-BERT (our method) with variants for ablation studies We report performance of several variants of the MSet-BERT model to assess the impact of different session contexts in session-based search intent classification:

- **BERT w all context (Ablation):** Enhances the baseline BERT model by incorporating complete session context.
- **PairWise-BERT w all context (Ablation):** Builds upon the PairWise-BERT model by integrating full session context.

- **MSet-BERT w/o context:** Uses only the search queries as input, serving as a baseline to understand the model's performance without any session context.
- **MSet-BERT w prev-query:** Includes the previous query in the session as part of the context.
- **MSet-BERT w page-context:** Incorporates the context of the web page the user search happens.

5.3.5 Evaluation Metrics

The performance is measured using appropriate multi-label metrics, including Precision, and F1 score. In addition, I also order the probability of the intents and evaluate the rank using metrics including Hit Rate@3 and NDCG@3 for multiple intents retrieval.

Perplexity measures the intent distribution of a given query, its intent distribution can be represented as a series of probabilities $P(label_1), P(label_2), ..., P(label_n)$, which is calculated from query click weights in search logs. The calculation of Perplexity is based on these probability values and is defined as:

$$\text{Perplexity} = 2^{-\sum_{i=1}^{n} P(label_i) \log_2 P(label_i)} \tag{5.7}$$

Precision measures the proportion of correctly predicted positive observations to the total predicted positives for each query:

$$Precision = \frac{TP}{TP + FP} \tag{5.8}$$

where TP is the number of true positives and FP is the number of false positives.

The F1 score is the harmonic mean of Precision and Recall, providing a balance between them. It is particularly useful in scenarios where I have an uneven class

distribution. The F1 score is given by:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.9}$$

Hit Rate@3, often used in ranking problems, measures the proportion of times the correct label is within the top 3 predictions. It is a recall-based measure at the top of the ranking list and is calculated as follows:

$$HitRate@3 = \frac{1}{N} \sum_{i=1}^{N} I(y_i \in top3(\hat{Y}_i)) \tag{5.10}$$

where N is the number of samples, y_i is the true label for the i^{th} sample, $\hat{Y}_i$ is the set of top 3 predicted labels, and I is the indicator function.

Normalized Discounted Cumulative Gain at rank 3 (NDCG@3) evaluates ranking quality, considering the position of the correct intent label in the predicted ranking list. Higher ranks receive more weight. NDCG is calculated as:

$$NDCG@3 = \frac{1}{N} \sum_{i=1}^{N} \frac{DCG@3_i}{IDCG@3_i} \tag{5.11}$$

where $DCG@3_i$ is computed from the predicted probability vector $\mathbf{y}_i$ against the true label vector, representing a weighted vector of correct classification. $IDCG@3_i$ is the ideal ranking gain, the highest possible $DCG@3$ given the set of intent labels. The DCG computation considers the relevance score derived from ground truth $\mathbf{y}_i$ and the position of the label in the predicted vector $\hat{\mathbf{y}}_i$.

Table 5.4: Comparative clustering performance analysis of query representations from BERT, PairWise-BERT, and MSet-BERT models on HS and TripClick test datasets using adjusted rand index (ARI) and normalized mutual information (NMI). The best performing values are highlighted in bold.

Dataset	**Models**	**ARI**	**NMI**
HS	BERT	0.0540	0.0859
	PairWise-BERT	0.0690	0.1161
	MSet-BERT	**0.0723** (+4.78%)	**0.1181** (+1.72%)
TripClick	BERT	0.00189	0.0120
	PairWise-BERT	0.00213	0.0113
	MSet-BERT	**0.00235** (+10.33%)	**0.0122** (+1.67%)

Table 5.5: Evaluation of model performance on the HS and TripClick datasets across multiple metrics: Precision, F1, Hit Rate@3, and NDCG@3. The highest performing values are marked in bold, with asterisks denoting significant improvements over the best baseline model. Statistical significance is determined by a t-test for N queries in the test dataset, with $p < 0.05$.

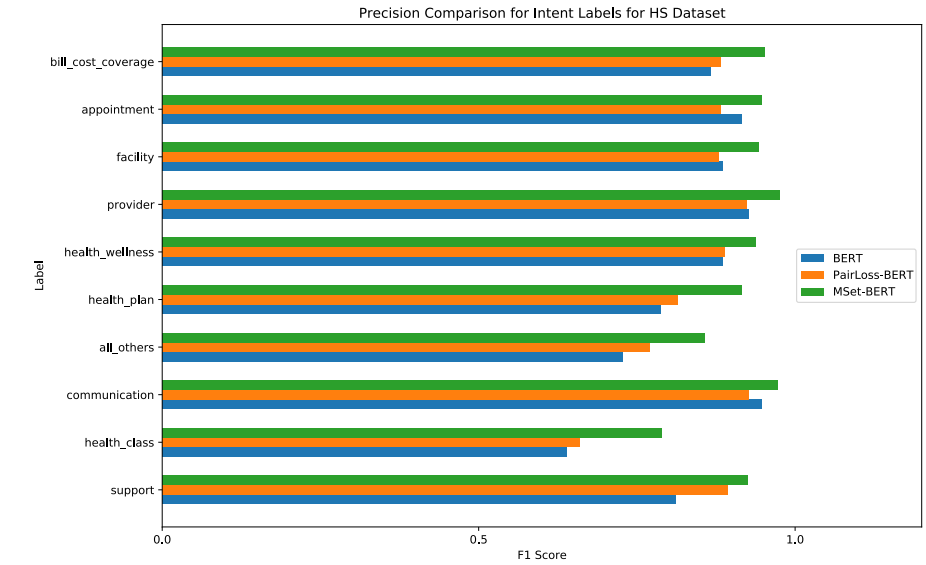
Dataset	**Model**	**Precision**	**F1**	**Hit Rate@3**	**NDCG@3**
HS Dataset	BERT	0.937	0.932	0.976	0.937
	PairWise-BERT	0.943	0.936	0.988	0.940
	MSet-BERT	**0.970*** (+2.86%)	**0.969*** (+3.53%)	**0.992*** (+0.40%)	**0.975*** (+3.72%)
TripClick Dataset	BERT	0.870	0.813	0.947	0.853
	PairWise-BERT	0.881	0.840	0.956	0.875
	MSet-BERT	**0.895*** (+1.59%)	**0.854*** (+1.67%)	**0.965*** (+0.94%)	**0.886*** (+1.26%)

