Predicting Unemployment Rate Using Literacy Rate with Neural Network

Xiaoxuan Peng

xiaoxuan.blair.peng@gmail.com

School of Social Science, University of California, Irvine, 92697, United States of America

Abstract-The outbreak of the epidemic profoundly affected multiple industries, resulting in a sharp surge in unemployment during its peak. While many sectors have seemingly recovered, unemployment rates in certain regions have not reverted to their pre-epidemic levels. Intriguingly, the literacy rate has remained stable, prompting an examination of its relationship with unemployment. This paper employs a neural network model, incorporating data from 1990 to 2022, that integrates the literacy rate, country code, and year to predict future unemployment trends. However, the model does not accurately predict unemployment based on literacy rates alone, indicating a lack of direct correlation. Consequently, this suggests the necessity of considering a broader range of variables when forecasting future unemployment rates.

Keywords: Literacy, Unemployment, Neural Network, Mathematics

1. Introduction

The forecasting of the unemployment rate has become increasingly crucial for understanding how joblessness affects markets and the broader economy. Governments depend on unemployment rate predictions to budget for public services such as welfare and unemployment benefits. The unemployment rate can also offer insights into analyzing labor markets, formulating social stability programs, and making monetary policy decisions. On a personal level, knowledge of future unemployment trends can aid in career decisions, financial planning, and risk assessment. The demand for highly skilled employees rose by 19% from 1998 to 2018, and proficiency in literacy will continue to be vital for forthcoming job opportunities [1]. According to labor force surveys and the annual reports of the Palestinian Central Bureau of Statistics, unemployment percentages among the age groups of 15-24, 25-34, and 35-44 are closely tied to the literacy rate [2]. A clear and growing association between literacy and employment rates is evident. The relationship between these two factors also fluctuates based on countries and specific years. Both the social environment and a country's historical context can potentially influence the direct link between employment and literacy rates. Thus, accounting for these two confounding factors, this paper seeks to integrate literacy, country, and years into a comprehensive three-dimensional model to predict the likely unemployment rate in the foreseeable future.

To train data for improved accuracy in predicting the unemployment rate, this study utilized neural networks. Artificial neural networks comprise a node layer, which includes one input layer, several hidden layers, and one output layer. Each node (or artificial neuron) is connected

ICEMME 2023, November 17-19, Beijing, People's Republic of China

DOI 10.4108/eai.17-11-2023.2342667

to another node and is associated with specific weights and biases (thresholds) [3]. Thus, neural networks can identify intricate patterns and features, extracting significant insights or even uncovering hidden relationships from vast and complex historical datasets. By training on and discerning patterns in these data, the network can make predictions for future, unseen data. In this paper, I developed a three-dimensional numerical model and employed feed-forward neural networks to process extensive data on literacy, country, and years. This approach aimed to make predictions on future unemployment rates based on new data.

2. Aim and Research Question

Given this context, the objective of this paper is to investigate the relationship between literacy rates and unemployment rates. Initially, I considered incorporating other significant factors such as gender and the economic status of particular countries into the unemployment rate predictions. However, the paper ultimately focused solely on literacy rates as the primary factor. I propose the following research questions, employing a comparative analysis approach with an emphasis on differences and similarities: (1) How has the relationship between literacy and unemployment in a single country evolved over decades? (2) Are there notable differences in unemployment rates between countries or regions with varying literacy levels?

Comparing countries with high literacy rates to those with low rates can help determine if a consistent global trend exists between literacy and unemployment. Such comparisons can spotlight countries that are anomalies (e.g., a nation with high literacy yet high unemployment), which would necessitate further exploration into distinct national circumstances.

