Data Analytics for the COVID-19 Epidemic*

Ranran Wang* Gang Hu* Chi Jiang*
ran.ran.wang@stu.zuel.edu.cn hugang@stu.zuel.edu.cn cjiang@stu.zuel.edu.cn
*Zhongnan University of Economics and Law, Wuhan, China

Huimin Lu† Senior Member, IEEE Yin Zhang‡ Senior Member, IEEE
dr.huimin.lu@ieee.org yin.zhang.cn@ieee.org
†Kyushu Institute of Technology, Japan ‡University of Electronic Science and Technology of China, Chengdu, China

Abstract—With the spread of COVID-19 worldwide, people¡¯s production and life have been significantly affected. Artificial intelligence and big data technologies have been vigorously developed in recent years. It is very significant to use data science and technology to help humans in a timely and accurate manner to prevent and control the development of the epidemic, maintain social stability and assess the impact of the epidemic. This paper explores how data science can play a role from the perspectives of epidemiology, social networking, and economics. In particular, for the existing epidemic model SIR, we present a parameter learning method using particle swarm optimization (PSO) and the least squares method, and use it to predict the trend of the epidemic. Aiming at the social network data, we provide a specific method to realize sentiment analysis during the epidemic and propose an explainable fake news detection technique based on a variety of data mining methods.

Index Terms—Data science, epidemic, data mining, sentiment analysis, fake news detection

I. Introduction

DURING the past 300 years, data science has been an important auxiliary tool in epidemiological research during the human struggle against epidemics. The application of data science to epidemic disease control is, of course, not just to the daily data of the epidemic. It is also an important tool to help us understand the epidemic's infectious characteristics, their patterns, and the effectiveness of control strategies. In summary, the advantages that data science can play in an epidemic include the following:

- Predict the development of the epidemic and assist in prevention and control. For example, big data technology is used to issue epidemic early warning to specific populations and regions, and the data analysis technology is used to assist the deployment of materials.
- Assist the medical rescue process and assist drug screening. For example, big data technology can be used to assist in tracking susceptible people and screening effective drugs for prevention or treatment.
- Use social networks to understand public opinion and maintain social order during the epidemic. For example, use relevant data to analyze the sentiment of the masses, detect false information, and so on.
- Use economic data to estimate the possible impact of a pandemic, so that the government can take precise measures and society can respond in a timely manner.

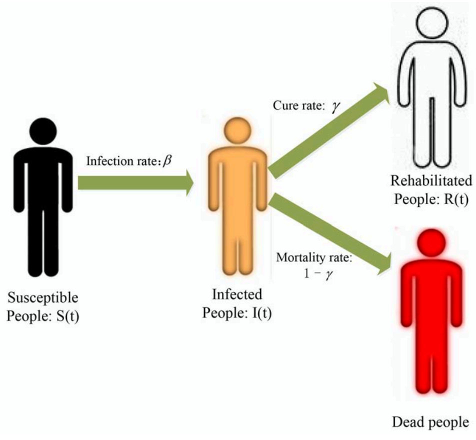

Fig. 1. Classic SIR Model

COVID-19, a virus causing a terrible pandemic outbreak, was identified at the end of 2019. After spreading for several months, it has caused serious losses to various countries in the world. With today's relatively advanced data science, if we can use big data technology to advance a rapid, accurate and timely grasp of the development trend of the epidemic or to predict the response of the people and take more effective prevention and control measures, we can expect to reduce the negative impact of epidemics on people's production and life [1]. During the COVID-19 epidemic, many researchers have begun to use data science technologies such as artificial intelligence and big data to help people fight the epidemic. For example, Wu et al. [2] predicted the spread of the virus in China and around the world using monthly flight booking data from official aviation guides and turnover data from more than 300 prefecture-level cities in mainland China from

Supported by the National Natural Science Foundation of China under Grant 61702553.
Corresponding Author: Yin Zhang