5.4 Results

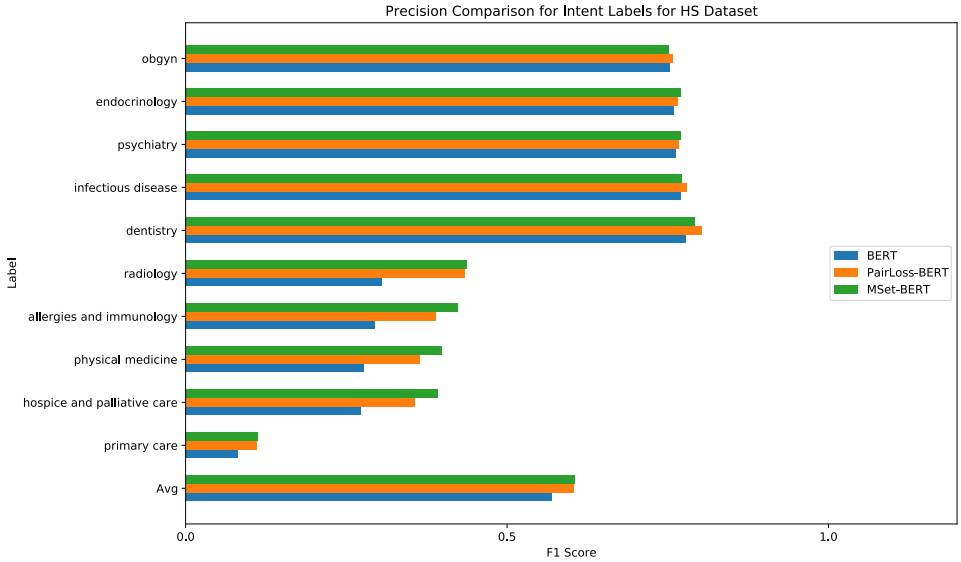
In this section, I describe the performance analysis of the MSet-BERT model, underscoring its advancements in multi-label search intent classification and session-based search intent classification over baseline models.

Table 5.6: Comparative performance of MSet-BERT models in session-based intent classification on the HS and TripClick datasets, highlighting the impact of different context integration strategies (no context, previous query, page context, and all contexts). Performance metrics include Precision, F1, Hit Rate@3, and NDCG@3, with the best scores highlighted in bold and marked with an asterisk (*) to indicate significant improvement. Statistical significance is determined by a t-test for N queries in the test dataset, with $p < 0.05$. Session length of 4, as per [89], was used for this analysis.

Dataset (Concordance)	**Model (Ablation)**	**Precision**	**F1**	**Hit Rate@3**	**NDCG@3**
HS (33%)	BERT w all context (Abl)	0.715	0.722	0.856	0.658
	PairWise-BERT w all context (Abl)	0.736	0.749	0.866	0.678
	MSet-BERT w/o context	0.707	0.708	0.864	0.653
	MSet-BERT w prev-query	0.721	0.753	0.871	0.669
	MSet-BERT w page-context	0.759	0.751	0.869	0.667
	MSet-BERT w all context	**0.776*** (+5.44%)	**0.784*** (+4.67%)	**0.880*** (+1.61%)	**0.769*** (+13.42%)
TripClick (88%)	BERT w all context (Abl)	0.829	0.814	0.931	0.765
	PairWise-BERT w all context (Abl)	0.846	0.820	0.936	0.772
	MSet-BERT w/o context	0.831	0.813	0.929	0.762
	MSet-BERT w prev-query	0.844	0.824	0.941	0.778
	MSet-BERT w all context	**0.868*** (+2.60%)	**0.840*** (+2.44%)	**0.953*** (+1.81%)	**0.826*** (+7.00%)



a) F1 scores in HS Dataset



b) F1 scores in TripClick Dataset

Figure 5.2: Comparative analysis of F1 scores for different intent types within the HS and TripClick dataset, providing insights into the model's performance in accurately classifying and retrieving relevant search intents.

5.4.1 Performance Comparison

In the study of the MSet-BERT model for multi-label search intent classification and session-based search intent classification, I observed significant improvements over baseline models. In the evaluation of query representation clustering, as detailed in

Table 5.4, the MSet-BERT model consistently outperforms the BERT and PairWise-BERT model on the ARI and NMI score for query representation clustering. Specifically, for the HS dataset, MSet-BERT achieves a 4.78% higher ARI and a 1.72% improvement in NMI compared to the best baseline. This trend is similarly observed in the TripClick dataset, where MSet-BERT shows a substantial improvement of 10.33% in ARI and 1.67% in NMI scores over the other models. These results underscore the efficacy of MSet-BERT in clustering query representations from search logs to match users' search intent. In the evaluation of the downstream intent classification task, the MSet-BERT model also demonstrate significant improvement over the baseline models. Specifically, as in detailed Table 5.5 and Figure 5.2, MSet-BERT demonstrates an improvement of 2.86% in precision, 3.53% in F1 score, 0.40% in Hit Rate@3, and a notable 3.72% in NDCG@3 for HS dataset. Simlarly, for the TripClick dataset, the model exhibits improvements of 1.59% in precision, 1.67% in F1 score, 0.94% in Hit Rate@3, and 1.26% in NDCG@3 compared to the best performing baseline model. These improvements underscore the effectiveness of MSet-BERT in enhancing the robustness and accuracy of search intent classification for health-related queries.