3. Data

The data for the present study are derived from the World Bank (data.worldbank.org), which displays the literacy rate for adults (total percentage of people aged 15 and above) [4] and the total unemployment rate (percentage of the total labor force) [5]. Since the data are organized according to the national division of disciplines, "mathematics" at this level refers to mathematical sciences. This encompasses topics such as applied mathematics and mathematics education, in addition to pure mathematics. The literacy data records individuals aged 15 and above who can both read and write, understanding a short, simple statement about their daily life. The unemployment data represents the portion of the labor force that is without work but is available for and seeking employment. To create a three-dimensional input for my neural network, I formatted my data into Country Code, Year, and Literacy Rate, with the aim of predicting the Unemployment Rate. Moreover, I input 80% of the data as the training data and 20% as the testing data. The training and testing data is being divided into 64 bachs, and the epoch (every evaluation period in the case of iteration trainer) is set to 1000. As shown in Figure 1, which is the graph of the world's data of literacy rate over years from 1980 to 2009, the literacy rate data starts at around 67% in 1980, rising to 87% in 2020, with a general trend of increasing without any significant occasional decreases. Meanwhile in Figure 2, which is the graph of the world's data of unemployment rate over years from 1980 to 2009, the unemployment data is less stable, beginning at 4.9% in 1980 and reaching 5.8% of the total labor force by 2022. The graph also depicts drastic changes in the unemployment rate over time.

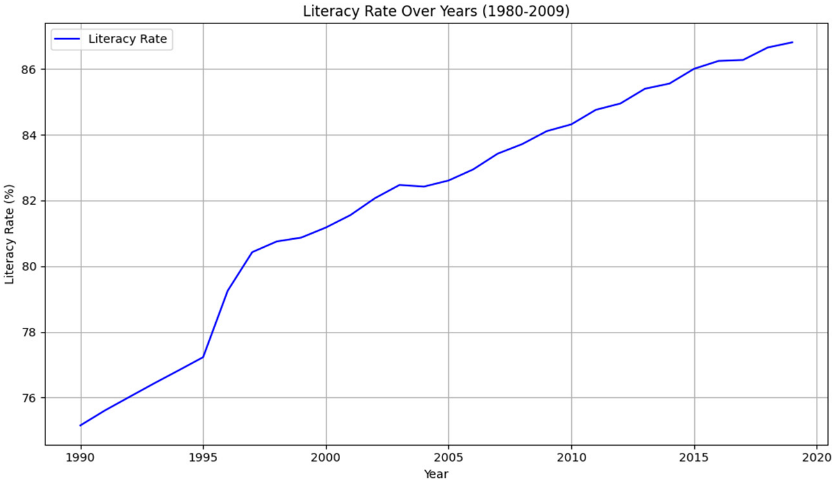

Figure 1. Literacy rate over years(1980-2009). The graph above shows the change of literacy rate over years of the world.

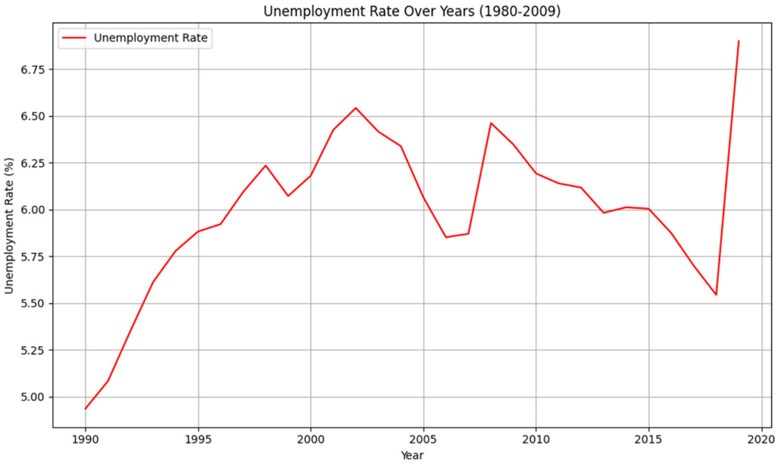

Figure 2. Unemployment rate over years (1980-2009). The graph above shows the change of unemployment rate over years of the world.

4. Model

The activation function transforms the weighted sum of a neuron's input signal into an output signal, aiming to introduce nonlinearity into the neural network. This output signal then serves as an input for the subsequent layer. Mathematically, this can be represented as

$Z = Activation\ Function(\sum(weights \times input + bias))$. Thus, if inputs are the activation can be described as $x_1 + x_2 + x_3 ... x_n$ and the weights are $w_1 + w_2 + w_3 ... w_n$, the activation can be described as $Activation\ Function(x_1w_1 + x_2w_2 + x_3w_3 ... x_nw_n) + bias)$.