978-1-7281-7303-0/20/$31.00 ©2020 IEEE
DOI 10.1109/COMPSAC48688.2020.00-83

Tencent. Huang et al. [3] collected and analyzed data from patients with confirmed COVID-19 infections and data from relevant clinical visits through real-time RT-PCR and next-generation sequencing, and concluded that COVID-19 infection caused severe respiratory diseases, similar to severe acute respiratory syndrome coronavirus, and was associated with ICU entry and high mortality. Zhao et al. [4] used an exponential growth model to model the epidemic curve of COVID-19 cases in mainland China between December 1, 2019 (solstice) and January 24, 2020, and confirmed that the initial growth stage of the outbreak followed an exponential growth pattern. Fong et al. [5] constructed a polynomial neural network (PNN cf) with correction feedback under the condition of insufficient available data. To support the decision-making and logistics planning of the health care system, Yan et al. [6] used a database of blood samples from 404 infected patients in Wuhan, China, and used machine learning tools to select three biomarkers to predict the survival rate of individual patients. During the outbreak of COVID-19, artificial intelligence and big data technology have helped people solve many problems.

It is against this background that this paper will summarize the role of data science and technology in epidemics from the perspective of epidemiological data, social network data, and socioeconomic data, and propose how to use advanced data mining techniques to analyze epidemic situations to scientifically and rationally lay the foundation for later development and to provide a reference experience for unknown future epidemics.

The remainder of the paper is organized as follows: Section II will analyze and predict the development of the epidemic from the perspective of epidemiology. Section III and Section IV start from social network data and explains how data science and technology can be used to monitor public opinion and detect fake news. Section V summarizes the whole paper.

II. Prediction of COVID-19 development based on SIR model

In this section, we will introduce how to use the SIR model optimized by least squares and particle swarm optimization to predict the development of COVID-19 in combination with epidemic data.

The SIR model is a classic epidemic model. Its basic process is shown in Figure 1. The model first assumes that there is an upper limit on the number of people infected $S(t)$ during the spread of the epidemic. Among the susceptible people $S(t)$, a daily rate β of susceptible people will transform into diseased people. The total population affected is $I(t)$, and among the total population affected, a daily ratio γ of the population will be transformed into a cured population. The final total cured population is $R(t)$. In 1927, after studying historical infectious disease data, Kermark et al. [7] used mathematical models to summarize the SIR model, which has the following form:

$$\frac{\partial S}{\partial t} = -\beta \cdot S \cdot I \tag{1}$$

$$\frac{\partial I}{\partial t} = \beta \cdot S \cdot I - \gamma \cdot I \tag{2}$$

$$\frac{\partial R}{\partial t} = \gamma \cdot I \tag{3}$$

The key of the SIR model lies in the determination of the model parameters. In machine learning, there are very reliable ways to determine parameters, such as the least squares method and the particle swarm optimization algorithm. Here, we consider using the least squares method and particle swarm optimization to learn the parameters β and γ.

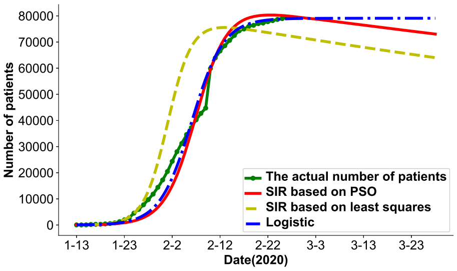

Fig. 2. Forecasting the Development Trend of Epidemic

A. The basic principle of particle swarm optimization

PSO [8] was first proposed by Eberhart and Kennedy in 1995, and its basic concept originated from the research of the foraging behavior of birds. A particle is used to simulate an individual bird, each particle can be regarded as a search individual in the N-dimensional search space, the current position of the particle is a candidate solution of the corresponding optimization problem, and the flight process of the particle is the search process of the individual. The particle velocity can be dynamically adjusted according to the optimal position of the particle history and the population history. Particles have only the two properties of speed and position, with speed representing how fast they move and position representing the direction they move. The optimal solution sought by each particle individually is called the individual extreme value, and the optimal individual extreme value in the particle swarm is the current global optimal solution. The model continues to iterate, updating the speed and location. Finally, the optimal solution satisfying the termination condition is obtained.

The formulas to update the speed and position are as follows:

$$\begin{aligned} V_{id} = \omega V_{id} + C_1 random(0,1)(P_{id} - X_{id}) \\ + C_2 random(0,1)(P_{gd} - X_{id}) \end{aligned} \tag{4}$$

$$X_{id} = X_{id} + V_{id} \tag{5}$$

Where ω is the inertia factor and its value is nonnegative. When it is large, the global optimization ability is strong and the local optimization ability is strong. For a relatively small ω, the global search ability is weak and the local search ability is strong. C_1 and C_2 are acceleration constants. The former is the individual learning factor of each particle, while the latter is the social learning factor of each particle. Generally, $C_1 = C_2 = 2$, but it does not have to be 2. Generally, $C_1 = C_2 \in [0, 4]$. P_{id} represents the d-dimension of the individual extreme value of the i-th variable and P_{gd} represents the d-dimension of the global optimal solution.

B. The basic principle of the least squares method

The least squares method is a mathematical optimization technique that finds the best function matching of data by minimizing the sum of the squares of the errors. The unknown data can be obtained easily and the sum of the squared errors between the obtained data and the actual data can be minimized. In principle, the position of the line is determined by the "minimum sum of squares of residuals" (in mathematical statistics, residuals refer to the difference between the actual observed and estimated values).

For example, for the unary linear regression model, it is assumed that n groups of observed values $(X_1, Y_1), ((X_2, Y_2), \ldots, (X_n, Y_n)$ are obtained from the population. For the n points in the plane, an infinite number of curves can be used to fit. Linear regression requires the sample regression function to fit the set of values as much as possible, that is, the line should be at the center of the sample data as much as possible. Therefore, the criterion for selecting the best fitting curve can be determined as minimizing the total fitting error (that is, the total residual).

Suppose the fitting line is y = ax + b for any sample point (x_1, y_1) and the error is $e = y_i - (ax_i + b)$. When $S = \sum_{i=1}^{n} e_i^2$ is the smallest, the fitting degree is the highest. To solve the problem, the first partial derivatives are obtained as follows:

$$\frac{\partial S}{\partial b} = -2(\sum_{i=1}^{n} y_i - nb - a\sum_{i=1}^{n} x_i) \tag{6}$$

$$\frac{\partial S}{\partial a} = -2(\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i - a\sum_{i=1}^{n} x_i^2) \tag{7}$$

Set the above two equations equal to 0, and obtain the final solution, as follows::

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{8}$$

$$b = \bar{y} + a\bar{x} \tag{9}$$

The basic idea of using the least squares estimation coefficient in the SIR model is to first estimate the value interval of $s(0)$ and β, and then find the group of values with the smallest residual sum of squares between the predicted value and the real value calculated in the interval. This is used to predict the estimated value.

C. Results of COVID-19 trend prediction

We also consider the classic logistic regression model. The SIR model based on particle swarm optimization and the SIR model based on the least squares method and the logistic regression model are compared. The final prediction is shown in Figure 2. In the experiment shown in Figure 2, we used the data on the number of patients with a new type of coronary pneumonia nationwide between January 13 and February 28 and trained the logistic regression model and the parameter estimation of the SIR model. The estimations use a particle swarm algorithm and the least squares method, respectively. After obtaining the trained model or estimated parameters, we re-predicted the number of patients between January 13 and March 27 and plotted it as a curve. Additionally, in the graph, we present a line chart of the actual epidemic trend as a comparison with other forecast curves.

It can be seen from the results that the logistic regression model is obviously better than the SIR model based on particle swarm optimization in terms of conformity with the real situation. One possibility we considered is that logistic regression is a mathematical model based entirely on machine learning. During the data fitting process, it can fully fit the changes of the curve. Therefore, its changes are more in line with the historical curve. The SIR model is an empirical model. In addition to the parameters learned in the epidemic data, it has an independent mathematical model based on historical epidemic data observations. Therefore, in predicting the trend of the epidemic, it adheres to the general law of the development of the COVID-19 epidemic, rather than deliberately fitting the trend of the curve.

III. Sentiment analysis based on social network data during the epidemic

To better explain how to use data mining technology to assist public opinion analysis, we use the Weibo data as an example. From the aspects of data preparation, data preprocessing, feature embedding, classification model, and experimental results, we will introduce in detail how to use social network data to analyze the sentiment polarity of online people.

A. Data preparation and preprocessing

We crawled Weibo data related to the COVID-19 topic between January 1, 2020 and February 20, 2020. The final dataset includes the Weibo ID, Weibo posting time, publisher account, and Weibo text. In total, we sorted 100,000 pieces of data. We hope to use data science related methods to determine the sentiment polarity of the sorted texts and label them into the three categories of 1 (positive), 0 (neutral), and -1 (negative).