Session-based Intent Classification Performance

In session-based intent classification, predicting user user search intents is a more difficult task because it requires the model to handle the situations where the user search behavior deviates or disagrees with the search-query-based intent. In the evaluation of MSet-BERT in the session-based search intent classification task, as described in Table 5.6, the MSet-BERT model demonstrate notable effectiveness compared with all baseline models. Specifically, the MSet-BERT model exhibits a marked improvement in precision (up to 5.44% for HS and 2.60% for TripClick), F1 score (increased by 4.67% for HS and 2.44% for TripClick), Hit Rate@3 (improved by 1.61% for HS and 1.81% for TripClick), and NDCG@3 (a substantial gain of 13.42% for HS and

7.00% for TripClick). The ablation studies without context or with only a part of session contexts also underscore teh effectiveness of the MSet-BERT in harnessing context integration strategies for session-based intent classification.

5.5 Analysis & Discussion

In this section, I'll analyze the performance of the proposed methodology in more detail and discuss the limitations when adapting the methodology to general-purpose intent recognition tasks.

Ambiguity and Difficulty in Session-based Intent Classification

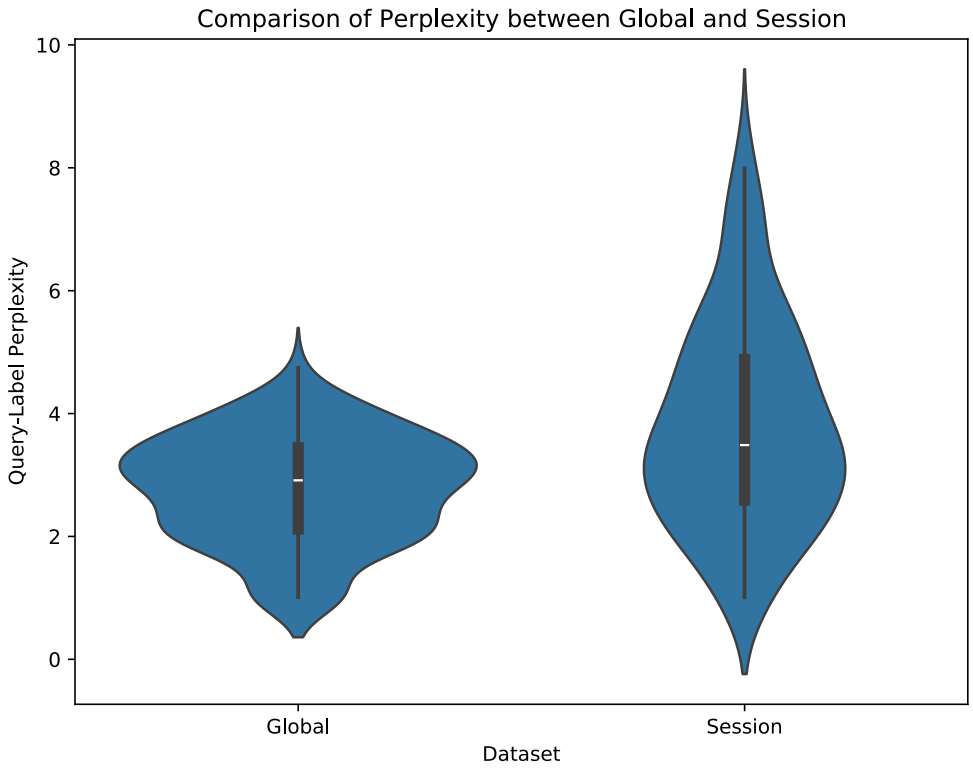


Figure 5.3: Comparison of query perplexity for the global and session-specific intent classification of 140 common queries in the HS dataset.

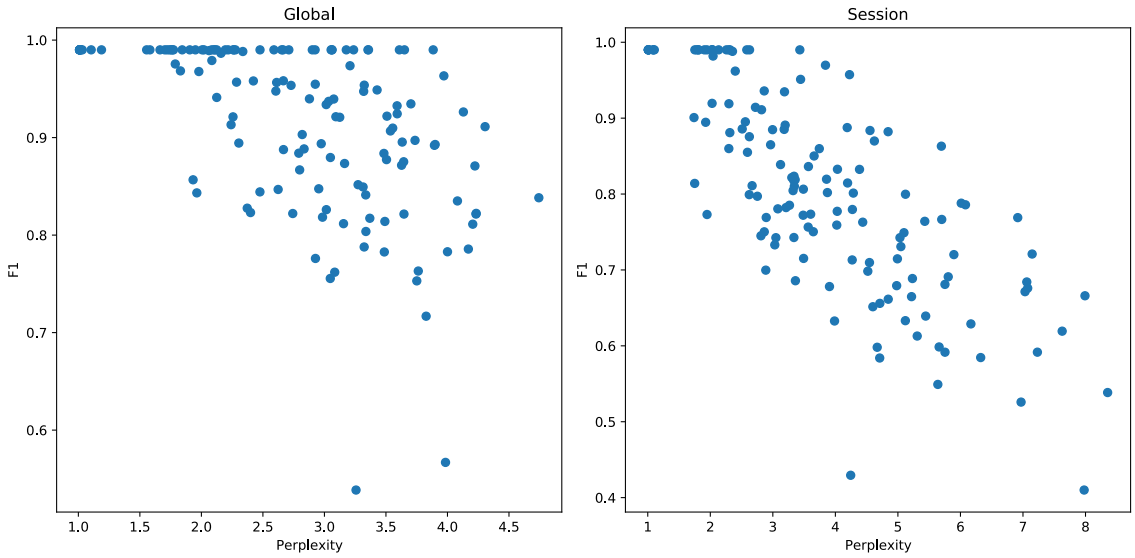


Figure 5.4: Comparison of query perplexity and F1 scores for the global and session-specific intent classification of 140 common queries in the HS dataset.