The coefficients in the equation are termed weights. The bias is a constant value added to the product of the input and weights, which allows the output value to be skewed more positively or negatively. I have chosen the Rectified Linear Unit (ReLU) function as the activation function, shown in Figure 3, as it avoids vanishing gradient issues and is frequently employed in convolutional neural networks and deep learning models. The ReLU function's output ranges from 0 to positive infinity. Because it lacks exponential components, the ReLU function performs computations swiftly. However, its positive side can extend towards extreme values, leading to computational challenges during the training phase.

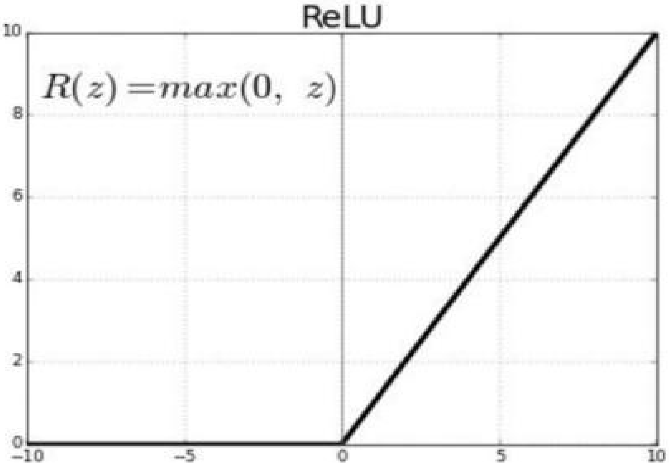

Fig.3. The rectified linear activation function

In order to reduce the error in the neural network, using backpropagation would be an effective way in updating all the weights.

$$ReLU(x) = \{0, if\ x < 0, x, otherwise.\}$$

$$\frac{d}{dx}ReLU(x) = \{0, if\ x < 0, 1, otherwise.\}$$

For the neural network model, I choose to use the feed-forward model, which is most commonly used and considered as the simplest neural network model. Here, the two arbitrary layers input are referred to as Layer (k) indexed by i and Layer (k+1) indexed by j, the feed-forward equations will appear to be as shown below.

$$z_j^{(k+1)} = \Sigma_{\forall i} W_{ij}^{(k)} r_i^{(k)}$$

$$r_j^{(k+1)} = ReLU(z_j^{(k+1)})$$

Following the chain rule, the neuron input before applying ReLu appear to be shown as below:

$$\frac{\partial C}{\partial z_j^{(k+1)}} = \frac{\partial C}{\partial r_j^{(k+1)}} \frac{\partial r_j^{(k+1)}}{\partial z_j^{(k+1)}} = \frac{\partial C}{\partial z_j^{(k+1)}} Step(z_j^{(k+1)})$$

For the training data set, I choose to use the Mean Squared Error function, which is calculating the average of squared differences between the actual and the predicted value. Since the goal is to predict continuous numerical values, the use of MSE function reduces the complexity of the training data set.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} \quad (Yi - \hat{Y}i)^2$$

As for the testing data set, I choose to use the Mean Absolute Error function, which is the average of absolute differences between the actual and the predicted value. MAE calculates the average size of errors in a series of forecasts without taking into account the data's directions.

Unlike MSE, MAE is relatively robust to outliers, which could be more advantageous in performing testing data regression.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \quad |Yi - \hat{Y}i|$$

In order to optimize the predicted values, I use Adam optimizer, which is one of the most efficient algorithm optimization techniques for gradient descent. The 'exponentially weighted average' of the gradients is taken into account in this approach to speed up the gradient descent algorithm. The technique converges quicker towards the minima when averages are used. Taking the momentum function of the algorithm and the Root Mean Square Propagation, the adaptive learning algorithm that tries to improve Adam algorithm, I get

$$M_t = \beta_1 m_{t-1} + (1 - \beta_1)[\frac{\delta L}{\delta W_t}]V_t = \beta_2 V_{t-1} + (1 + \beta_2)[\frac{\delta L}{\delta W_t}]^2$$

In this function, Mt is the aggregate of gradients at time t, Vt is the sum of the square of past gradients, and is the moving average parameter. Based on the aforementioned procedures, Mt and Vt were both initialized as 0. It is noted that because both 1 and 2 1, Mt and Vt are 'biased towards 0'. The 'bias-corrected' Mt and Vt are computed by this optimizer to solve this issue.

