Forecasting Emerging Pandemics with Transfer Learning and Location-aware News Analysis

Jing Chen
Stevens Institute of Technology
Hoboken, NJ, USA
jchen6@stevens.edu

Germán G. Creamer
Stevens Institute of Technology
Hoboken, NJ, USA
gcreamer@stevens.edu

Yue Ning
Stevens Institute of Technology
Hoboken, NJ, USA
yue.ning@stevens.edu

***Abstract*—Monitoring and forecasting epidemic diseases are of prime importance to public health organizations and policymakers in taking proper measures and adjusting prevention tactics. Early prediction is especially important to restrict the spread of emerging pandemics such as COVID-19. However, despite increasing research and development for various epidemics, several challenges remain unresolved. On the one hand, early-stage epidemic prediction for emerging new diseases is difficult because of data paucity and lack of experience. On the other hand, many existing studies ignore or fail to leverage the contribution of social factors such as news, geolocations, and climate. Even though some researchers have recognized the profound impact of social features, capturing the dynamic correlation between these features and pandemics requires an extensive understanding of heterogeneous formats of data and mechanisms. In this paper, we design TLSS, a neural transfer learning architecture for learning and transferring general characteristics of existing epidemic diseases to predict a new pandemic. We propose a new feature module to learn the impact of news sentiment and semantic information on epidemic transmission. We then combine this information with historical time-series features to forecast future infection cases in a dynamic propagation process. We compare the proposed model with several state-of-the-art statistics approaches and deep learning methods in epidemic prediction with different lead times of ground truth. We conducted extensive experiments on three stages of COVID-19 development in the United States. Our experiment demonstrates that our approach has strong predictive performance for COVID infection cases, especially with longer lead times.**

***Index Terms*—COVID-19, Epidemic Forecasting, Transfer Learning, Sentiment and Semantic Analysis**

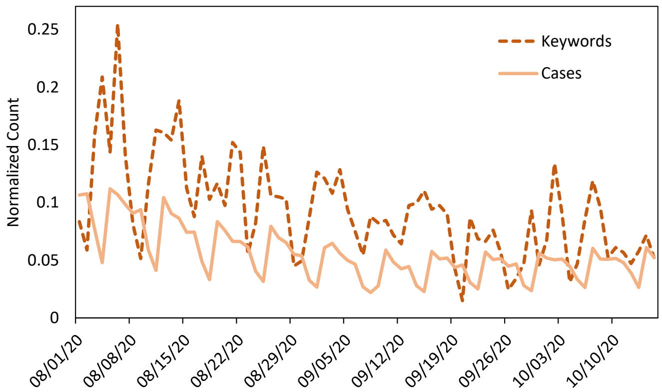
I. Introduction

In 2019, the Coronavirus disease (COVID-19) outbreak in Wuhan, China rapidly spread worldwide. It was a novel disease and evolved into a pandemic in a very short amount of time. In January 2020, it was declared as a public health emergency of international concern by WHO. Currently, according to WHO, more than 200 million cases have been confirmed, and more than 4 million deaths have been recorded [1]. Early detection of a new type of disease, such as COVID-19, using various indicators related to its progression provides valuable opportunities for local governments and public authorities to plan timely intervention, allocate medical resources, and adjust control strategies.

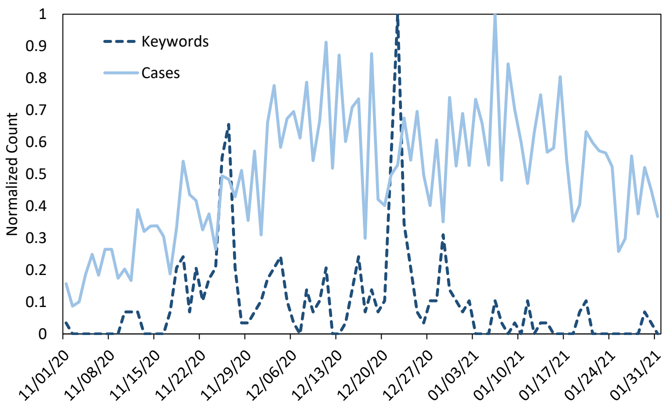
Given system latencies in data collection and monitoring, we primarily address the problem of COVID-19 forecasting with a leadtime from 1 to 14 days based on surveillance data of 50 states in the United States. It is intuitive to adapt tools that have been developed for existing contagious diseases to new emerging disease [34, 35]. Critical information and clues about disease emergence and persistence mechanisms may show similar patterns. In this work, we use Heterogeneous Transfer Learning (HTL) methods to learn general characteristics of existing epidemic diseases, such as influenza-like illnesses (ILI), and transfer the learned knowledge to predict an emerging epidemic disease, such as COVID-19. Existing studies have already proved the impact of social factors on public health such as the mediating effect of human awareness and behavioral changes [4, 6]. In this paper, we explore the impact of social factors on pandemic transmission from two aspects: sentiment and semantic information of news. We aim to explore two motivating questions: (1) will including public opinions and emotions (e.g., panic or optimism) improve epidemic forecasting performance? and (2) will the correlation of public opinions from different locations affect epidemic forecasting performance? Unstructured data such as text are under-utilized for prediction tasks despite that they include potentially valuable indicators. We use different sentiment and semantic analysis methods to extract relevant features. Considering the dynamic variation of news over time in multiple locations, we design a dynamic location-aware sentiment and semantic attention mechanism to learn the correlation between COVID-19 related news and the outbreak of COVID-19 cases.

Existing research on epidemic prediction has achieved great success in various aspects. Mathematical methods, such as stochastic processes, Markov chains, and compartmental models have outstanding performance in theoretical analysis of macroscopic regularities of epidemic diffusion like the threshold of an epidemic becoming urgent and the size/population of epidemic infections [14]. The assumption of homogeneous populations and small sets of variables is inadequate to capture the variety of factors associated with epidemic spreading processes [3]. Other statistical models such as Autoregressive (AR) and Autoregressive Moving Average (ARMA) are convenient and easy to use. Precise results can be obtained in short-term time-series analysis [31]. However, they do not perform well in relatively long-term COVID-19 trend forecasting because of the evolution of diseases and human environments over time. Some complex network methods, such as system dynamic models, numerical simulations of

978-1-6654-8045-1/22/$31.00 © 2022 IEEE



(a) California: Normalized COVID-19 daily case count and normalized occurrence of keywords (e.g., mask) in news articles.



(b) Delaware: Normalized COVID-19 daily case count and normalized occurrence of keywords (e.g., holidays) in news articles.

Fig. 1: Example demonstrating the potential correlation between news articles and COVID-19 spread.

epidemics, and recurrent neural networks (RNN) are emerging approaches to learn different spreading patterns, yet it is hard to incorporate social factors in these methods for epidemic predictions [9].