A core challenge in session-based intent classification is recognizing and addressing the ambiguity of intent for the same queries across different sessions. As described in Figure 5.3, the same queries when collected for global query intent (derived from the number of global co-occurrence in multiple sessions) and for session-based search intent (defined as the subsequent action in a single session) can exihigit different query intent perplexity, where higher perpleixty means higher ambiguity. Further analysis, as shown in Figure 5.4, indicates a notable inverse correlation betweeen the query perplexity and F1 scores, suggesting greater difficulty in classification. This implies that accurately identifying intents in session-based queries involves not only considering global intents but also intricately analyzing the context of each session.

General-purpose intent recognition models must accommodate a wide range of queries, which introduces noise and reduces the precision of intent categorization in health-specific applications. By focusing on health-related user interactions, such as click behaviors in search engines or conversational agents, this research achieves a more refined approach to identifying user intent within a healthcare context. Additionally, leveraging domain-specific data and tuning models accordingly offers the

flexibility to address challenges such as ambiguous queries and session-based variations in intent, which are common in health-related searches.

Chapter 6

Conclusion

This dissertation presents my work on modeling search trends and search interest for health. The details of the contributions are listed as follows:

6.0.1 Modeling Search Trend for Air Pollution Detection

In Chapter 3, I incorporated the information from different modalities for multimodal sequential learning with missing values. Theoretical analysis and experimental evaluation in this dissertation demonstrate that the proposed method can effectively model the health data with missing values by exploring intra-modal and inter-modal dynamics. Unlike previous state-of-the-art methods, this approach can effectively utilize the information from different modalities to improve the prediction performance. I investigated the novel problem of exploring intra-modality and inter-modality dynamics for multi-modal sequential learning with missing values. I propose a new framework, CMFN, which uses modality-specific and cross-modal information to impute missing values. To validate the framework, I tested it using both real-world and synthetic data on benchmark multi-modal learning datasets. Our results outperform existing state-of-the-art models, and ablation studies highlight the architectural advantages.

The study on air pollution prediction is the first to demonstrate that online search

interest data, although noisy, can complement ground sensor measurements to indicate several urban air pollutants when combined properly. I achieved this by presenting a traditional machine learning model (Random Forest) and a novel neural network model (DL-LSTM) that incorporates search interest data with traditional air pollution predictors (i.e., historical pollutant concentration, temperature, and relative humidity). Our results show that the proposed model benefited from the addition of search interest data, whereas traditional models often suffered from the added noise and increased dimensionality.

6.0.2 Modeling Search Trend for Infectious Disease Forecasting

In Chapter 4, a novel approach to pandemic forecasting is introduced, combining web search activity data and location relationships in a graph. The proposed framework, SMPNN, merges the best of existing message passing networks and location-based regression models. The method was validated using two real-world COVID-19 datasets and was shown to outperform prior state-of-the-art models, particularly in the early stages of outbreaks, by incorporating spatial graph features. This work makes significant advancements in the field of disease surveillance and forecasting, offering a new approach, methodology, datasets, and insights that integrate web search data and spatial information.

6.0.3 Search Intent Understanding

In Chapter 5, I proposed a novel method for learning query representation, focusing on the Health domain, resulting in an effective query embedding model MSet-BERT. The method proposed is general and makes use of a novel multiset loss function designed to capture the inherent ambiguity of health search queries, which results

in an enhanced search query representation. This MSet-BERT model demonstrated its advantages over prior open state-of-the-art models of BERT and PairWise-BERT models trained on search logs, for both intrinsic query clustering tasks, and for multi-label intent classification tasks.

Furthermore, this study investigated the effectiveness of different context representations to improve session-based intent prediction, demonstrating that the MSet-BERT model generalizes to the more challenging individual session-based intent prediction task, even when the individual search intent disagrees with the most popular intent for the query based on global statistics. The experiments show that adding context information from previous queries and clicked documents can improve the performance of the MSet-BERT model for the search session query intent recognition.

6.0.4 Limitations

Although the proposed methods effectively address the key questions of the dissertation, there are several limitations that should be considered. CMFN method in Chapter 3 is limited by the assumption that the missing values are missing at random (MAR) and the information can be retrieved from the other modalities. This assumption may not hold in real-world scenarios, where the missing values may be missing not at random (MNAR) or the information may not be available in other modalities. The proposed DL-LSTM model in Chapter 3 addresses the correlation between the search queries by incorporating the semantic information, but the noise in search trends in real-world scenarios may affect the prediction performance. The experiments for SMPNN model in Chapter 4 relies on the assumption that the cross-location information can be used to improve the prediction performance. However, the prediction performance may be affected by the quality of the cross-location information and the availability of the data. The proposed search query representation

learning and intent prediction method in Chapter 5 relies on the assumption that the search queries can be represented by the user click results in the sessions and the search intent can be predicted based on the session context. However, the prediction performance may be affected by the quality of the session context and the availability of the data.

6.0.5 Future Work

This dissertation opens up several avenues for future research on modeling search trends and search intent for online health monitoring. To extend the proposed methods, future work could focus on the following directions. For the CMFN model, the intra-modality and inter-modality dynamics can be further explored by considering the noise in the data. Future work could investigate the impact of the noise in the data on the prediction performance and propose new methods to address the noise. For the noise in the search trends data, future work should focus on the long-term impact of the noise on the prediction performance and propose methods to improve the prediction performance. For representation learning and intent prediction, future work could focus on the amguious search queries and the multiple intents in different sessions. Future work should explore scenarios where a single query may carry multiple distinct intents within a search session and develop new methods to effectively interpret and address these varying intents.