Then, we remove the abnormal data in the original data and perform word segmentation to obtain high-quality, standardized and valid data. This improves the reliability and accuracy of the analysis. We simply obtain the word cloud of positive emotion and the word cloud of negative emotion in the dataset through python tools.

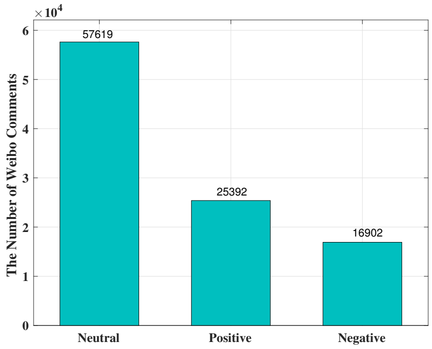

Fig. 3. Statistics of Sentimental Polarity

In addition, we developed statistics on the sentimental polarity in the entire dataset, as shown in Figure 3. As seen, the number of Weibo comments with positive emotions reaches more than 20,000, the number of Weibo comments with negative emotions is less than 20,000, and the number of Weibo comments with neutral emotions is close to 60,000. Overall, netizens' comments on epidemic related topics on Weibo tend to be positive.

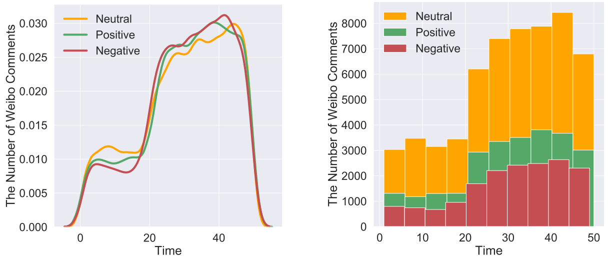

Fig. 4. News length statistics

Furthermore, we use the probability density estimation method to develop statistics on the relationship between the frequency of public opinion and time, as shown in Figure 4. The timeline starts on January 1, 2020. After January 19, the number of topics has increased significantly. On approximately February 7th, the number of Weibo comments reached a peak, and the sentimental tendency of the comments was mainly neutral, followed by positive comments, and then negative comments. The following week, the number of Weibo comments decreased rapidly.

B. Feature embedding

The next step is to embed the text features of the word segmentation into the word vector. Word2Vec is used. It can represent text as a relatively low-dimensional, real-valued vector. This process is called word embedding, which is a shallow neural network model. After the training is completed, the low-dimensional vector corresponding to each word can be obtained, and each vector can express the semantics of each word to a certain extent. Using this low-dimensional vector for training can achieve better training results.

C. Classification model

The essence of sentiment analysis is to classify the given text. Thus, after obtaining the vectorized representation of the text, we input the vectorized data into classification models. Here, we mainly use the following three methods to classify:

Bi-LSTM Bi-LSTM is a bidirectional Long short-term memory. It is a combination of a forward LSTM and a backward LSTM. The forward LSTM enters text from beginning to end, and you can learn the information prior to a word. The backward LSTM enters from the end to the beginning, and you can learn the information following a word; thus, the two-way LSTM can learn the context information of the text well.

TextCNN Convolutional neural networks (CNN), which usually consist of one or more convolutional layers and fully connected layers and may include pooling layers, have achieved great success in the field of computer vision. Although CNN lacks the ability to process serialized data, it has gradually been introduced into the field of natural language processing due to its parallelism and great advantages in local feature extraction. TextCNN comes from reference [9] and uses a convolution kernel with sizes of 3, 4 and 5 to extract text features for classification.

Bert The publication of Bert by Google [10] has aroused a huge response in the field of natural language processing. Bert can be said to be a trend of pretraining models. By adopting Bert and a fine-tuning model, excellent results have been achieved in 11 natural language processing tasks. This paper also draws on this model and uses Google's pretrained model in Chinese to perform sentiment analysis.

D. Experimental results

We use 100,000 data entries as the training set, 10,000 data entries as the test set, and an F1-score as our evaluation index. The results are shown in Table I . The results of TextCNN and Bi-LSTM are close, and the results of Bert are significantly better than both, further verifying that Bert is powerful.