Several significant challenges remain in epidemic forecasting especially with limited data. First, most mathematical and statistical methods cannot capture cultural/societal temporal dependencies directly and efficiently, using limited input data from new diseases. For instance, holidays and festivals caused a large-scale COVID-19 outbreak from November 2020 to January 2021. Second, it is hard to extract features from text data and capture their internal association with numerical data over time. Specifically, traditional word embedding methods cannot accurately learn public opinions and emotions in news articles relevant to each location's epidemics. Third, sentiment and semantic information of news are updated over time. Dynamic news information has not been exhaustively explored with limited feature extraction techniques. Figure 1 is a motivating example to explain the motivation of using news articles as an ancillary feature in epidemic prediction. In figure 1a, we observe that COVID-19 infection cases decreased as well as COVID-19 related keywords (e.g., mask) in news articles from 08/01/2020 to 10/15/2020 in California. It demonstrates the potential correlation between public policies and COVID-19 outbreaks. The flu outbreak commonly occurs in Fall and Winter, but COVID-19 cases show inconsistent outbreak patterns. Figure 1b shows two peaks of holiday-related COVID-19 news on 11/26/2020 and 12/22/2020 respectively in Delaware. Correspondingly, there are two massive COVID-19 outbreaks on 12/03/2020 and 12/30/2020. It demonstrates that potential influence of social factors (e.g., public sentiments reflected in news articles) may be helpful in interpreting non-traditional patterns of epidemic outbreaks.

In this paper, we focus on predicting daily infection cases of an emerging epidemic (i.e., COVID-19) using limited historical time-series data and epidemic-related news articles. We design a neural transfer learning architecture for learning and transferring common characteristics of an existing infectious disease (i.e., ILI) to forecast an emerging disease (i.e., COVID-19) that has sparse data. Meanwhile, we propose a module to learn the impact of news sentiment and semantic information on epidemic transmission to interpret non-traditional outbreak patterns. We then combine learned semantic/sentiment representations with sequential dependencies that are captured from recurrent neural networks in local time-series data to forecast future infection cases in a dynamic propagation process. Our contributions are summarized below:

- We propose a novel heterogeneous transfer learning architecture (HTL) to forecast future infection cases of the emerging pandemic COVID-19 based on historical data.
- We extract sentiment and semantic features from COVID-19 related news articles using Valence Aware Dictionary for Sentiment Reasoning (VADER) and Sentence Transformers (SBERT). We explore the impact of social factors on epidemic prediction and explain the unconventional variations of COVID-19 outbreaks.
- We design a module to learn the association of public opinions among different locations. We adopt a dynamic location-aware attention mechanism to capture the ever-changing correlation between COVID-19 related news and infection rates.
- We conduct experiments on three stages of real-world COVID-19 data with different time settings. We compare the proposed approach with a broad range of state-of-the-art models to demonstrate the effectiveness of the proposed model in COVID-19 forecasting.

II. Related Work

A. Epidemic Forecasting

In epidemic forecasting, many different approaches are used for understanding spreading patterns and evaluating disease control policies. Time-series regression is one of the primary directions of the problem formulation to model and simulate epidemic diseases. There are several data-driven and statistical approaches that have been developed for time series forecasting [8], such as Auto-regression (AR), Autoregressive Moving Average (ARMA), and Seasonal Auto-Regressive

Integrated Moving Average with eXogenous factors (SARIMAX) [10, 28, 29]. Sahai et al. [42] initially applied the ARIMA model to predict the COVID-19 outbreak in the top five most affected countries and proposed thought-provoking suggestions for further improvement of epidemic forecasting. Jain et al. [21] used exponential smoothing, ARIMA, and SARIMA models to predict COVID-19 diffusion in terms of the impact of festive seasons. Other complex deep learning methods have garnered increasing interest in epidemic forecasting and demonstrated the great learning ability to capture temporal dependencies [2, 44]. Xiao et al. [45]. proposed a novel data-driven framework C-Watcher, which monitors every neighborhood in a target city and predicts infection risks before a COVID-19 outbreak in that city.

There are two general types of validation methods in time-series analysis: direct forecasting and iterative forecasting. Direct forecasting methods use historical data as one input to predict a future value. Iterative methods recursively use historical data as several inputs to update a model for future predictions. One example is walk-forward validation [38] and its principle is to walk through several training sets and test sets over time to optimize a model and adapt the model to real-world data updates.

B. Heterogeneous Transfer Learning

HTL aims to leverage knowledge extracted from a source domain to a different but related target domain, which may have different data distributions and label spaces [11]. Moon and Carbonell [33] proposed Attentional Heterogeneous Transfer which leverages both unlabeled source and target data to enhance the discriminative power of feature mapping. Its simulation studies have been implemented in many real-world transfer learning tasks. Rodríguez et al. [41] designed a COVID augmented ILI deep network (CALI-NET) which is a transfer learning framework to forecast flu cases where flu and COVID co-exist. However, CALI-NET cannot be directly applied to forecast new epidemic diseases due to the restriction of input data and the lack of training flexibility. Horry et al. [19] demonstrated that neural transfer learning architectures can improve the COVID-19 detection ability from standard medical images such as X-Ray, Ultrasound, and CT scans. The optimization of intelligent image classification models can reduce the burden on medical professionals. Pal and Kar [37] applied Fuzzy Transfer Learning in time series to predict stock market prices and demonstrated that knowledge transfer improves prediction accuracy with smoothing dependent attributes over time.

C. Social Media Impact

With the increasing growth of Nature Language Processing (NLP) techniques [36, 43], word embedding methods such as word2vec [17], Global Vectors for Word Representation (GloVe) [39], and Bidirectional Encoder Representations from Transformers (BERT) [22] have improved the performance of many NLP tasks. Many sentiment/topic analysis approaches such as Latent Dirichlet allocation (LDA) [7], BERT based sentiment classifier (BERTsent) [26], VADER [20] are also commonly used to extract attitude, emotions, and opinions from text. Textual data such as tweets, reports, and news become complementary features in many time-series analyses. Mahdikhani [30] proposed a new framework to detect public emotions during the pandemic and after vaccination using text embedding methods. They then analyzed the impact of public emotions on the retweetability of posted tweets during the COVID-19 pandemic. Kim and Ahn [23] used weekly counts of influenza-related keywords in news articles in a support vector machine (SVM) to predict if the number of future patients increases or decreases. They then combined with the data of cases to improve the accuracy of future influenza patients' prediction. In summary, many researchers have reported a correlation between social factors and pandemic patterns [13, 25]. These previous studies are relevant precedents for the design of a pandemic prediction model based on social factors.

III. Methodology

A. Problem Formulation

As shown in Figure 2, we formulate the problem of emerging epidemic forecasting as a heterogeneous transfer learning model with three parts: a transfer learning module, a simulation module, and a prediction module. The purpose of the transfer learning module is to learn the general characteristics of existing infectious diseases (e.g., flu) from a source model Cola-GNN [12] and transfer the knowledge to our proposed model for fine-tuning on new epidemic disease data (e.g., COVID-19). The simulation module is for capturing the temporal dependencies of COVID-19 cases and encoding sentiment and semantic features in news articles over time. In the prediction module, we combine the transfer learning knowledge, temporal dependencies, and sentiment/semantic features to predict future infection cases.