Building on these directions, another promising area for future research is leveraging user interaction data from conversational agents, such as ChatGPT or other AI-powered chat platforms, to enhance health-related search intent modeling. As chat agents become increasingly popular, they offer a rich source of data on user behavior, particularly in health information-seeking contexts. This user interaction data, if made accessible, could provide valuable insights into users' evolving informational needs and engagement patterns, allowing for a more nuanced understanding

of search intent. Additionally, these interactions often occur in multi-round conversations, where users reformulate or clarify queries over the course of the session. This multi-round data could be treated as session-based interactions, closely aligning with the session-based intent recognition models developed in this dissertation. By incorporating conversational reformulations and session dynamics, the user-behavior data-based models developed here could be adapted to capture intent more dynamically, providing real-time responses and insights. This approach would enable health monitoring systems to become more responsive and personalized, enhancing their utility in supporting public health and individual health information needs in an era of conversational AI.

Overall, this dissertation contributes to modeling search trends and search interests for health by proposing new methods to address critical challenges in each application scenarios.

Bibliography

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, 2006.

[2] Qingyao Ai, Tao Yang, Huazheng Wang, and Jiaxin Mao. Unbiased learning to rank: online or offline? *ACM Transactions on Information Systems (TOIS)*, 39(2):1–29, 2021.

[3] Emily L Aiken, Andre T Nguyen, and Mauricio Santillana. Towards the Use of Neural Networks for Influenza Prediction at Multiple Spatial Resolutions. *arXiv*, 2019.

[4] Benjamin M Althouse, Samuel V Scarpino, Lauren Ancel Meyers, John W Ayers, Marisa Bargsten, Joan Baumbach, John S Brownstein, Lauren Castro, Hannah Clapham, Derek AT Cummings, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ data science*, 4(1): 1–8, 2015.

[5] Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, 32(3):259–280, 2018. ISSN 0895-3309. doi: 10.1257/jep.32.3.259.

[6] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 185–194, 2012.

[7] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] Cole Brokamp, Roman Jandarov, MB Rao, Grace LeMasters, and Patrick Ryan. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. *Atmospheric Environment*, 151:1–11, 2017.

[9] David A Broniatowski, Michael J Paul, and Mark Dredze. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672, 2013.

[10] Erik Brynjolfsson, Tomer Geva, and Shachar Reichman. Crowd-squared: amplifying the predictive power of search trend data. *Brynjolfsson, E., Geva, T., & Reichman, S., Crowd-Squared: Amplifying the Predictive Power of Search Trend Data. MIS Quarterly (Forthcoming)*, 2015.

[11] Sheen Mclean S. Cabaneros, John Kaiser Calautit, and Ben Richard Hughes. A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software*, 119:285–304, 2019. ISSN 1364-8152. doi: 10.1016/j.envsoft.2019.06.014.

[12] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional Recurrent Imputation for Time Series. *arXiv*, 2018.

[13] Herman Anthony Carneiro and Eleftherios Mylonakis. Google trends: a web-

based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564, 2009.

[14] Yan Carrière-Swallow and Felipe Labbé. Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4):289–298, 2013.

[15] Centers for Disease Control and Prevention. Data brief 482. `https://www.cdc.gov/nchs/products/databriefs/db482.htm`, 2024. Accessed: 2024-05-14.

[16] Emily H Chan, Vikram Sahai, Corrie Conrad, and John S Brownstein. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS neglected tropical diseases*, 5(5):e1206, 2011.

[17] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2 (3):1–27, 2011.

[18] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

[19] Sheng Chen, Guangyuan Kan, Jiren Li, Ke Liang, and Yang Hong. Investigating china's urban air quality using big data, information theory, and machine learning. *Polish Journal of Environmental Studies*, 27(2), 2018.

[20] Xinyu Chen, Zhaocheng He, Yixian Chen, Yuhuan Lu, and Jiawei Wang. Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model. *Transportation Research Part C: Emerging Technologies*, 104:66–77, 2019. ISSN 0968-090X. doi: 10.1016/j.trc.2019.03.003.

[21] Cynthia Chew and Gunther Eysenbach. Pandemics in the Age of Twitter:

Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, 5 (11):e14118, 2010. doi: 10.1371/journal.pone.0014118.

[22] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv*, 2014.

[23] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[24] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

[25] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. *Synthesis lectures on information concepts, retrieval, and services*, 7(3): 1–115, 2015.

[26] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[27] Aaron J Cohen, Michael Brauer, Richard Burnett, H Ross Anderson, Joseph Frostad, Kara Estep, Kalpana Balakrishnan, Bert Brunekreef, Lalit Dandona, Rakhi Dandona, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. *The Lancet*, 389(10082):1907–1918, 2017.

[28] Patrick Copeland, Raquel Romano, Tom Zhang, Greg Hecht, Dan Zigmond,

and Christian Stefansen. Google disease trends: An update. In *International Society of Neglected Tropical Diseases 2013*, page 3, 2013.

[29] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

[30] Audrey De Nazelle, Edmund Seto, David Donaire-Gonzalez, Michelle Mendez, Jaume Matamala, Mark J Nieuwenhuijsen, and Michael Jerrett. Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environmental Pollution*.

[31] Google DeepMind. Introducing gemini: our largest and most capable ai model, 2024. URL `https://blog.google/technology/ai/google-gemini-ai/`. Accessed: October 20, 2024.

[32] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Graph Message Passing with Cross-location Attentions for Long-term ILI Prediction. *arXiv*, 2019.

[33] Srinivas Devarakonda, Parveen Sevusu, Hongzhang Liu, Ruilin Liu, Liviu Iftode, and Badri Nath. Real-time air quality monitoring through mobile sensing in metropolitan areas. In *Proc. of SIGKDD international workshop on urban computing*.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[35] Qian Di, Yan Wang, Antonella Zanobetti, Yun Wang, Petros Koutrakis, Christine Choirat, Francesca Dominici, and Joel D Schwartz. Air pollution and

mortality in the medicare population. *New England Journal of Medicine*, 376 (26):2513–2522, 2017.

[36] Shengdong Du, Tianrui Li, Yan Yang, and Shi-Jinn Horng. Multivariate time series forecasting via attention-based encoder–decoder framework. *Neurocomputing*, 388:269–279, 2020.