Lastly, I add an Exponential Learning Rate Scheduler, which halves the learning rate by the same gamma factor every epoch, to ensure the learning rate keeps getting smaller.

$$Learning\ rate(epoch) = Initial\ Learning\ rate\ \times (1 - \frac{Decay\ Rate}{100})^{epoch}$$

5. Results

As shown in Fig. 4, the loss value graph shows that it has not been reduced to negative log-likelihood and residual sum of squares for classification and regression respectively. With such a high loss value, it is implying that the model is still not trained correctly which means it can not be used for predictions.

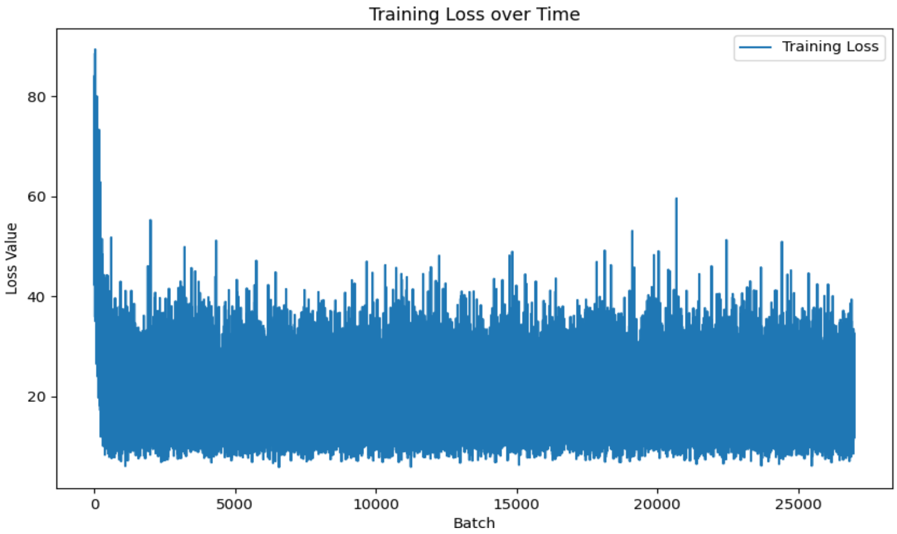

Figure 4. Loss value of each of the bach for 1000 epoch of the complete dataset.

Knowing that the model is not being trained correctly, I graph the regression plot of specific regions and global data.