Given current time k, the objective is to forecast new infection cases at a future time point $k + h$ using historical data from a past time window $[k - T : k]$ where h represents the leadtime of the prediction and T is the historical window size. In the following, we ignore k in the notations for clarity. First, we utilize historical daily COVID-19 time-series data $\boldsymbol{X} \in \mathbb{R}^{N\times T}$ where N is the number of locations. We pre-train the news articles data to extract sentiment and semantic features. We have pre-trained news sentiment data $\boldsymbol{V} \in \mathbb{R}^{N\times T}$ where $\boldsymbol{V}_{i,t}$ is a scalar value to represent the average emotion score of day t's news for location i. We have pre-trained semantic data $\boldsymbol{S} \in \mathbb{R}^{N\times T\times 50}$ where $\boldsymbol{S}_{i,t} \in \mathbb{R}^{50\times 1}$ represents an average semantic vector of day t's news for location i. Details of these vectors are described in Section B and Section C. Table I summarizes important notations.

B. Heterogeneous Transfer Learning

We learn the general characteristics of existing pandemics (e.g., flu) from state-of-the-art models and transfer the learned knowledge to predict new emerging diseases (e.g., COVID-19). Given that our problem is a spatio-temporal prediction

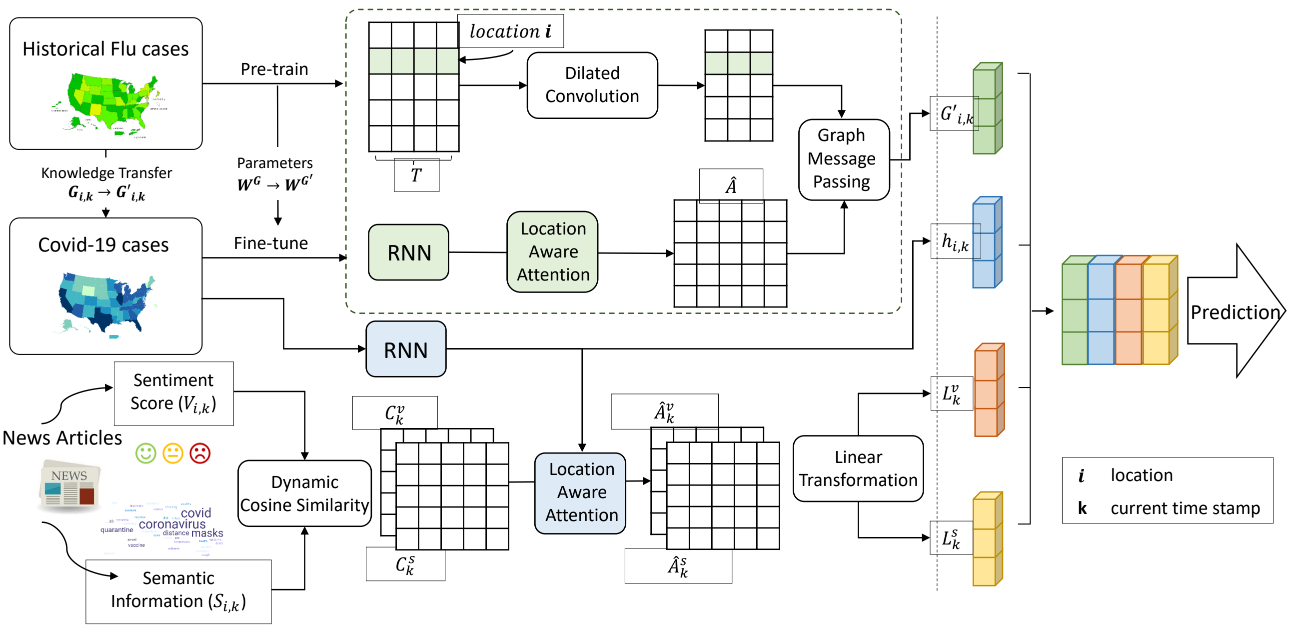


Fig. 2: Overview of the proposed framework TLSS. COVID-19 time-series data are imported to the top transfer learning module to recognize the spread patterns of existing infectious disease. COVID-19 news sentiment and semantic data are imported to the bottom module to learn how public opinion and emotions impact the COVID-19 outbreak.

TABLE I: Important notations and descriptions

Notation	Description
T	window size of one training input
k	time index
N	number of locations
h	horizon/leadtime of a prediction
D, F	feature dimensions
X$\in \mathbb{R}^{N\times T}$	infection cases for N locations of window size T
V$\in \mathbb{R}^{N\times T}$	sentiment scores for N locations of window size T
S$\in \mathbb{R}^{N\times T\times 50}$	semantic embeddings for N locations of window size T
$C_k^v \in \mathbb{R}^{N\times N}$	dynamic sentiment cosine similarity of N locations
$C_k^s \in \mathbb{R}^{N\times N}$	dynamic semantic cosine similarity of N locations
$G_{i,k} \in \mathbb{R}^{N\times F}$	learned representations from the source model
$G'_{i,k} \in \mathbb{R}^{N\times F}$	learned representations from the target model

task, we choose cross-location attention-based graph neural networks (Cola-GNN) [12] as the base model. Cola-GNN is designed to learn time series embeddings and geolocation correlations for predicting long-term influenza-like illness.

In our framework, we set up the modified Cola-GNN model as the source model and design a new module that incorporates social factors in the target model. The proposed model aims to efficiently and flexibly learn the characteristics of a new epidemic virus (i.e., COVID-19) based on relevant existing infectious diseases (i.e., ILI). We first initialize part of our target model parameters based on the pre-trained Cola-GNN model using ILI data. We then fine-tune the parameters in the target model using COVID-19 data. As depicted in Fig. 2, parts of the weights from the source model are shared with the target model: $W^G \rightarrow W^{G'}$. We address the COVID-19 data insufficiency problem by using the general epidemic characteristics, which are learned from rich ILI historical data. This transformation is defined as $G_{i,k} \rightarrow G'_{i,k}$ where i is the index for a location and k is the index for a time stamp. The hidden states of the modified Cola-GNN model are captured in the fine-tuning process on COVID-19 dataset. In this way, we project hidden features extracted from the heterogeneous feature space into a shared latent space. Then we concatenate these features with location-aware dynamic sentiment and semantic features. Finally, we reconstruct all the features as the input for final predictions.

C. Location-Aware Dynamic Sentiment Analysis

In this study, we dynamically model public concerns and sentiments in different locations during the outbreak and spreading of infectious disease from news data. We collect COVID-19 related news articles from the Global Database of Events, Language, and Tone (GDELT) [27] by keyword filtering. Since GDELT data are unlabeled, we use an unsupervised sentiment analysis method, VADER [20], to pre-train the data and generate a sentiment score for each news article. Sentiment scores measure emotions, sentiments, and attitudes of word vectors [5]. VADER provides a sentiment score for each word and then we sum up the scores of words in an article to represent the sentiment of this article. Given a location, we apply average pooling to the sentiment scores of news articles in a day. Each location will have a scalar value to represent the average polarity of news at the current time stamp.