[37] Jeffrey L Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2-3):195–225, 1991.

[38] J Friedman, T Hastie, and R Tibshirani. The elements of statistical learning. volume 1 springer; new york, ny, usa: 2001.

[39] Cornelius Fritz, Emilio Dorigatti, and David Rügamer. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly covid-19 cases in germany. *Scientific Reports*, 12(1):3930, 2022.

[40] Isaac Chun-Hai Fung, Zion Tsz Ho Tse, and King-Wa Fu. The use of social media in public health surveillance. *Western Pacific surveillance and response journal*.

[41] Isaac Chun-Hai Fung, Zion Tsz Ho Tse, Chi-Ngai Cheung, Adriana S Miu, and King-Wa Fu. Ebola and the social media. 2014.

[42] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[43] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009. ISSN 0028-0836. doi: 10.1038/nature07634.

[44] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, pages 1764–1772, 2014.

[45] Helia Hashemi, Hamed Zamani, and W Bruce Croft. Guided transformer: Leveraging multiple external sources for representation learning in conversational search. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval*, pages 1131–1140, 2020.

[46] Helia Hashemi, Hamed Zamani, and W Bruce Croft. Learning multiple intent representations for search queries. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 669–679, 2021.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. of CVPR*, pages 1026–1034, 2015.

[48] A Herdağdelen, A Dow, S Bogdan, M Payman, and A Pompe. Protecting privacy in facebook mobility data during the covid-19 response. *Facebook Research*, 2020.

[49] Shawndra Hill, Raina Merchant, and Lyle Ungar. Lessons learned about public health from online crowd surveillance. *Big Data*, 1(3):160–167, 2013.

[50] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[51] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735—1780, 1997.

[52] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[53] Bernard J Jansen, Danielle L Booth, and Amanda Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*, pages 1149–1150, 2007.

[54] Seung-Pyo Jun, Hyoung Sun Yoo, and San Choi. Ten years of research change using google trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, 130:69–87, 2018.

[55] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[56] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2017. *ArXiv abs/1609.02907*, 2017.

[57] Tadeusz Kufel et al. Arima-based forecasting of the dynamics of confirmed covid-19 cases for selected european countries. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, 15(2):181–204, 2020.

[58] Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. Leveraging sentence-level information with encoder lstm for semantic slot filling. *arXiv preprint arXiv:1601.01530*, 2016.

[59] Hoeyun Kwon and Caglar Koylu. Revealing associations between spatial time series trends of covid-19 incidence and human mobility: an analysis of bidirectionality and spatiotemporal heterogeneity. *International journal of health geographics*, 22(1):33, 2023.

[60] Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *2010 2nd international workshop on cognitive information processing*, pages 411–416, 2010.

[61] Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector-tracking epidemics on twitter. In *Proc. of ECML*, pages 599–602, 2010.

[62] Vasileios Lampos, Daniel Preoţiuc-Pietro, and Trevor Cohn. A user-centric model of voting intention from social media. In *Proceedings of ACL*, pages 993–1003, 2013.

[63] Vasileios Lampos, Andrew C Miller, Steve Crossan, and Christian Stefansen. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific reports*, 5(1):1–10, 2015.

[64] Vasileios Lampos, Maimuna S. Majumder, Elad Yom-Tov, Michael Edelstein, Simon Moura, Yohhei Hamada, Molebogeng X. Rangaka, Rachel A. McKendry, and Ingemar J. Cox. Tracking COVID-19 using online search. *npj Digital Medicine*, 4(1):17, 2021. doi: 10.1038/s41746-021-00384-w.

[65] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[66] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *science*, 343(6176):1203–1205, 2014.

[67] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[68] Elad Levi, Eli Brosh, and Matan Friedmann. Intent-based prompt calibration: Enhancing prompt optimization with synthetic boundary cases. *arXiv preprint arXiv:2402.03099*, 2024.

[69] Victor O K Li, Jacqueline C K Lam, Yun Chen, and Jiatao Gu. Deep Learning Model to Estimate Air Pollution Using M-BP to Fill in Missing Proxy Urban Data. *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–6, 2017. doi: 10.1109/glocom.2017.8255004.

[70] Donghai Liang, Rachel Golan, Jennifer L Moutinho, Howard H Chang, Roby Greenwald, Stefanie Ebelt Sarnat, Armistead G. Russell, and Jeremy A Sarnat. Errors associated with the use of roadside monitoring in the estimation of acute traffic pollutant-related health effects. *Environmental Research*, 165:210–219, 2018.

[71] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning Representations from Imperfect Time Series Data via Tensor Rank Regularization. *arXiv*, 2019.

[72] Chen Lin, Joyce C Ho, and Eugene Agichtein. Cross-modal memory fusion network for multimodal sequential learning with missing values. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 312–319. Springer, 2021.

[73] Chen Lin, Joyce C Ho, and Eugene Agichtein. Cross-modal memory fusion network for multimodal sequential learning with missing values. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 312–319. Springer, 2021.

[74] Chen Lin, Safoora Yousefi, Elvis Kahoro, Payam Karisani, Donghai Liang, Jeremy Sarnat, Eugene Agichtein, et al. Detecting elevated air pollution lev-

els by monitoring web search queries: Algorithm development and validation. *JMIR Formative Research*, 6(12):e23422, 2022.

[75] Chen Lin, Jianghong Zhou, Jing Zhang, Carl Yang, and Eugene Agichtein. Graph neural network modeling of web search activity for real-time pandemic forecasting. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 128–137. IEEE, 2023.

[76] Yijun Lin, Nikhit Mago, Yu Gao, Yaguang Li, Yao-Yi Chiang, Cyrus Shahabi, and José Luis Ambite. Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning. In *Proc. of ACM SIGSPATIAL*, pages 359–368, 2018.

[77] Jingjing Liu, Panupong Pasupat, Yining Wang, Scott Cyphers, and Jim Glass. Query understanding enhanced by hierarchical parsing structures. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 72–77. IEEE, 2013.

[78] Xinyu Liu. Query sub-intent mining by incorporating search results with query logs for information retrieval. In *2023 IEEE 8th International Conference on Big Data Analytics (ICBDA)*, pages 180–186. IEEE, 2023.

[79] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[80] Google LLC. Google covid-19 search trends symptoms dataset, 2022. URL `http://goo.gle/covid19symptomdataset`.

[81] Google LLC. Explore covid-19 symptoms search trends, 2022. URL `https://pair-code.github.io/covid19_symptom_dataset`.

[82] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1101–1104, 2019.

[83] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proc. of NeurIPS*, NIPS'13, 2013.

[84] Gabriel J Milinovich, Gail M Williams, Archie CA Clements, and Wenbiao Hu. Internet-based surveillance systems for monitoring emerging infectious diseases. *The Lancet infectious diseases*, 14(2):160–168, 2014.

[85] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: harvesting opinions from the web. pages 169–176, 2011. doi: 10.1145/2070481.2070509.

[86] Fulufhelo V Nelwamondo, Shakir Mohamed, and Tshilidzi Marwala. Missing data: A comparison of neural network and expectation maximization techniques. *Current Science*, pages 1514–1521, 2007.

[87] National Library of Medicine. Pubmed, 2021. URL `https://pubmed.ncbi.nlm.nih.gov/`.

[88] OpenAI. Chatgpt (mar 14 version) [large language model], 2023. URL `https://chat.openai.com/chat`. Accessed: October 20, 2024.

[89] Diego Ortiz, José G Moreno, Gilles Hubert, Karen Pinel-Sauvagnat, and Lynda Tamine. Exploring the value of multi-view learning for session-aware query representation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022)*, pages 304–315. ACL: Association for Computational Linguistics, 2022.

[90] George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. Transfer Graph Neural Networks for Pandemic Forecasting. *arXiv*, 2020.

[91] Bharat A. Panuganti, Aria Jafari, Bridget MacDonald, and Adam S. DeConde. Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter. *Otolaryngology–Head and Neck Surgery*, 163(3):491–497, 2020. ISSN 0194-5998. doi: 10.1177/0194599820932128.

[92] Michael J Paul and Mark Dredze. *Social monitoring for public health*. Morgan & Claypool Publishers, 2017.

[93] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[94] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448, 2016. doi: 10.1109/icdm.2016.0055.

[95] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*, 2019.

[96] Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. Tripclick: the log files of a large health web search engine. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2507–2513, 2021.

[97] IBM Research. What is retrieval-augmented generation?, 2023. URL `https://research.ibm.com/blog/retrieval-augmented-generation-RAG`. Accessed: October 20, 2024.

[98] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. Trec-covid: rationale and structure of an information retrieval shared task for covid-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436, 2020.

[99] Yves Rybarczyk and Rasa Zalakeviciute. Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences*, 8(12):2570, 2018.

[100] Mauricio Santillana, André T Nguyen, Mark Dredze, Michael J Paul, Elaine O Nsoesie, and John S Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513, 2015.

[101] Jeremy A Sarnat, Stefanie Ebelt Sarnat, W Dana Flanders, Howard H Chang, James Mulholland, Lisa Baxter, Vlad Isakov, and Halûk Özkaynak. Spatiotemporally resolved air exchange rate as a modifier of acute air pollution-related morbidity in atlanta. *Journal of Exposure Science and Environmental Epidemiology*, 23(6):606, 2013.

[102] Stefanie Ebelt Sarnat, Jeremy A Sarnat, James Mulholland, Vlad Isakov, Halûk Özkaynak, Howard H Chang, Mitchel Klein, and Paige E Tolbert. Application of alternative spatiotemporal metrics of ambient air pollution exposure in a time-series epidemiological study in atlanta. *Journal of Exposure Science and Environmental Epidemiology*, 23(6):593, 2013.

[103] Ari Seifter, Alison Schwarzwalder, Kate Geis, and John Aucott. The utility of "google trends" for epidemiological research: Lyme disease as an example. *Geospatial health*, pages 135–137, 2010.

[104] Frans Snik, Jeroen HH Rietjens, Arnoud Apituley, Hester Volten, Bas Mijling, Antonio Di Noia, Stephanie Heikamp, Ritse C Heinsbroek, Otto P Hasekamp, J Martijn Smit, et al. Mapping atmospheric aerosols with a citizen science network of smartphone spectropolarimeters. *Geophysical Research Letters*, 41 (20):7351–7358, 2014.

[105] Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli, and Mike Bendersky. Quill: Query intent with large language models using retrieval augmentation and multi-stage distillation. *arXiv preprint arXiv:2210.15718*, 2022.

[106] Xianfeng Tang, Huaxiu Yao, Yiwei Sun, Charu Aggarwal, Prasenjit Mitra, and Suhang Wang. Joint Modeling of Local and Global Temporal Dynamics for Multivariate Time Series Forecasting with Missing Values. 2019.

[107] Yue Teng, Dehua Bi, Guigang Xie, Yuan Jin, Yong Huang, Baihan Lin, Xiaoping An, Dan Feng, and Yigang Tong. Dynamic forecasting of zika epidemics using google trends. *PloS one*, 12(1):e0165085, 2017.

[108] Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, George Karypis, Thien Q Tran, and Jun Sakuma. Seasonal-adjustment Based Feature Selection Method for Predicting Epidemic with Large-scale Search Engine Logs. *arXiv*, pages 2857–2866, 2019. doi: 10.1145/3292500.3330766.

[109] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro,

Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[110] Sverre Vedal, Michael Brauer, Richard White, and John Petkau. Air pollution and daily mortality in a city with low levels of pollution. *Environmental health perspectives*, 111(1):45–52, 2003.

[111] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. Select-Additive Learning: Improving Generalization in Multimodal Sentiment Analysis. *arXiv*, 2016.

