Regression Model for Prediction of Epidemic Outbreaks

Punna Sai Teja Shashank
Computer Science and Engineering, Hyderabad
Shashankpunna21@gmail.com

Akkaladevi Vaibhavi
Computer Science and Engineering, Hyderabad
vaibhavi.hyd@gmail.com

Jenega Vaishnavi
Computer Science and Engineering, Hyderabad.
vaishu9121@gmail.com

Dr M.A Jabbar
Professor
Computer Science and Engineering
Vardhaman College of Engineering, Hyderabad
jabbar.meerja@gmail.com

ABSTRACT:

Epidemic diseases are contagious diseases that can spread into the entire nation if the contagion measurement had reached the outbreak level and there is a chance to wipe out the entire population. So, there must be a way to track these epidemics to take counter measurements to restrict its spreading. There are many ways to track these epidemics. Nowadays, machine learning is applied in almost all the fields since data is generating rapidly from various sources. The data which is useful for tracking epidemics is also very high around us. Various machine learning techniques can be implemented to find out the status of epidemics. Regression is a machine learning algorithm that is used to find the relationship between a dependent variable and an independent variable. For fitting a regression line, the most common method is the method of least-squares. In this method, the best-fitting line is calculated for the observed data such that the sum of the squares of the vertical deviations from each data point to the line is minimum. In this paper, we proposed a Regression-based prediction system that is used to predict the status of an epidemic in a region. Linear regression is used for this purpose to predict the number of deaths in a region on a specific date. Various libraries of python like NumPy, pandas, matplotlib, and sci-kit-learn are used to develop the model.

KEYWORDS:
Epidemic Prediction, Electronic Health Record, Linear Regression, Covid-19, Machine Learning, Influenza

1.INTRODUCTION:

Epidemic Outbreaks are one of the major threats in the 21st century. They occur when there is a sudden increase in the number of cases of a disease. They spread very rapidly. The time period of this epidemic is usually unknown. It can last up to days, weeks, months, or years. Scientists are still finding out a way to know the time period of these epidemics. It kills almost half of the population if the contagion measurement had reached the outbreak level and expands into the entire nation. Epidemic Outbreaks have been confirmed to create severe world health issues and have caused many life losses and wiped out 41% of the population globally. Examples of epidemic outbreak seen earlier are Cholera, Dengue, Influenza, Bird flu, Pneumonia, Covid-19, etc. In recent years, there have been advances in collecting health care data. Epidemics are usually caused by direct contact from one person to another and contaminated air. By considering the causes and features of epidemics, there should be a way to monitor or track these epidemics to take counter measurements to control it.

For this reason, we proposed a predictive model, to predict the status of the epidemic in a region using social media data. Since machine learning is used in many fields to extract valuable information which requires a huge amount of data and social media is generating a huge amount of data every minute. So, we applied machine learning on social media data...
for prediction. We collected the dataset from the data.CDC.gov website. We imported different types of modules and libraries like NumPy, pandas, and sci-kit learn which helps in Pre-processing, regression, and data cleaning. Data cleaning helps to discard unnecessary information in the dataset. We implemented this prediction system by applying certain machine learning algorithms like linear regression. It is a commonly used technique to build a relationship between two sets of variables. The result is a linear regression equation that can be used to predict the disease outbreak. The early detection and prediction of the spread of epidemics can be very useful and faster response can take place. We also used Data Visualization for a better understanding of the data. The prediction model on the resultant data gives us information about the number of deaths in a given region or location on a certain day, week, or month. This paper deals with the methodology to predict the status of the epidemic with the application of Linear Regression. Electronic Health data is chosen for simulation.

The remaining part of the paper is ordered in this way. Section 2 is assigned to related work. Section 3 describes the methodology and Section 4 contains about our proposed work. The experimental result is discussed in Section 5. We conclude our remarks and gives direction in Section 6.

2.RELATED WORKS:

Machine Learning in the healthcare field can be used for effective disease prediction, diagnosis, and treatments. This section discusses related works on epidemic outbreak prediction by mining social media data.