To capture geographical correlations of public emotions, we calculate the cosine similarity of news sentiment scores in a past time window between every two locations. Once obtained the sentiment scores $\boldsymbol{V}$ for each location, we select data in the past time window of size T given current time stamp k to calculate a dynamic sentiment cosine similarity between two locations. We will update $\boldsymbol{C}_k^v$ for each time stamp k.

$$\boldsymbol{C}_k^v[i,j] = \frac{\boldsymbol{V}_{i,[k-T:k]} \cdot \boldsymbol{V}_{j,[k-T:k]}}{||\boldsymbol{V}_{i,[k-T:k]}|| \cdot ||\boldsymbol{V}_{j,[k-T:k]}||} \tag{1}$$

where $\boldsymbol{V}_{i,[k-T:k]} \in \mathbb{R}^T$ and $\boldsymbol{V}_{j,[k-T:k]} \in \mathbb{R}^T$ represent the sentiment scores of location i and location j, respectively for the past time window $[k-T:k]$.

Inspired by Cola-GNN, we adopt a location-aware attention mechanism to learn temporal dependencies as well as sentiment dependencies of locations from historical data. To measure the impact of location i on location j from the hidden states learned from RNN, we define $\boldsymbol{a_{i,j}}$ in attention coefficient matrix $\boldsymbol{A}$ as:

$$\boldsymbol{a_{i,j}} = \boldsymbol{v}^T g(\boldsymbol{W}^s \boldsymbol{h}_i + \boldsymbol{W}_t \boldsymbol{h}_j + \boldsymbol{b}^s) + b^v, \tag{2}$$

where $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ are the last hidden state $\boldsymbol{h}_k$ of an RNN model for location i and location j, $\boldsymbol{W}^s$, $\boldsymbol{W}^t \in \mathbb{R}^{d_a \times D}$, $\boldsymbol{v} \in \mathbb{R}^{d_a}$, $b^s \in \mathbb{R}^{d_a}$, and $b^v \in \mathbb{R}$ are trainable parameters. d_a is a hyperparameter to control the dimensions of other parameters. To combine sentiment cosine similarity matrix $\boldsymbol{C}_k^v$ and attention coefficient matrix $\boldsymbol{A}$, we define an element gate learned from the general attention matrix that evolves over time. We then propose the location-aware sentiment attention mechanism as:

$$\hat{\boldsymbol{A}}_k^v = \sigma(\boldsymbol{W}_m \boldsymbol{A} + b^m \mathbf{1}_N \mathbf{1}_N^\top) \odot \boldsymbol{C}_k^v + (\mathbf{1}_N \mathbf{1}_N^\top) \odot \boldsymbol{A}, \tag{3}$$

where $\boldsymbol{W}_m \in \mathbb{R}^{N \times N}$ and $b_m \in \mathbb{R}$ are trainable parameters. We apply a linear transformation to location-aware sentiment attention matrix $\hat{\boldsymbol{A}}_k^v$ as:

$$\boldsymbol{L}_k^v = \boldsymbol{W}^v \hat{\boldsymbol{A}}_k^v + \boldsymbol{b}^v, \tag{4}$$

where $\boldsymbol{L}_k^v \in \mathbb{R}^{N \times D}$ is a dynamic matrix that changes over different time stamps, and $\boldsymbol{W}^v$ and $\boldsymbol{b}^v$ are the trainable parameters. We then project $\boldsymbol{L}_k^v$ to the shared latent space and concatenate it with other embeddings for predicting infection cases at time point $k+h$.

D. Location-Aware Dynamic Semantic Analysis

Besides the sentiment features of news, the semantic information of news articles plays an essential role in estimating epidemic outbreaks. For example, the publicity of disease prevention policies will help control the diffusion of epidemics. To accurately extract semantic information from news, we use Sentence-BERT (SBERT) [40] to construct an embedding vector of each news article. SBERT is a modification of the BERT network using siamese and triplet networks to derive semantically meaningful sentence embeddings. We set the maximum input length as 500 words in the pre-trained SBERT base model to generate output embeddings with a default size of 786. We apply Principle Component Analysis (PCA) to the existing model and reduce the output embedding size from 768 to 50 dimensions. Using SBERT and PCA, we create a vector $s \in \mathbb{R}^{50 \times 1}$ for each news article to represent its semantic information.

To capture geographical correlations of public opinions, we calculate the cosine similarity of the news semantic matrices for each pair of locations, since the sentence embeddings trained by SBERT are semantically meaningful and can be compared using cosine similarity. According to the tensor of semantic embeddings $\boldsymbol{S}$ for each location at current time stamp k, we select data with window size T to calculate a dynamic semantic cosine similarity between two locations:

$$\boldsymbol{C}_k^s[i,j] = \frac{\text{concat}(\boldsymbol{S}_{i,[k-T:k]}) \cdot \text{concat}(\boldsymbol{S}_{j,[k-T:k]})}{||\text{concat}(\boldsymbol{S}_{i,[k-T:k]})|| \cdot ||\text{concat}(\boldsymbol{S}_{j,[k-T:k]})||}, \tag{5}$$

where $\boldsymbol{S}_{i,[k-T:k]} \in \mathbb{R}^{T \times 50}$ represents the semantic information of location i for a time-span from $k-T$ to k. We apply a concatenation function *concat* to reshape the matrix into a vector of dimension $\mathbb{R}^{T*50}$.

We also use the location-aware attention mechanism, as introduced in Section III.C, to calculate the dynamic location-aware semantic matrix $\hat{\boldsymbol{A}}_k^s \in \mathbb{R}^{N \times N}$ with a time-span of size T. We apply linear transformation to this matrix to calculate $\boldsymbol{L}_k^s$ with trainable parameters $\boldsymbol{W}^s$ and $\boldsymbol{b}^s$. We project $\boldsymbol{L}_k^s$ to the shared latent space and concatenate it with other embeddings for forecasting infection cases at time point $k+h$.

$$\boldsymbol{L}_k^s = \boldsymbol{W}^s \hat{\boldsymbol{A}}_k^s + \boldsymbol{b}^s, \tag{6}$$

E. Prediction and Optimization

For a location i, given a time stamp k, we learn RNN hidden states ($\boldsymbol{h}_{i,k} \in \mathbb{R}^D$) from historical infection cases (or rates) of window size T, as well as the linear transformations of the sentiment ($\boldsymbol{L}_{i,k}^v \in \mathbb{R}^D$) and semantics ($\boldsymbol{L}_{i,k}^s \in \mathbb{R}^D$) features learned from news data. We combine the output of the sentiment analysis ($\boldsymbol{L}_{i,k}^v \in \mathbb{R}^D$), the output of the semantic analysis ($\boldsymbol{L}_{i,k}^s \in \mathbb{R}^D$), the hidden states ($\boldsymbol{h}_{i,k} \in \mathbb{R}^D$) from the RNN model, and the hidden states learned from Cola-GNN ($\boldsymbol{G}'_{i,k} \in \mathbb{R}^F$). We feed them to the output layer for prediction:

$$\hat{y}_{i,k} = \phi(\boldsymbol{\theta}^\intercal [\boldsymbol{h}_{i,k}; \boldsymbol{L}_{i,k}^v; \boldsymbol{L}_{i,k}^s; \boldsymbol{G}'_{i,k}] + b^\theta), \tag{7}$$

where ϕ is the activation function, and $\boldsymbol{\theta} \in \mathbb{R}^{D+D+D+F}$, b^θ are trainable parameters. D is the dimension of RNN hidden states and sentiment/semantic outputs, and F is the dimension of the transferred hidden states from the source model.