[112] Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. Combining knowledge with deep convolutional neural networks for short text classification. In *IJCAI*, volume 350, pages 3172077–3172295, 2017.

[113] Lijing Wang, Aniruddha Adiga, Jiangzhuo Chen, Adam Sadilek, Srinivasan Venkatramanan, and Madhav Marathe. Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12191–12199, 2022.

[114] Shuting Wang, Xin Xu, Mang Wang, Weipeng Chen, Yutao Zhu, and Zhicheng Dou. Richrag: Crafting rich responses for multi-faceted queries in retrieval-augmented generation. *arXiv preprint arXiv:2406.12566*, 2024.

[115] Yaqing Wang, Song Wang, Yanyan Li, and Dejing Dou. Recognizing medical search query intent by few-shot learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 502–512, 2022.

[116] Zhongyuan Wang, Kejun Zhao, Haixun Wang, Xiaofeng Meng, and Ji-Rong Wen. Query understanding through knowledge-based conceptualization. In *IJCAI*, 2015.

[117] Ryen W White, Paul N Bennett, and Susan T Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1009–1018, 2010.

[118] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[119] Feng Xie, Zhong Zhang, Liang Li, Bin Zhou, and Yusong Tan. EpiGNN: Exploring Spatial Transmission with Graph Neural Network for Regional Epidemic Forecasting. *arXiv*, 2022. doi: 10.48550/arxiv.2208.11517.

[120] Shihao Yang, Mauricio Santillana, and Samuel C Kou. Accurate estimation of influenza epidemics using google search data via argo. *PNAS*, 112(47):14473–14478, 2015.

[121] Xi Yang, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

[122] Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. St-mvl: filling missing values in geo-sensory time series data. 2016.

[123] Elad Yom-Tov, Vasileios Lampos, Thomas Inns, Ingemar J. Cox, and Michael Edelstein. Providing early indication of regional anomalies in COVID-19 case counts in England using search engine queries. *Scientific Reports*, 12(1):2373, 2022. doi: 10.1038/s41598-022-06340-2.

[124] Elad Yom-Tov, Vasileios Lampos, Thomas Inns, Ingemar J. Cox, and Michael Edelstein. Providing early indication of regional anomalies in COVID-19 case counts in England using search engine queries. *Scientific Reports*, 12(1):2373, 2022. doi: 10.1038/s41598-022-06340-2.

[125] Chunyuan Yuan, Yiming Qiu, Mingming Li, Haiqing Hu, Songlin Wang, and Sulong Xu. A multi-granularity matching attention network for query intent classification in e-commerce retrieval. In *Companion Proceedings of the ACM Web Conference 2023*, pages 416–420, 2023.

[126] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv*, 2016.

[127] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. *arXiv*, 2017.

[128] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory Fusion Network for Multi-view Sequential Learning. *arXiv*, 2018.

[129] Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Factorized Multimodal Transformer for Multimodal Sequential Learning. *arXiv*, 2019.

[130] Scott L Zeger, Duncan Thomas, Francesca Dominici, Jonathan M Samet, Joel Schwartz, Douglas Dockery, and Aaron Cohen. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental health perspectives*, 108(5):419–426, 2000.

[131] Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14365–14373, 2021.

[132] Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N Bennett, Nick Craswell, and Saurabh Tiwary. Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74, 2019.

[133] Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, and Alexander Baklanov. Real-time air quality forecasting, part i: History, techniques, and current status. *Atmospheric Environment*, 60:632–655, 2012.

[134] Yuzhou Zhang, Laith Yakob, Michael B Bonsall, and Wenbiao Hu. Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data. *Scientific reports*, 9(1):3262, 2019.

[135] Xiaosong Zhao, Rui Zhang, Jheng-Long Wu, and Pei-Chann Chang. A deep recurrent neural network for air quality classification. *J. Inf. Hiding Multimed. Sig. Proc.*, 9:346—354, 2018.

[136] Xiaosong Zhao, Rui Zhang, Jheng-Long Wu, and Pei-Chann Chang. A deep recurrent neural network for air quality classification. *J. Inf. Hiding Multimed. Sig. Proc.*, 9:346–354, 2018.

[137] Jianghong Zhou. *Improving Interactive Search with User Feedback*. PhD thesis, Emory University, 2022.

[138] Jianghong Zhou and Eugene Agichtein. Rlirank: Learning to rank with reinforcement learning for dynamic search. In *Proceedings of The Web Conference 2020*, pages 2842–2848, 2020.

[139] Jianghong Zhou, Jiangqun Ni, and Yuan Rao. Block-based convolutional neural network for image forgery detection. In *Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, 2017, Proceedings 16*, pages 65–76. Springer, 2017.

[140] Jianghong Zhou, Eugene Agichtein, and Surya Kallumadi. Diversifying multi-aspect search results using simpson's diversity index. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2345–2348, 2020.

[141] Jianghong Zhou, Sayyed M Zahiri, Simon Hughes, Khalifeh Al Jadda, Surya Kallumadi, and Eugene Agichtein. De-biased modeling of search click behavior with reinforcement learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1637–1641, 2021.

[142] Jianghong Zhou, Sayyed Zahiri, Simon Hughes, Surya Kallumadi, Khalifeh Al Jadda, and Eugene Agichtein. User click modelling in search queries, May 5 2022. US Patent App. 17/514,522.

[143] Wenjin Zhu, Jianzhou Wang, Wenyu Zhang, and Donghuai Sun. Short-term effects of air pollution on lower respiratory diseases and forecasting by the group method of data handling. *Atmospheric environment*, 51:29–38, 2012.

[144] Bin Zou, Vasileios Lampos, and Ingemar Cox. Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data. pages 2505—2516, 2019.

[145] Bin Zou, Vasileios Lampos, and Ingemar Cox. Transfer learning for unsupervised influenza-like illness models from online search data. In *Proc. of WWW*, pages 2505–2516, 2019.

[146] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.