Predicting Outbreak Detection in Public Health Surveillance: Quantitative Analysis to Enable Evidence-Based Method Selection is proposed by David L Buckeridge, Anya Okhatovskaia, Samson W Tu, Martin Joseph O'Connor[1]. They developed and validated quantitative models for predicting the ability of commonly used surveillance algorithms which are used to detect different types of outbreaks using simulated outbreak data. They used the BioSTORM software test-bed developed by their group to encode the C-algorithms and to evaluate their ability. They recorded characteristics of each outbreak signal and the parameters of the algorithms along with the results for detection.

Detecting and Tracking Disease Outbreaks by Mining Social Media Data is proposed by Yusheng Xie1,2 Zhengzhang Chen1,2 Yu-Cheng Kunpeng Zhang Kathy Lee Ankit Agrawal Wei-Keng Liao Alok Choudhary2 [2]. They stated the real-time challenge in detecting and tracking epidemic outbreaks. They aimed to provide a platform for researchers in the domain to work on the real-time challenge by releasing relevant datasets from social media. They made four distinct social media datasets publicly available:
1. Health-related blogs (BL)
2. Facebook public walls of medical brands and organizations (FBW)
3. Facebook public posts mentioning health-related terms (FBP)
4. Media uploads from Instagram (INST).

Using online social networks to track a pandemic: A systematic review is proposed by Muhammad Sadiq Khan Kasturi Dewi Varathan Ghulam Mujtaba Abdelkodose M Al-Kabsi[3]. They conducted a systematic review of the use of Online Social Network to track a pandemic and presented challenges and future implications of using them. The electronic literature search is conducted for eligible English articles published between 2004 and 2015 and concluded that the outcomes of online social network surveillance systems have demonstrated high correlations with the findings of official surveillance systems with some limitations.
Comparing Social media and Google to detect and predict severe epidemics is proposed by Loukas Samaras, Elena Garcia-Bariocanal\& Miguel-Angel Sicilia [4]. The authors aimed at gathering evidence on which kind of data sources lead to better results. Data on influenza in Greece are collected from Google and Twitter and compared to influenza data from the official authority of Europe. The data were analyzed by using two models: the ARIMA model and a customized approximate model. Google data showed a high Pearson correlation and a relatively low Mean Absolute Percentage Error \(R = 0.933, \text{MAPE} = 21.358\) whereas Twitter results gave \(R = 0.943, \text{MAPE} = 18.742\).

3. METHODOLOGY:

This section will discuss the methodology adopted by our proposed methodology. As stated earlier, this research paper aims to develop a predictive model to predict the status of an epidemic in a region. We used the Deaths involving coronavirus disease 2019 (COVID-19), pneumonia, and influenza data set given by the Centers for Disease Control and Prevention (Data.CDC.gov), which is partitioned into training and test set.

3.1 DATASET:

Our dataset contains the details of the number of deaths by various epidemics like Covid-19, Influenza, and Pneumonia in different regions. Our dataset contains 1782 rows and 14 columns before pre-processing, which is 33 rows for a location. After removing the uninterested columns and rows containing missing values the resulting dataset contains 720 rows and 11 columns. To predict the status of the epidemic in a location, the rows of only that location are extracted to build the model. Here the independent variable is the date attribute and the dependent variable is the Covid-19 deaths. The dataset used in our methodology is shown in figure 1.

![Figure 1: Dataset used for our model](image)

3.2 LINEAR REGRESSION:

To find the relationship between a dependent variable i.e. COVID-19 deaths and an independent variable i.e. date, the best and simple regression algorithm is linear regression. There are several variations of linear regression. If there is only one explanatory variable, then it is called simple linear regression and if there is more than one explanatory variable then it is called multiple linear regression [9].
For fitting a regression line, the least square method is used. The regression line is to be drawn for the observed data. This line is drawn such that the sum of the squares of the vertical deviation from each point to the line is minimum [9].