We compare the prediction value of each location with the corresponding ground truth and optimize a regularized L1-norm loss via gradient descent:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \sum_{k=1+T}^{n_i} |y_{i,k} - \hat{y}_{i,k}| + \lambda \mathcal{R}(\Theta), \tag{8}$$

where $y_{i,k}$ represents the ground truth, and $\hat{y}_{i,k}$ represents the prediction value. Θ represents all training parameters and $\mathcal{R}(\Theta)$ is the regularization term.

IV. Experimental Evaluation

A. Experiment Setup

1) Datasets:

- **COVID-19 Cases Data** This data is collected from CDC US-COVID-19-Cases [16]. It contains COVID-19 daily confirmed cases (new patient counts) in 50 states of the United States. The data includes three growing stages from 4/1/2020 to 9/30/2021.
- **COVID-19 Original News Data** This data is collected from GDELT. It includes COVID-19 related news articles in 50 states of the United States from 4/1/2020 to 9/30/2021.
- **Pre-trained News Sentiment Data** We pre-train the COVID-19 original news data using VADER to extract the relevant sentiment score of each news article.
- **Pre-trained News Semantic Data** We pre-train the COVID-19 original news data using SBERT to extract the relevant semantic features of each news article.

We use the walk-forward cross-validation method to split the COVID-19 data into three growing stages (e.g., data1 from 4/1/2020 to 9/30/2020, data2 from 10/1/2020 to 3/31/2021, and data3 from 4/1/2021 to 9/30/2021) based on the existing studies on the outbreak seasonality of epidemics [32]. Each stage covers at least one outbreak peak of COVID-19 respectively. We then split each stage of COVID-19 data into training, validation, and test set in chronological order at a ratio of 70%-10%-20%. We normalize all data between 0 and 1 range for measuring variables at different scales based on the training data. Validation data is used to avoid overfitting and to determine the number of epochs to run.

TABLE II: Data description. Size means number of locations multiplied by the number of dates and the dimension of daily data.

Dataset	Size	Max	Min
COVID-19 Cases	50*548*1	34425	0
COVID-19 Original News	863k	-	-
Pre-trained News Sentiment	50*548*1	1	0
Pre-trained News Semantic	50*548*50	1	-1

2) Evaluation Metrics:

- **The Root Mean Squared Error (RMSE)** measures the error between the predicted values from a model and ground truth observations after converting the normalized values into the real range:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2},$$

  where $\hat{y}_i$ is a predicted value and y_i is the corresponding ground truth value.
- **Diebold-Mariano (DM) Test** measures the accuracy of two forecast methods $(\hat{y}_1, \hat{y}_2)$, which compares the error between the predicted values from two models and ground truth: $\text{DM} = (\hat{y}_{1,i} - y_i)^2 - (\hat{y}_{2,i} - y_i)^2$, where $\hat{y}_{1,i}$ and $\hat{y}_{2,i}$ are predicted values from two models.
- **Leadtime:** is the period between the current end of the input time window and the target prediction time. For example, leadtime=7 means the input is $[x_1, x_2, \ldots, x_T]$, and the output is x_{T+7}.

3) Comparison methods: We compare our proposed framework with the following state-of-the-art approaches and their derivative models:

- **Autoregressive (AR)** AR is a basic statistical method in time-series prediction. The prediction values for future time stamps are modeled as a linear transformation of historical data in a past time window. In this work, we train individual AR models for different locations. We set the hyperparameter p to be the window size T.
- **Autoregressive Moving Average (ARMA)** ARMA is derived from the AR model and combines with a moving average (MA) term to better estimate the trend cycle of historical data in time series prediction. In this work, the size of moving average smoothing window is set as 2.
- **Recurrent Neural Network (RNN)** RNN is a state-of-the-art deep learning model for temporal sequence learning, widely used in epidemic forecasting. In this work, we apply multiple versions of RNN with different inputs: **RNN-multilocation** uses an input vector of COVID-19 infection cases of multiple locations in the United States. **RNN-Global** is an RNN model for COVID-19 case prediction regardless of geolocation. The parameters are shared across different regions. **RNN-Global-keywords** uses a pre-trained news dataset as complementary features, which contains the count of high-frequency vocabulary about COVID-19 in each news article. In **RNN-Global-doc2vec**, we initially pre-train the news dataset using doc2vec to learn the news embeddings and represent each news as a fixed-length vector. We combine the news embeddings with infection cases of each location and feed them to the RNN-Global model for predicting future COVID-19 cases. All locations share one RNN model.
- **Cross-Location Attention Based Graph Neural Network (Cola-GNN)** Cola-GNN is a deep learning method for learning time-series embeddings in long-term epidemic forecasting. We utilize it in COVID-19 cases prediction in multiple locations, considering the propagation at the population level (e.g., geolocation). We also use it as the source model of our proposed framework.
- **Transfer Learning Network with News Sentiment and Semantic Analysis (TLSS-Frozen) TLSS**-Frozen is derived from our proposed method **TLSS** when transferring all characteristics of influenza to represent COVID-19 in epidemic prediction. The target model's parameters are frozen and not updated based on the pre-trained source model Cola-GNN using influenza data.

4) Hyper-parameter setup: In the experiments, we evaluate multiple leadtime values from [1, 7, 14], and different input window sizes from [9, 15, 20]. For the target model and baselines involving the RNN module, the size of hidden states is selected from [12, 20, 32, 64], and the number of hidden

TABLE III: Comparison of RMSE on three stages of the COVID-19 data with leadtime=1, 7, 14 days. Boldface and underlined indicate the best and the second-best result of each column. Relative gain is the improvement of TLSS compared with the second best result. The p-value of the Diebold-Mariano test is to measure the significant improvement of TLSS with other baselines when window size = 20 on data3.