IV. Fake news detection during epidemic

On social networks, the dissemination of fake information can have a very negative impact on society as a

TABLE I
Sentiment Analysis Results

Model	F1-Score
Bert	0.7159
Bi-LSTM	0.6111
TextCNN	0.6115

whole. Especially during the COVID-19 epidemic, real and fake news continued to appear in social media. Ordinary people were limited by their vision, knowledge and other reasons, and it was difficult to distinguish the authenticity of the news. During epidemic prevention and control, some rumors may cause damage to the stability of society and cause actual loss of personality. To this end, the use of machines to detect the true and fake news in advance and to accurately and timely convey the relevant information about the epidemic to people is vital to winning this war. The fake news detection task using data mining technology usually implements fake news classification through several steps such as data preprocessing, feature engineering, and news classification.

To explain in detail how to use data mining-related methods to complete the fake news detection task during the epidemic, this paper uses the data [1] published by the Internet Fake News Detection During the Epidemic jointly organized by the Beijing Municipal Economic and Information Technology Bureau and the Big Data Expert Committee of the Chinese Computer Society to complete the model training. This dataset comes from the Weibo platform and contains a total of 49,910 news articles. During the experiment, we used 20% of the dataset to test and the remainder for model training. Text is the main carrier of news information, and research on news text helps to effectively identify fake news. Therefore, we mainly detect news text during the detection process. Next, we will explain how data science can serve news detection in epidemic situations from the aspects of preprocessing, feature selection, classification of news, and experimental results.

A. Feature selection

After preprocessing, we selected some available features. To achieve interpretable fake news detection, we expended significant effort in feature selection.

We considered the following features:

News length: According to statistics, the events in our lives are in a normal distribution, although sometimes the distribution is not standard. The same applies to news length. From the statistics, we find that the average length of each news sentence is approximately 120 words. From the distribution of news sentence lengths for real and fake news, we find that the sentence lengths of real and fake news in the current dataset are not much different; thus, the sentence length is not considered to be a feature.

Special symbols: Next, we develop statistics on some special symbols such as “@” and personal pronouns such as “you”, “me”, and “he” in real and fake news.

According to the news statistics, real news has fewer question marks and exclamation marks than fake news. This may be because there is a large amount of exaggerated rhetoric in fake news, which is used to emphasize the writers¡¯ opinions and enhance the authenticity of the news to achieve falsehood. Therefore, fake news tends to use words with special symbols to arouse readers¡¯ specific emotions or to attract the readers¡¯ attention. According to the statistical results , it can be seen that there is no significant difference between the “?” of the real and fake news in the current dataset, so it is not a feature for detection.

In addition, from a psychological perspective, we studied the use of personal pronouns in real and fake news. Deceptive news is mixed with personal sexiness, while real news expresses an objective perspective, often reducing the use of some personal pronouns to increase its objectivity. Because the news itself has a certain grammar, it will be more in line with some norms. According to the statistical results , it can be seen that real and fake news in the Weibo dataset are significantly different; this difference serves as the basis for news detection.

The news data we detect comes from Weibo and there are usually some special symbols in Weibo. For this reason, we counted the frequency of occurrence of some special symbols with special functions, and compared the true and fake news. In social news, for example, the ’#’ symbol is used as a symbol for some topics that people can discuss, using “@” to refer to other users and ‘[]’ to add a symbol to the news as a title. We have counted these three symbols. According to statistics, to achieve digitalization, there will be many media on Weibo. Their news often needs a topic or title to make their content more specific and more in line with the norms. Most fake news will not follow this convention; the content format is more casual. According to the difference between the true and fake news in the current dataset shown by the statistical results, “whether ’#’ is included” and “whether ‘[]’ is included” are taken as a feature of the news detection basis.

In addition, we calculated whether phone numbers appeared in the news. Phone numbers are rarely seen in real news. This is because fraudulent information providers will leave their contact information. In real news, phone numbers are rarely seen because people protect their personal information. According to the statistical results ,there is a significant difference between phone numbers in true and fake news in the current dataset, which is therefore a feature of news detection basis.

Based on the statistics of the above features, we think that these special symbols and personal pronouns are useful for distinguishing between true and fake news;

[1] https://www.datafountain.cn/competitions/422/datasets

thus, we have designed four features for each news type in addition to the text itself. They are called symbol features and are: ‘with or without personal pronouns”, “with or without ’#’”, “with or without ’[]’”, and ’[]’”with or without phone number”.

Text content feature: In this part, we use word2vec as described in section 3.2. Prior to execution, we set the length of the news to 120 (the statistical average of the text length), and then used word2vec to train the d-dimensional word vector. In the end, each news article is represented as a sentence vector (that is, a vector obtained by averaging 120 word vectors) and the symbol features of the current text are combined to obtain a $1 \times (d + 4)$ representation vector for each news article.

B. Classification of news

This part mainly uses binary classification models to divide news into true and fake. In the classification process, we tried several classification models, such as fastText, Bi-LSTM, and TextCNN.

FastText is a text classifier released by Facebook AI Research in 2016. The fastText model architecture is similar to CBOW in word2vec (with similar three layers of input, hidden, and output), but the task is different. CBOW¡¯s task is to predict intermediate words, and fastText is used to predict the category of the sentences. The input of fastText is the word sequence (a paragraph or a sentence) that outputs the probability that the sentence belongs to different categories [11].

Since it is also a classification task, it is slightly different from sentiment analysis in that the former task is classified into three categories and fake news detection is a dichotomy task. Thus, we use these two models here; the details of the model will not be repeated.

C. Experimental results

The F1-score is used as our evaluation index. The results of each classification model for fake news detection are shown in Table II. The results of TextCNN and Bi-LSTM were similar, while the results of fastText were slightly worse. However, during the experiment, it was found that the fastText training only took approximately two minutes while the other two models needed more than eight times the training time. In particular, the parameter learning time of Bi-LSTM was longer due to its more complex network structure.

TABLE II
Performance of Fake News Detection Model

Model	F1-Score
FastText	0.9870
Bi-LSTM	0.9997
TextCNN	0.9998

V. Conclusion

The global epidemic of COVID-19 has had an unpredictable impact on the entire human community. To minimize the impact of the epidemic on human society, this paper explains the role of data science from the perspectives of epidemic development, public opinion control and economic impact. Further, we tried to predict the development of the epidemic situation by using COVID-19 data, using the least squares and particle swarm optimization methods combined with the existing epidemic model SIR. Using data from Weibo, the paper presents a sentiment analysis model with high accuracy and a fake news detection method with strong explanatory ability. This series of work not only explains the important role that data science can play in outbreaks but also provides lessons for future pandemics.

References

[1] B. McCall, “Covid-19 and artificial intelligence: protecting health-care workers and curbing the spread,” The Lancet Digital Health, 2020.

[2] J. T. Wu, K. Leung, and G. M. Leung, “Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study,” The Lancet, vol. 395, no. 10225, pp. 689–697, 2020.

[3] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu et al., “Clinical features of patients infected with 2019 novel coronavirus in wuhan, china,” The Lancet, vol. 395, no. 10223, pp. 497–506, 2020.

[4] S. Zhao, S. S. Musa, Q. Lin, J. Ran, G. Yang, W. Wang, Y. Lou, L. Yang, D. Gao, D. He et al., “Estimating the unreported number of novel coronavirus (2019-ncov) cases in china in the first half of january 2020: a data-driven modelling analysis of the early outbreak,” Journal of clinical medicine, vol. 9, no. 2, p. 388, 2020.

[5] S. J. Fong, G. Li, N. Dey, R. G. Crespo, and E. Herrera-Viedma, “Finding an accurate early forecasting model from small dataset: A case of 2019-ncov novel coronavirus outbreak,” International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp. 132–40, 2020.

[6] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jin, M. Zhang et al., “A machine learning-based model for survival prediction in patients with severe covid-19 infection,” medRxiv, 2020.

[7] W. O. Kermack and A. G. McKendrick, “A contribution to the mathematical theory of epidemics,” Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, vol. 115, no. 772, pp. 700–721, 1927.

[8] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in Proceedings of ICNN’95-International Conference on Neural Networks, vol. 4. IEEE, 1995, pp. 1942–1948.

[9] Y. Kim, “Convolutional neural networks for sentence classification,” arXiv preprint arXiv:1408.5882, 2014.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[11] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” arXiv preprint arXiv:1607.01759, 2016.