![Best Fit line](image)

Figure 2. Best Fit line
To decide the best fit line, we use the following formula: \( \sum(Y_0 - Y_p)^2 \)

3.3 Evaluation Measures:

To evaluate the accuracy of our model we used the following evaluation measures:

- Explained variance score
- Max error
- Mean absolute error
- Mean squared error
- Mean squared logarithmic error
- Median absolute error
- \( R^2 \) error, the coefficient of determination

4. PROPOSED APPROACH:

As stated earlier, this paper aims to propose a predictive model to predict the status of an epidemic in a region. A dataset containing details of the number of deaths by various epidemics like Covid-19, Influenza, Pneumonia in different regions on different days is collected. The dataset is preprocessed to remove noise and redundancy. Further, the dataset is divided into two parts called train dataset and test dataset in the ratio of 70:30, and the linear regression algorithm is applied to the training dataset to predict the number of deaths in a region on a specific date. Lastly, our prediction model is tested using various performance measures.

Steps in our proposed method are as follows:

Step 1: Load the Dataset
Step 2: Apply preprocessing to the data set
Step 3: Visualize the data by various plots
Step 4: Build the predictive model using Linear Regression
Step 5: Predict the number of deaths by an epidemic in a region on a given date.

Algorithm: Predictive Model to predict the status of epidemic in a region

Loading is done in step 1 with the help of python’s library called pandas. The dataset is loaded into a datatype called data frame which is like a two-dimensional array with indexes
on both columns and rows with the help of which we can perform effective operations on datasets.

Pre-processing is done in step 2 to remove the uninterested columns and rows containing missing and misleading values with the help of python’s libraries namely NumPy and pandas.

Since in our dataset, the date column is having dates of different formats i.e. dd/mm/yyyy and dd-mm-yyyy. We made all these dates to a single format i.e. dd-mm-yyyy. And also, the column COVID-19 deaths had commas in the data. So, we removed those commas and converted the string type to integer type.

Data Visualization is done in step 3 to observe the patterns, trends, and outliers in the data using the python’s library called matplotlib. We observed the relation between date and number of COVID-19 deaths by plotting various graphs like bar, line, and scatter. The outputs of bar graphs and scatter graphs are shown in figure 2 and figure 3.

Dataset is divided into training and testing datasets in the ratio of 70:30 and the Linear regression algorithm is applied to the training dataset in step 4 to build our prediction model. Further, we estimated the accuracy of our prediction model with the help of various performance measures in step 5.

The data flow in our proposed approach is pictorially shown in figure 4.
5. Experimental Results:

Experiments for the proposed model is conducted on the Deaths involving coronavirus disease 2019 (COVID-19), pneumonia, and influenza data set given by the Centers for Disease Control and Prevention (Data.CDC.gov). The dataset is partitioned into training and testing datasets. The training dataset is used to build the predictive model and the testing dataset is used to test the instances. The proposed model is evaluated using the following metrics:

1. R-squared ($R^2$): \[ \frac{\text{var}(y - y^*)}{\text{var}(y)} \]

2. Explained variance score: \[ 1 - \frac{\text{Var}(y_i - y_i^*)}{\text{Var}(y)} \]

3. Max Error: \[ \max(|y_i - y_i^*|) \]

4. Mean absolute error: \[ \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - y_i^*| \]

5. Mean Squared error: \[ \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - y_i^*)^2 \]

6. Mean squared logarithmic error: \[ \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (\log_e(1+y_i) - \log_e(1-y_i))^2 \]

7. Median absolute error: \[ \text{median}(|y_i - y_i^*, \ldots, y_n - y_n^*|) \]

The results obtained by our prediction model on the test dataset is shown in table 1.

| X_TEST | Y_TEST | PREDICTED OUTPUT |
|--------|--------|------------------|
| 737491 | 13131  | 5300.36          |
| 737456 | 3787   | 5623.94          |
| 737484 | 571    | 5023.02          |
| 737561 | 0      | 4791.89          |
| 737575 | 5610   | 5762.61          |
| 737463 | 5313   | 6039.96          |
| 737477 | 0      | 4699.44          |
| 737680 | 7301   | 5947.51          |
| 737470 | 17052  | 5207.91          |
| 737554 | 6478   | 5993.74          |

The results of different evaluation metrics for our model is shown in table 2.

| ACCURACY MEASURE       | SCORE   |
|------------------------|---------|
| Explained variance score | 0.03    |
| Max Error              | 6618.01 |
| Mean absolute Error    | 4075.24 |
| Mean squared Error     | 21670427.64 |
| Mean squared Logarithmic Error | 15.03 |
| Median absolute Error  | 3889.99 |
| R-square Score         | -1.80   |
6. CONCLUSION:
In this paper, we proposed a model for predicting the status of an epidemic in a specific area, region, or location by predicting the number of deaths occurring in a specific location. The use of conventional methods is not sufficient enough for faster and accurate results. So, we use this new approach in finding out the epidemic in earlier stages. Data cleaning, pre-processing, and machine learning algorithms like linear regression are used for this model. We use the Least Square Method for fitting the regression line. Data Visualization is used to understand the data and relationships between dependent and independent variables in a better way. We have also tested our predicted model using various kinds of performance measures. We evaluated the accuracy of linear regression by using R-squared(R2), Root-mean squared error, Residual Standard Error, Mean Absolute Error. By implementing the steps of our approach, we have proved it to be successful in predicting the epidemic.

REFERENCES:
[1]. Predicting Outbreak Detection in Public Health Surveillance: Quantitative Analysis to Enable Evidence-Based Method Selection David L. Buckeridge, MD, PhD1, Anna Okhmatovskaia PhD1, Samson Tu MS2, Martin O’Connor MS2, CsongorNyulas MS2 , Mark A. Musen, MD PhD2 1 McGill University, Montreal, QC; 2 Stanford University, Palo Alto, CA
[2]. Detecting and Tracking Disease Outbreaks by Mining Social Media Data Yusheng Xie1,2 Zhengzhang Chen1,2 Yu Cheng Kunpeng Zhang Kathy Lee Ankit Agrawal Wei-keng Liao Alok Choudhary2
[3]. Using online social networks to track a pandemic: A systematic review Mohammed Ali Algaradi, Muhammad Sadiq Khan, Kasturi DewiVarathan, Ghulam Mujtaba, AbdelkodoseM.Al-Kabisi. Journal of Biomedical InformaticsVolume 62, August 2016, Pages 1-11
[4]. Comparing Social media and Google to detect and predict severe epidemics Loukas Samaras, Elena Garcia-Barriocanal &Miguel-Angel Sicilia. Available at: https://www.nature.com/articles/s41598-020-61686-9
[5]. A review of influenza detection and prediction through social networking sites Ali Alessa and MiadFaezipour. Available at: https://pubmed.ncbi.nlm.nih.gov/29386017/
[6]. TRENDING NOW:Using Social Media to Predict and Track Disease Outbreaks. Charles W. Schmidt, MS. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3261963/
[7]. How social media can be used to track disease outbreaks. Article by healthcareglobal.com. Available at: https://www.healthcareglobal.com/technology-and-ai-3/how-social-media-can-be-used-track-disease-outbreaks.
[8]. Can Social Media Predict Disease Outbreaks? An Article in healthadministrationdegrees.com. Available at: https://www.healthcareglobal.com/technology-and-ai-3/how-social-media-can-be-used-track-disease-outbreaks
[9]. Linear Regression Article in wikipedia. Available at: https://en.wikipedia.org/wiki/Linear_regression.
[10]. Forecasting Influenza Levels using Real-Time Social Media Streams Kathy Lee, Ankit Agrawal, Alok Choudhary EECS Department, Northwestern University, Evanston, IL {Kathy lee,ankitag.choudhari}@eecs.northwestern.edu
[11]. The Real-World Benefits of Machine Learning in Healthcare by Ed Corbett Medical Officer. Article in healthcatalyst.com. Available at: https://www.healthcatalyst.com/clinical-applications-of-machine-learning-in-healthcare
[12]. Machine Learning Applications Enhancing Healthcare Sector 2020 by Prashant Kathuria. Blog in upgrade.com. Available at:https://www.upgrad.com/blog/machine-learning-applications-in-healthcare/
[13]. Numpy documentation in numpy.org. Available at:https://numpy.org/doc/
[14]. Pandas documentation in pandas.pydata.org. Available at: https://pandas.pydata.org/docs/
[15]. Matplotlib documentation in matplotlib.org. Available at: https://matplotlib.org/3.3.2/contents.html
[16]. Scikit-Learn documentation in scikit-learn documentation. Available at:https://scikit-learn.org/stable/

[17]. Regression Model Accuracy Metrics: R-square, AIC, BIC, Cp and more. By kassambara. Article in STHDA.com. Available at:http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/