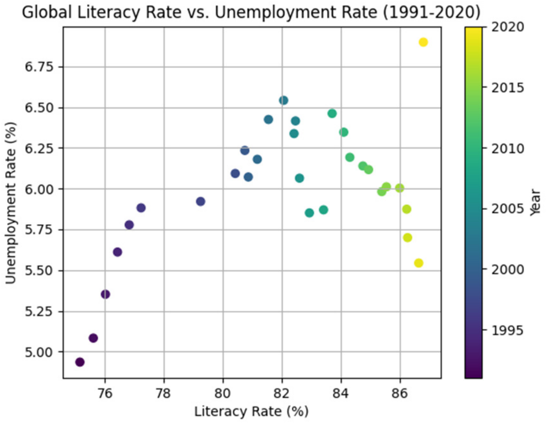

Fig. 5. Global Literacy Rate vs. Unemployment Rate

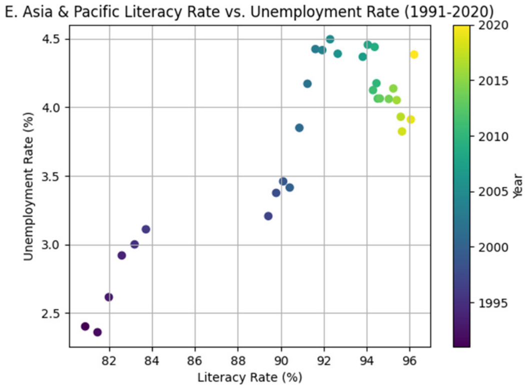

Fig. 6. Asia & Pacific Literacy Rate vs. Unemployment Rate

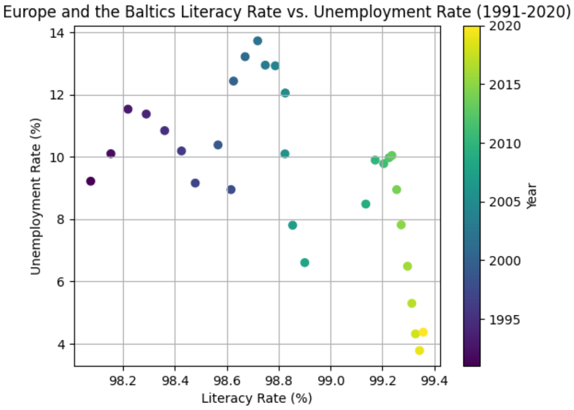

Fig.7. Europe and the Baltics Literacy Rate vs. Unemployment Rate

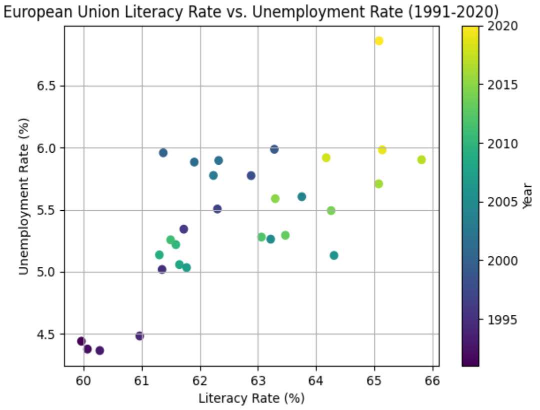

Fig. 8. Asia & Pacific Literacy Rate vs. Unemployment Rate

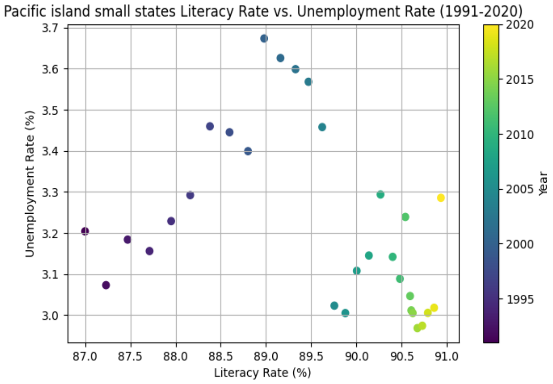

Fig.9. Pacific Island Small States Literacy Rate vs. Unemployment Rate

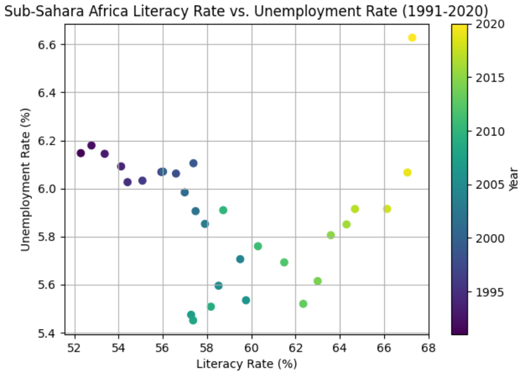

Fig. 10. Sub-Saharan Africa Literacy Rate vs. Unemployment Rate

The image above illustrates that the relationship between literacy rate and unemployment rate varies across regions. Fig.5 shows the global data of literacy rate versus unemployment rate; Fig.6, Fig.7, Fig.8, Fig.9, and Fig.10 relatively demonstrate the relationship of literacy rate versus unemployment rate using the data of Asia & Pacific region, Europe & the Baltics region, European Union countries, Pacific island small states, and Sub-Saharan Africa countries. These relationships are non-linear; an increase in the literacy rate doesn't necessarily lead to a consistent rise or fall in the unemployment rate. For instance, the graph for the East Asia & Pacific region (Fig.6) reveals an intriguing trend: as the literacy rate climbs, the unemployment rate also increases. This contradicts common expectations.