RMSE	Leadtime = 1												
	Window size = 9				Window size = 15				Window size = 20				
	data1	data2	data3	Average	data1	data2	data3	Average	data1	data2	data3	Average	p-value
AR	509	883	928	773	534	855	881	757	543	891	854	763	
ARMA	520	863	912	765	536	850	866	751	540	884	852	759	
RNN-Multilocation	**483**	827	950	753	498	829	870	732	518	858	862	746	***
RNN-Global	672	1079	1136	962	716	1096	1108	973	720	1103	1105	976	***
RNN-Global-keywords	704	1128	1189	1007	718	1145	1169	1011	730	1129	1163	1007	***
RNN-Global-doc2vec	707	1284	1228	1073	719	1272	1219	1070	725	1250	1189	1055	***
Cola-GNN	512	**816**	840	723	500	878	**800**	726	508	834	798	713	***
TLSS-Frozen	510	849	869	743	502	826	848	725	516	843	827	729	***
TLSS	502	820	**837**	**720**	**495**	**817**	807	**706**	**502**	**831**	**793**	**709**	-
% relative gain	-	-	0.4%	0.4%	1%	1.1%	-	2.6%	1.2%	1.6%	0.6%	0.6%	

RMSE	Leadtime = 7												
	Window size = 9				Window size = 15				Window size = 20				
	data1	data2	data3	Average	data1	data2	data3	Average	data1	data2	data3	Average	p-value
AR	548	1332	1049	973	586	1274	995	952	622	1347	947	972	***
ARMA	574	1231	1024	943	595	1260	977	944	615	1323	936	958	***
RNN-Multilocation	549	1136	1094	926	561	1127	1017	902	553	1090	988	877	**
RNN-Global	769	1493	1512	1258	778	1474	1456	1236	777	1482	1428	1229	***
RNN-Global-keywords	788	1555	1608	1317	786	1564	1559	1303	841	1506	1494	1280	***
RNN-Global-doc2vec	809	1666	1655	1377	818	1657	1614	1363	825	1684	1603	1371	***
Cola-GNN	546	1132	965	881	585	1156	953	898	551	1089	926	855	*
TLSS-Frozen	554	**1035**	1034	874	573	1112	996	894	544	1031	973	849	***
TLSS	**543**	1108	**941**	**864**	**559**	**1091**	**931**	**860**	**536**	**1031**	**912**	**826**	-
% relative gain	0.6%	-	2.5%	1.2%	0.4%	1.9%	2.3%	3.8%	1.5%	5.3%	1.5%	2.7%	

RMSE	Leadtime = 14												
	Window size = 9				Window size = 15				Window size = 20				
	data1	data2	data3	Average	data1	data2	data3	Average	data1	data2	data3	Average	p-value
AR	628	2071	1509	1403	664	2122	1361	1382	753	2094	1292	1380	**
ARMA	627	2032	1467	1375	662	2119	1331	1371	718	2103	1282	1368	**
RNN-Multilocation	**600**	1845	1481	1309	593	1846	1366	1268	597	1970	1286	1284	
RNN-Global	974	1769	1988	1577	943	1762	1925	1543	963	1768	1873	1535	***
RNN-Global-keywords	968	1782	2124	1625	938	1772	2079	1596	978	1796	2025	1600	***
RNN-Global-doc2vec	994	1874	2234	1701	982	1916	2281	1726	970	1936	2340	1749	***
Cola-GNN	644	1711	1392	1249	611	1759	1354	1241	594	1911	1279	1261	
TLSS-Frozen	705	1649	1625	1326	607	1634	1513	1251	575	1551	1410	1179	***
TLSS	606	**1638**	**1331**	**1192**	**570**	**1615**	**1284**	**1156**	**543**	**1548**	**1226**	**1106**	-
% relative gain	-	0.7%	4.4%	4.6%	3.9%	1.2%	3.5%	6.8%	5.6%	0.2%	4.1%	6.2%	

[a]p-values (* , **, *** indicate statistical significance at $p < 0.10$, $p < 0.05$, and $p < 0.01$)

layers is selected from [1, 2, 3]. All models are trained using the Adam optimizer [24] with a weight decay of 5e-4 and a dropout rate of 0.2. The initial learning rate is searched from the set [0.001, 0.005, 0.01], the training epoch is 1500, and the batch size is 32. We implement early stopping based on the validation loss.

When pre-training the source model Cola-GNN, the input window size T is 30, and and we keep the leadtime consistent with **TLSS**. For the RNN module, the size of hidden states is selected from [10, 20, 30] and the number of hidden layers is selected from [1, 2, 3]. In the multi-scale dilated convolution module in Cola-GNN, we set 10 filters, and long-term and short-term dilation rates equal to 2 and 1. The model is trained using the Adam optimizer with a weight decay of 5e-4 and a dropout rate of 0.2. The initial learning rate is 0.001, the training epoch is 1500, and the batch size is 32. The trained parameters of dilated convolution layers in the source model are shared with the target model **TLSS** and set as initializations.

B. Results

We evaluate the model performance on leadtime = [1, 7, 14] and window size T = [9, 15, 20] for three stages = [data1, data2, data3] of COVID-19. In table III, all the model results are summarized based on RMSE. The large difference in RMSE values across different stages of COVID-19 is because of the epidemic seasonality and other uncertainties. The large-scale outbreak peak periods of COVID-19 are in data2 (10/1/2021-3/31/2022) and data3 (4/1/2022-9/30/2022) stages. When leadtime = 1, our proposed method overcomes most baselines with relatively stable and optimal performance. Except for RNN-Global models, most approaches exhibit relatively good performance in capturing temporal patterns, which is due to the small information gap between the history window and the predicted time. When the leadtime is 7, our approach outperforms all other models in all three stages of the COVID-19 spread with different window sizes. Compared with some deep learning methods (RNN-multilocation, Cola-GNN, and **TLSS**), traditional statistical models AR and ARMA have decreasing performance when leadtime becomes longer. This shows the impact of model complexity on the time series forecasting with limited input. When the leadtime is 14, our model successfully captures the potential correlation between news articles and time series to achieves more accurate predictions than other models in most cases.

Only RNN-multilocation, Cola-GNN, and **TLSS**-Frozen have competitive prediction abilities with **TLSS** in some cases. The RNN-Global models are not performing well because they do not consider spatio-temporal dependencies. This demonstrates the importance of geolocation correlation in epidemic prediction at the population level. Specifically, RNN-Global-keywords and RNN-Global-doc2vec do not achieve good performance in the three stages of COVID-19 even with social factors (e.g., COVID-19 keywords and news embeddings). It suggests that the complex relationship between text data and numerical data cannot be easily captured. **TLSS** has better performance (higher % relative gain) with a large historical window because of the latency of news publication and the incubation time of the COVID-19 virus. Using the Diebold-Mariano Test [15, 18], we compare the improvement of **TLSS** with other baselines when window size = 20 on data3, which has the most observations among the three stages of COVID-19. Source model Cola-GNN achieves good performance with larger leadtime because it is designed for long-term epidemic prediction. Our proposed model **TLSS** shows statistically significant improvement in COVID-19 prediction in most cases, especially with leadtime = 7.

Overall, our proposed model outperforms all baseline methods in most situations. The models directly using news data (pre-trained embeddings or keywords) have weak performance. It suggests that temporal dependencies are hard to capture in text data. It is still a challenge to deeply learn the dynamic impact of social factors on epidemic diseases.

C. Ablation Tests

To evaluate the contribution of each component in our framework, we implement an ablation test on the three stages of COVID-19 data with the following settings: 1) **TLSS** w/o transfer learning: Remove the transfer learning architecture from the proposed method, only use COVID-19 time-series data as input in the model. 2) **TLSS** w/o sentiment analysis: Remove the sentiment analysis module. 3) **TLSS** w/o semantic analysis: Remove the semantic analysis module. We use RMSE to evaluate the model performance with each module in table IV. We observe that **TLSS** achieves the best performance on average for forecasting COVID-19 infection cases in three stages. We also implement the DM-Test to measure the improvement of each component within **TLSS** on data3. In most cases, the proposed model shows significant improvement with transfer learning architecture, because it can capture the regularity of existing diseases and transfer the learned knowledge to the target model for emerging disease predictions. Models that involve news sentiment and semantic analysis produce good results, especially when using a larger window size of historical data. The news sentiment has a more obvious impact than news semantic information on a COVID-19 outbreak in most cases, which indicates public emotion is crucial for epidemic prevention. The ablation test results show that our model can accurately predict epidemic transmission by capturing social factors' impact and by learning the general characteristics of existing pandemics.

TABLE IV: Ablation test result in three stages of COVID-19 with a leadtime of 14.

RMSE	**Window size = 9**				
	data1	data2	data3	Average	p-value
TLSS	606	**1638**	**1331**	**1192**	-
w/o transfer learning	593	1808	1409	1270	*
w/o sentiment	**592**	1864	1702	1386	***
w/o semantic	607	1852	1395	1285	***
RMSE	**Window size = 15**				
	data1	data2	data3	Average	p-value
TLSS	**570**	1615	**1284**	**1156**	-
w/o transfer learning	563	1778	1377	1239	*
w/o sentiment	597	1639	1552	1263	**
w/o semantic	574	**1569**	1340	1161	***
RMSE	**Window size = 20**				
	data1	data2	data3	Average	p-value
TLSS	**543**	**1548**	**1226**	**1106**	-
w/o transfer learning	576	1584	1341	1167	**
w/o sentiment	568	1566	1411	1182	*
w/o semantic	548	1785	1362	1232	

[a]p-values (* , **, *** indicate statistical significance at $p < 0.10$, $p < 0.05$, $p < 0.01$)

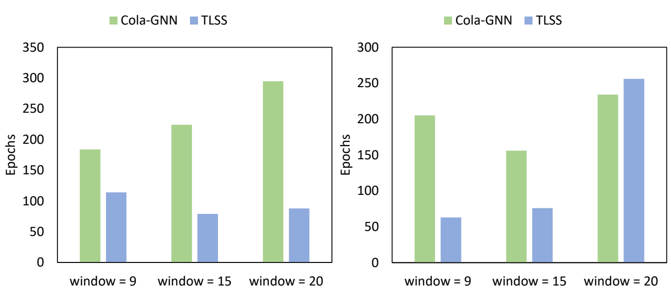
D. Model Complexity

In table V, we compare the runtime and the number of parameters for each model on the third stage of COVID-19 (data3) because it has the most observations among all three stages. In this work, models including AR, ARMA, and RNN-Global have the most efficient runtime because of lower dimension inputs and simpler model complexities. Our proposed model **TLSS** contains multiple features such as COVID-19 cases, news sentiments, and news semantics. Compared with other baseline models, it shows no significant adverse efficiency on training time. It achieves better performance than its source model Cola-GNN.

TABLE V: Model parameter and runtime comparison. Runtime is the time spent on a single GPU per epoch.

Model	Parameters	Runtimes
AR	500	0.02
ARMA	900	0.03
RNN-Multilocation	13K	0.33
RNN-Global	481	0.06
RNN-Global-keywords	16K	0.24
RNN-Global-doc2vec	16K	0.24
Cola-GNN	22K	0.86
TLSS-Frozen	18K	0.91
TLSS	18K	0.72

In figure 3, we compare the training epochs of the source model Cola-GNN and our proposed model **TLSS** on the COVID-19 dataset. Our proposed method achieves the best result with fewer training epochs in most cases. All programs are implemented using Python 3.9.4 and PyTorch 1.11.0 in an Ubuntu server with an Nvidia 1080Ti GPU.



(a) Leadtime = 7 (b) Leadtime = 14

Fig. 3: Average training epochs in the three stages of COVID-19 data with window size T.

E. Statistical Summary of Models

Figure 4 shows the statistical analysis of average RMSE for deep learning models (RNN-Multilocation, Cola-GNN, and **TLSS**) when leadtime = 14. We calculate RMSE on 10 randomized trials to observe the model improvement compared with two competitive baselines. Our proposed model **TLSS** achieves more accurate predictive performance with different window sizes.

V. CONCLUSION

In this work, we introduce the challenges of forecasting emerging epidemic disease (i.e., COVID-19) and propose a novel framework **TLSS** to address these issues. We design a heterogeneous transfer learning architecture to learn the standard patterns from existing relevant infectious diseases (i.e., ILI) and transfer the knowledge to a target model for COVID-19 prediction. We implement a social feature learning module to analyze the impact of social factors (e.g., news) on the spread of pandemics from both sentiment and semantic aspects. We adopt a location-aware attention mechanism to capture the dynamic correlation between news text data and time-series numerical data over time. We evaluate our proposed model on three stages of COVID-19 propagation and demonstrate its effectiveness and accuracy in predicting future infection cases with different lead times. In the future, we will extend the proposed framework to other types of diseases and investigate other source models. Furthermore, we will explore more complex factors such as local policies, cultures, and climate changes.

ACKNOWLEDGMENTS

This work is supported in part by the US National Science Foundation under grants 1948432 and 2047843.

REFERENCES

[1] WHO coronavirus (covid-19) dashboard. URL https://covid19.who.int/.

[2] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B. Aditya Prakash. Epideep: Exploiting embeddings for epidemic forecasting. In *25th ACM SIGKDD*, KDD '19, page 577–586, New York, NY, USA, 2019.

[3] Asif Afzal, C Ahamed Saleel, Suvanjan Bhattacharyya, N Satish, Olusegun David Samuel, and Irfan Anjum Badruddin. Merits and limitations of mathematical modeling and computational simulations in mitigation of COVID-19 pandemic: A comprehensive review. *Arch. Comput. Methods Eng.*, 29(2): 1311–1337, 2022.

[4] Hani Al-Dmour, Ra'ed Masa'deh, Amer Salman, Mohammad Abuhashesh, and Rand Al-Dmour. Influence of social media platforms on public health protection against the COVID-19 pandemic via the mediating effects of public health awareness and behavioral changes: Integrated model. *J. Med. Internet Res.*, 22(8):e19996, August 2020.

[5] Alexandra Balahur, Rada Mihalcea, and Andrés Montoyo. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech Language*, 28:1–6, 01 2014.

[6] Mesfin A Bekalu, Rachel F McCloud, and K Viswanath. Association of social media use with social well-being, positive mental health, and self-rated health: Disentangling routine use from emotional connection to use. *Health Educ. Behav.*, 46 (2_suppl):69–80, December 2019.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003.

[8] Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer-Verlag, Berlin, Heidelberg, 1986.

[9] Jianguo Chen, Kenli Li, Zhaolei Zhang, Keqin Li, and Philip S. Yu. A survey on applications of artificial intelligence in fighting against covid-19, 2020.

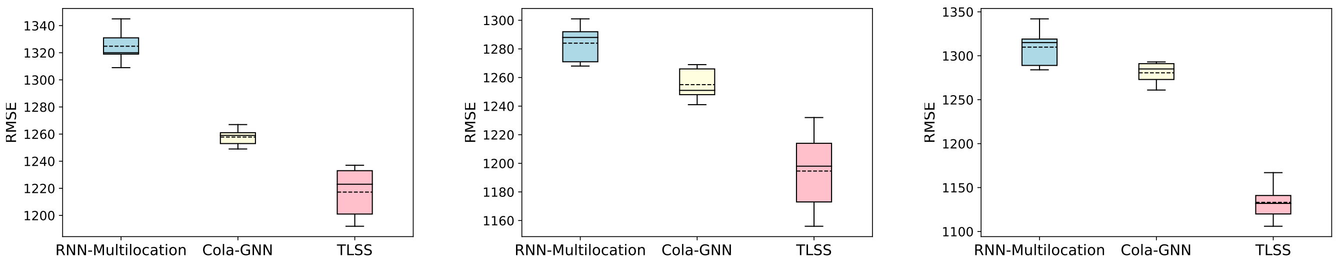
[10] Fatoumata Dama and Christine Sinoquet. Time series analysis and modeling to forecast: a survey, 2021.

[11] Oscar Day and Taghi M Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4(1):29, September 2017.

[12] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Cola-gnn: Cross-location attention based graph neural networks for long-term ili prediction. In *29th ACM CIKM*, CIKM '20, page 245–254, New York, NY, USA, 2020.

[13] Erhu Du, Eddie Chen, Ji Liu, and Chunmiao Zheng. How do social media and individual behaviors affect epidemic transmission and control? *Science of The Total Environment*, 761: 144114, 2021.

[14] Wei Duan, Zongchen Fan, Peng Zhang, Gang Guo, and Xiaogang Qiu. Mathematical and computational approaches to



(a) Window Size = 9 (b) Window Size = 15 (c) Window Size = 20

Fig. 4: Statistics analysis of the prediction results from RNN-Multilocation, Cola-GNN, and TLSS.

epidemic modeling: a comprehensive review. *Frontiers of Computer Science*, 9(5):806–826, October 2015.

[15] f. x. and r. s. mariano. comparing predictive accuracy. *journal of business and economic statistics*, 13:253–263.

[16] CDC Case Task Force. United states covid-19 cases and deaths by state over time.

[17] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722, 2014.

[18] David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, Jun 1998.

[19] Michael J Horry, Subrata Chakraborty, Manoranjan Paul, Anwaar Ulhaq, Biswajeet Pradhan, Manas Saha, and Nagesh Shukla. COVID-19 detection through transfer learning using multimodal imaging data. *IEEE Access*, 8:149808–149824, August 2020.

[20] C J Hutto and E E Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM-14*. Ann Arbor, MI, 2014.

[21] Aayush Jain, Tanay Sukhdeve, Himanshu Gadia, Satya Prakash Sahu, and Satya Verma. Covid19 prediction using time series analysis. In *2021 ICAIS*, pages 1599–1606, 2021.

[22] Rohit Kumar Kaliyar. A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of bert. In *2020 10th Confluence*, pages 336–340, 2020.

[23] Juhyeon Kim and Insung Ahn. Weekly ILI patient ratio change prediction using news articles with support vector machine. *BMC Bioinformatics*, 20(1):259, May 2019.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[25] Sameer Kumar, Chong Xu, Nidhi Ghildayal, Charu Chandra, and Muer Yang. Social media effectiveness as a humanitarian response to mitigate influenza epidemic and COVID-19 pandemic. *Annals of Operations Research*, January 2021.

[26] Rabindra Lamsal, Aaron Harwood, and Maria Rodriguez Read. Twitter conversations predict the daily confirmed covid-19 cases, 2022.

[27] Kalev Leetaru and Philip A. Schrodt. Gdelt: Global data on events, location, and tone. *ISA Annual Convention*, 2013.

[28] Zhenyu Liu, Zhengtong Zhu, Jing Gao, and Cheng Xu. Forecast methods for time series data: A survey. *IEEE Access*, 9:91896–91912, 2021.

[29] G. Mahalakshmi, S. Sridevi, and S. Rajaram. A survey on forecasting of time series data. In *ICCTIDE'16*, pages 1–8, 2016.

[30] Maryam Mahdikhani. Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of covid-19 pandemic. *IJIM Data Insights*, 2(1):100053, 2022.

[31] Spyros Makridakis. A survey of time series. *ISR*, 44(1):29–70, 1976.

[32] Cory Merow and Mark C Urban. Seasonality and uncertainty in global COVID-19 growth rates. *Proc. Natl. Acad. Sci. U. S. A.*, 117(44):27456–27464, November 2020.

[33] Seungwhan Moon and Jaime Carbonell. Completely heterogeneous transfer learning with attention - what and what not to transfer. In *IJCAI-17*, pages 2508–2514, 2017.

[34] David M Morens and Anthony S Fauci. Emerging infectious diseases in 2012: 20 years after the institute of medicine report. *MBio*, 3(6), December 2012.

[35] David M Morens, Peter Daszak, and Jeffery K Taubenberger. Escaping pandora's box - another novel coronavirus. *N. Engl. J. Med.*, 382(14):1293–1295, April 2020.

[36] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *JAMIA*, 18(5):544–551, 09 2011.

[37] Shanoli Samui Pal and Samarjit Kar. Fuzzy transfer learning in time series forecasting for stock market prices. *Soft Comput.*, 26(14):6941–6952, jul 2022.

[38] Robert Pardo. *WALK-FORWARD ANALYSIS*. John Wiley amp; Sons, Inc., 2011.

[39] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, Doha, Qatar, October 2014. ACL.

[40] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019.

[41] Alexander Rodríguez, Nikhil Muralidhar, Bijaya Adhikari, Anika Tabassum, Naren Ramakrishnan, and B. Aditya Prakash. Steering a historical disease forecasting model under a pandemic: Case of flu and COVID-19. *CoRR*, abs/2009.11407, 2020.

[42] Alok Kumar Sahai, Namita Rath, Vishal Sood, and Manvendra Pratap Singh. ARIMA modelling & forecasting of COVID-19 in top five affected countries. *Diabetes Metab. Syndr.*, 14 (5):1419–1427, September 2020.

[43] Yasir Ali Solangi, Zulfiqar Ali Solangi, Samreen Aarain, Amna Abro, Ghulam Ali Mallah, and Asadullah Shah. Review on natural language processing (nlp) and its toolkits for opinion mining and sentiment analysis. In *2018 IEEE 5th ICETAS*, pages 1–4, 2018.

[44] Lijing Wang, Jiangzhuo Chen, and Madhav V. Marathe. Defsi: Deep learning based epidemic forecasting with synthetic information. In *AAAI*, 2019.

[45] Congxi Xiao, Jingbo Zhou, Jizhou Huang, An Zhuo, Ji Liu, Haoyi Xiong, and Dejing Dou. C-watcher: A framework for early detection of high-risk neighborhoods ahead of covid-19 outbreak. *AAAI*, 35(6):4892–4900, May 2021.