Conversely, in the Europe and the Baltics graph (Fig.7), the relationship is more inconsistent and doesn't lend itself to straightforward predictions. Such fluctuations in unemployment during specific years in these charts might not solely arise from variations in the literacy rate. Other factors, like a country's economic or political climate, could be influencing these changes.

Consequently, a key insight emerges: the reason neural networks struggle to discern patterns in this data and accurately forecast future trends is that the literacy rate, by itself, isn't a reliable predictor of unemployment.

6. Discussion

In relation to my research inquiries, I've discerned that the relationship between literacy and unemployment—whether within a single nation or across multiple countries—cannot be reliably forecasted without taking additional factors into account. Notably, the pronounced disparities in unemployment rates observed among countries or regions with diverse literacy levels cannot be pinpointed solely based on a direct correlation between these two datasets.

To refine the accuracy of unemployment rate predictions within the current model, I can introduce multifaceted dimensions. One prime example is the economic growth (as indicated by GDP) of a nation, which often directly correlates with its job prospects. Technological advancements, if not complemented by an upskilled workforce, can also instigate job displacements. The expansion or contraction of specific industries, as well as the adaptability of the labor market, can sway job opportunities in various directions. Furthermore, demographics, particularly the age distribution of a population, can shed light on employment trajectories since different age groups exhibit distinct employment behaviors. On a broader scale, globalization recalibrates global demand, production, and trade equilibriums, subsequently influencing local employment landscapes. Government-led initiatives, encompassing aspects like healthcare and infrastructural developments, invariably mold productivity and employment metrics.

7. Conclusion

In this study, given the recurring challenges in training the neural network model, there's a need to continually refine and clean the data to make it more suitable for the modeling process. By making an in-depth analysis of the dataset, I might reveal more correlations, or perhaps introduce a feature selection process to determine the most influential factors for future studies. I can also improve the model's predictive capability by expanding the variables. Literacy was the primary factor of interest in this study. However, given its limitations, expanding the study to incorporate other socio-economic variables like GDP, inflation rate, or industrial growth might offer richer insights. To prevent overfitting of the model, there might be a need to explore alternative neural network architectures or introduce regularization techniques like dropout or L2 regularization.

Hyperparameter tuning, using techniques like GridSearch or RandomizedSearch, can also help optimize the model's performance. While deep learning has its merits, the problem might benefit from other machine learning models like Support Vector Machines, Random Forest, or Gradient Boosted Trees. This would be a worthwhile pursuit in subsequent research. In terms of the data, the outliers, such as countries with high literacy but also high unemployment, present an interesting avenue of further exploration. They could be the subject of a more in-depth case study to understand unique national circumstances. Given that the dataset and prediction scope is global, collaborating with international organizations or academic institutions in other countries can offer localized insights.

Once the model is refined and well-tuned to its accuracy, there's potential for building an interactive platform or tool where governments or individuals can input specific data and receive predictions, allowing for proactive decision-making. Unemployment prediction is a dynamic field, with yearly emerging data. Establishing a continuous learning mechanism, where the model is retrained periodically, will ensure its accuracy and relevance with the changing global perspective.

References

[1] OECD (2013), OECD Skills Outlook 2013: First Results from the Survey of Adult Skills, OECD Publishing, Paris, https://doi.org/10.1787/9789264204256-en.

[2] desLibris, Canada’s Top 10 Barriers to Competitiveness in 2016, Canadian Chamber of Commerce. Ottawa, ON, CA. Retrieved from https://canadacommons.ca/artifacts/1217309/canadas-top-10-barriers-to-competitiveness-in-2016/1770407/. CID: 20.500.12592/0sbrdq.

[3] Grossi, E., & Buscema, M. (2007). Introduction to artificial neural networks. European Journal of Gastroenterology &amp; Hepatology, 19(12), 1046–1054. https://doi.org/10.1097/meg.0b013e3282f198a0

[4] World Bank, World Development Indicators. (2022). Unemployment, total (% of total labor force) (modeled ILO estimate). Received from https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS.

[5] World Bank, World Development Indicators. (2022). Unemployment, total (% of total labor force) (modeled ILO estimate). Received from https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS.