Forecasting the United States Unemployment Rate by Using Recurrent Neural Networks with Google Trends Data

Sourav Kundu and Rajshekhar Singhania

Abstract—**We study the problem of obtaining an accurate forecast of the unemployment claims using online search data. The motivation for this study arises from the fact that there is a need for nowcasting or providing a reliable short-term estimate of the unemployment rate. The data regarding initial jobless claims are published by the US Department of labor weekly. To tackle the problem of getting an accurate forecast, we propose the use of the novel Long Short-Term Memory (LSTM) architecture of Recurrent Neural Networks, to predict the unemployment claims (initial jobless claims) using the Google Trends query share for certain keywords. We begin by analyzing the correlation of a large number of keywords belonging to different aspects of the economy with the US initial jobless claims data. We take 15-year weekly data from January 2004 to January 2019 and create two different models for analysis: a Vector Autoregressive Model (VAR) model combining the official unemployment claims series with the search trends for the keyword 'job offers' taken from Google Trends and an LSTM model with only the Google trends time series data for the complete set of identified keywords. Our analysis reveals that the LSTM model outperforms the VAR model by a significant margin in predicting the unemployment claims across different forecast horizons.**

Index Terms—**Unemployment claims, recurrent neural networks, Long Short-Term Memory Network (LSTM), Google trends, Vector Autoregression (VAR), SHAP (Shapley Additive Explanations).**

I. Introduction

In recent years, the advent of big data and advanced computing techniques has made it possible to derive insights and achieve accurate prediction results about the future in almost every possible domain. Numerous studies have used online search data to predict various social phenomenon such as recession, unemployment, inflation, along with other economic indicators [1]–[4]. This has been possible mainly due to the rise of online search companies particularly Google, Yahoo, Bing, etc. Currently, it has been estimated that Google handles almost 92% of the online search queries made throughout the world. The company provides a tool called 'Google Trends', it provides search query trends of various keywords over time. The data is provided in a weekly as well as yearly format. Google Trends has proven to be a reliable source of trend data for online searches and it is being extensively used by researchers around the world for prediction of various macroeconomic trends.

In this study, we focus our attention on using google trends data for prediction of unemployment claims in the United States by leveraging the Long-Short Term Memory (LSTM) architecture of Recurrent Neural Networks. It is well known that the information people provide through their internet search history can provide a good estimate of the economic indicator under consideration, In simple terms, as this study focuses on the unemployment claims as the indicator, hence, an increase in the number of claims would be accompanied by a surge in the number of internet searches focused on finding new job opportunities and reduction of searches for luxurious products and services. A major problem with the use of any type of neural network is the interpretation of the factors affecting the obtained results. To overcome this problem, we use the recently developed Shapley Additive explanations (SHAP) algorithm for identifying the most important keywords in the prediction of the unemployment claims.

The rest of the paper is organized as follows: Section II describes the methodology used for identifying a large number of keywords that may help in the prediction of unemployment claims. In Section III, we provide a brief overview of the two models used for comparison of results, namely, a Vector Autoregressive (VAR) Model and the LSTM model. The results of the two models are discussed in Section IV. Section V provides an overview of the SHAP algorithm and the importance of the various keywords identified in Section II. Section VI gives the conclusion and discusses the importance of using different categories of keywords for the prediction of the unemployment claims.

II. Data

The primary data sources for this study are the Google Trends database by Google Inc. and the Labor force statistics from the current population survey published by the US Bureau of Labor Statistics. We select the data by analyzing the correlation between the Search trends for a keyword with the US Initial Claims (USIC) during the time frame January 2004 to May 2019. Our analysis indicated a very high correlation between the actual US initial jobless claims (USIC) and search trends for keywords related to 5 broad aspects:

i) Unemployment benefits and filing procedure
ii) Lifestyle Indicators
iii) Job search

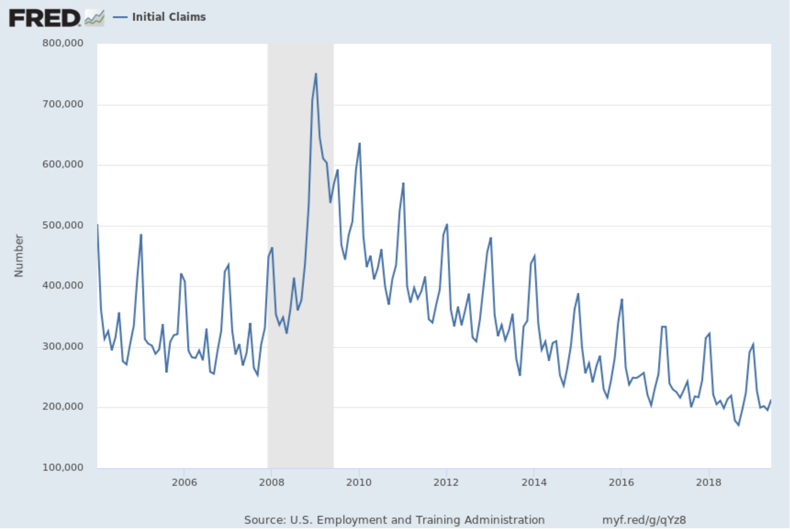
Manuscript received February 15, 2020; revised September 18, 2020.

Sourav Kundu and Rajshekhar Singhania are with the Indian Institute of Technology Kharagpur, India (e-mail: souravwinner@iitkgp.ac.in, raj@iitkgp.ac.in).

doi: 10.18178/ijtef.2020.11.6.679

iv) Welfare and Public Policies
v) Higher Studies

For selecting the data, a script was built in the python programming language, which automatically downloaded the data from the Google server using an API. Suggestions by Google for related keywords for each search were recorded and a repository of 547 keywords was built. It was then divided into two sets: Training data and Testing data. Out of the 185 data points, different data sets were classified into Training data and Testing data according to different prediction horizons. The reason for selecting these time frames is the fact that training data should have enough data points before and after the global recession of 2008 to learn the online search behavior of people according to the different stages of a recession. Figure 1 shows the number of US unemployment claims filed from January 2004 to May 2019.



Source: U.S. Employment and Training Administration myf.red/g/qYz8

Fig. 1. Number of US initial jobless claims between January 2004 and May 2019.

III. Models

A. Vector Autoregressive Model (VAR)

We use a Vector Autoregressive (VAR) model for prediction horizons of 1 month and 3 months. It is not feasible to use a VAR model for prediction horizons of greater periods as the number of parameters increases exponentially with an increase in the prediction horizon. Let Δy_1 and Δy_2 denote the differentiated series of the original unemployment claims and the Google Trend series for the keyword 'job offers' ($\Delta y_{i,t} = y_{i,t} - y_{i,t-1}, i = 1,2..$). The Order of the Autoregressive component of the VAR model was chosen on the basis of the Akaike Information Criteria (AIC) [5]. The VAR (8) model estimated for the variables $\Delta y_{1,t}$ and $\Delta y_{2,t}$ is:

$$\Delta y_{1,t} = a_1 \Delta y_{1,t-1} + a_2 \Delta y_{1,t-2} + a_3 \Delta y_{1,t-3} + b_4 \Delta y_{2,t-4} + \varepsilon_{1,t} \quad (1)$$

$$\Delta y_{2,t} = b_1 \Delta y_{2,t-1} + b_2 \Delta y_{2,t-2} + \varepsilon_{2,t} \quad (2)$$

where $a_i (i = 1, 2, 3)$ and $b_j (j = 1, 2, 4)$ are the parameters that need to be estimated.

In the Eqs. (1) and (2), each error term was determined by using a GARCH (1,1) model with the following equation

$$\varepsilon_t = \mu + \sigma_t z_t \quad (3)$$

where

$$z_t \sim N(0,1) \text{ and } \sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

Here, N (0,1) indicates the standardized normal distribution.

The benefit of using a VAR model lies in the fact that it estimates a relationship between the two time-series under consideration and provides a more accurate forecast of the unknown parameters of the model. As mentioned earlier, a major drawback of the VAR model is the limitation on the prediction horizon due to the increase in complexity of the model with an increase in the number of parameters. It is to be noted that the objective of this study is not to describe the VAR model, it is only used as a benchmark to show the improvement in results obtained with the application of an LSTM model.

B. Long Short-Term Memory Network (LSTM)

Long Short-Term Memory networks were proposed in 1997 [6] as a novel architecture to solve the problem of short-term memory in Recurrent Neural Networks (RNNs). Vanilla RNN models tend to perform quite well when the prediction depends on short-term dependencies. To update the values of weights assigned in a neural network, gradients are used. When the problem requires information from earlier time steps to be carried to later periods, RNNs tend to suffer from the problem of vanishing gradients. The problem of vanishing gradients arises when the values of gradients keep on decreasing during the back-propagation step of the RNNs, thereby reducing their contribution to the learning of the algorithm.

We use the LSTM model to capture the information from the search trends of earlier years in the prediction of the jobless claims. The two major components of LSTM networks are the cell states and the three types of gates used to control information flow in the network. Cell states act as a medium of information transfer throughout the network. The cell state carries relevant information throughout the sequence chain, ensuring the use of information from earlier time steps in the prediction of the output. The three types of cell states are classified as follows:

i) Previous cell state: It is used to describe the information that existed in the memory after the previous time step.

ii) Hidden cell state: This state provides the output of the previous cell.

iii) Input at the current time step: It describes the new information being fed to the network.

Gates are used to filter the relevant information and add it to the cell state. Activation functions, namely, sigmoid activation and hyperbolic tangent activation functions are used in different gates. The three types of gates used are described below and the different notations used are listed in Table I.

TABLE I: NOTATIONS USED TO DESCRIBE THE VARIOUS GATE EQUATIONS IN THE LSTM MODEL

Variable	Meaning
$x_t \in \mathbb{R}^d$	Input vector to the LSTM unit
$f_t \in \mathbb{R}^h$	Activation vector of the forget gate
$i_t \in \mathbb{R}^h$	Activation vector of the input gate
$c_t \in \mathbb{R}^h$	Cell state vector
$o_t \in \mathbb{R}^h$	Activation vector of the output gate
$h_t \in \mathbb{R}^h$	Output vector of the LSTM unit
$W \in \mathbb{R}^{h\times d}, U \in \mathbb{R}^{h\times h}$ and $b \in \mathbb{R}^h$	Weight matrices and bias parameter vectors
σ_g	Sigmoid function
σ_c	Hyperbolic tangent function

C. Forget Gate

The forget gate decides the information that is to be 'forgotten' or removed from the network. In this gate, information from the hidden cell state and the input being fed at the current step is passed through a sigmoid activation function. A value closer to 0 denotes the removal of information from the network.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \qquad (4)$$

Eq. (4) describes the activation function for the forget gate.

D. Input Gate

The input gate is used for updating the cell state. A three-step process is followed:

i) Application of a sigmoid function for regulation of the values that need to be added to the cell state. This is done using the information from the previous hidden state and the current input.

ii) Application of the hyperbolic tangent function (tanh) for the creation of a vector containing the set of all possible values that can be included in the cell state.

iii) The outputs from the first two steps are multiplied with each other to decide which information is to be added to the cell state.

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \qquad (5)$$

Eq. (5) describes the input gate activation function. After the operations on the forget gate and the input gate, the cell state is calculated by the following equation:

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \qquad (6)$$

Here, $\circ$ denotes the Hadamard product and the initial values are $c_0 = 0$ and $h_0 = 0$.

E. Output Gate

The output gate decides the next hidden state. The process followed in this gate can be divided into three steps:

i) Vector creation by application of the tanh activation function to the cell state.

ii) Creation of a regulatory filter by application of the sigmoid function to the hidden state and the current input,

iii) The outputs from the sigmoid function and the tanh function are multiplied to decide the information to be carried by the hidden state. The output from this gate is the new hidden state.

$$h_t = o_t \circ \sigma_c(c_t) \qquad (7)$$

Eq. (7) describes the output vector of the LSTM unit. Figure 2 shows a visualization of a complete LSTM unit with the three gates.

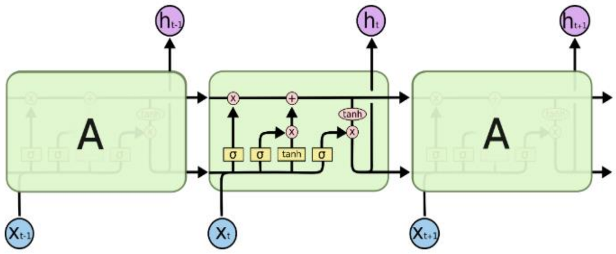


Fig. 2. A repeating module in an LSTM unit.

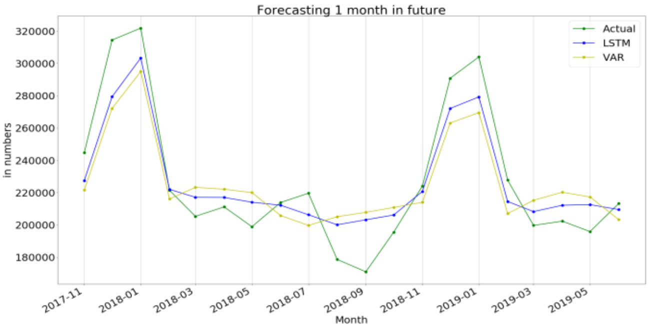


Fig. 3. Predictions made by the LSTM and the VAR model for a prediction horizon of 1 month.

IV. RESULTS

Table II shows the average MAPE obtained for the LSTM model and the VAR model. The LSTM model outperforms the VAR model by almost 3 percentage points. Fig. 3 shows the performance comparison of the VAR model with the LSTM model for a prediction horizon of 1 month. For example, the prediction of jobless claims for November 2017 was made using the data from January 2004 to October 2017, the prediction for December 2017 was made using data from January 2004 to November 2017 and so on. It is observed that the LSTM model outperforms the VAR model in every month. The same phenomenon is observed for a prediction horizon of 3 months as shown in Fig. 4.

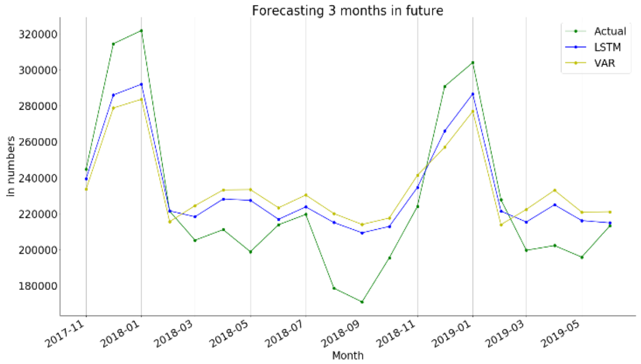


Fig. 4. Predictions made by the LSTM and the VAR model for a prediction horizon of 3 months.

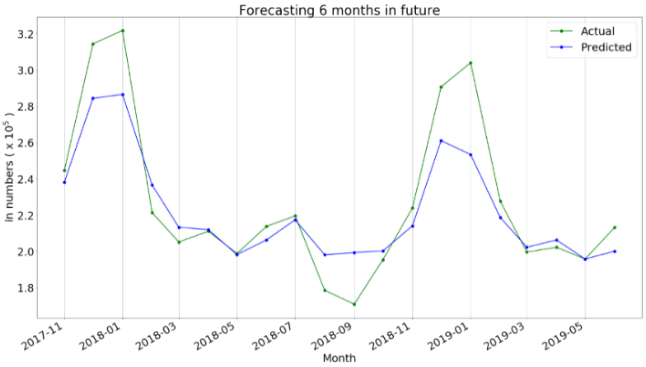


Fig. 5. Predictions made by the LSTM model for a prediction horizon of 6 months.

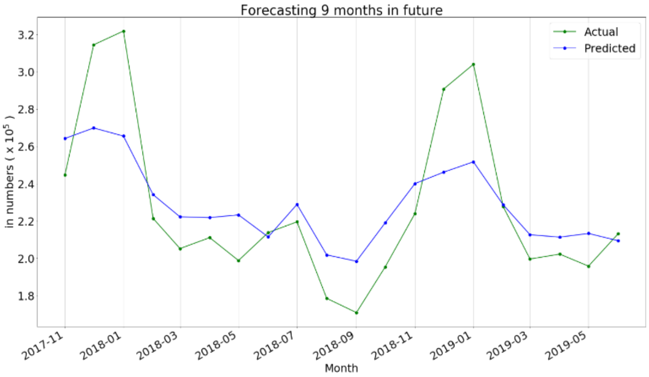


Fig. 6. Predictions made by the LSTM model for a prediction horizon of 9 months.

TABLE II: MAPE VALUES FOR THE LSTM AND THE VAR MODEL

Prediction horizon	**Average MAPE of LSTM**	**Average MAPE of VAR**
h = 1 month	6.24%	9.12%
h = 3 months	7.24%	10.72%

As discussed earlier, the computational complexity of providing forecasts for periods such as 6 months or 12 months is high for a VAR model but performing the same task for an LSTM model is feasible. Fig. 5, Fig. 6, Fig. 7 and Fig. 8 show the predictions made by the LSTM model for horizons of 6 months, 9 months, 12 months, and 24 months respectively. Table III shows the comparison between the average MAPE of the LSTM model and the VAR model for the different horizon periods. The LSTM model performs well with an average MAPE of around 7% for all the prediction horizons.

TABLE III: MAPE VALUES FOR THE LSTM MODEL FOR DIFFERENT PREDICTION HORIZONS

Prediction horizon	Average MAPE of LSTM
h = 6 months	5.69%
h = 9 months	8.97%
h = 12 months	5.23%
h = 24 months	7.86%

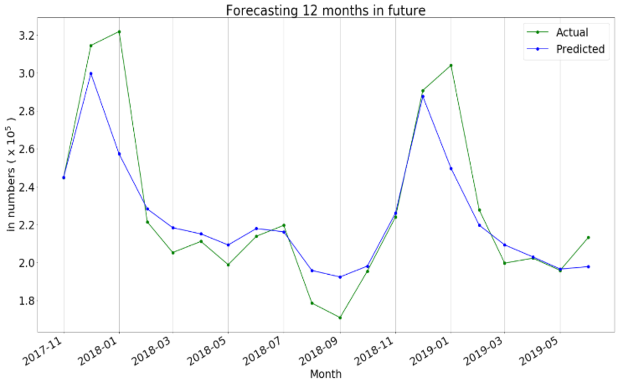


Fig. 7. Predictions made by the LSTM model for a prediction horizon of 12 months.

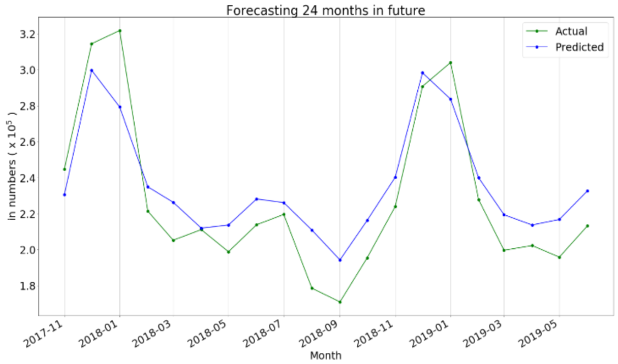


Fig. 8. Predictions made by the LSTM model for a prediction horizon of 24 months.

V. SHAP (SHAPLEY ADDITIVE EXPLANATIONS) ALGORITHM

It is a well-known fact that neural networks are classified as Blackbox algorithms, meaning the interpretability of these algorithms is not possible without an external method. To tackle this challenge, we use the recently developed SHAP algorithm [7]. The basic idea of the SHAP algorithm is to train a linear or interpretable model on top of the original model, leading to an approximation of the original model by the new model. Each feature, or in this case, keyword, is assigned a SHAP value to quantify the contribution of each feature to the overall model. For a detailed mathematical explanation of the algorithm and the calculation of SHAP values, the reader is encouraged to refer to [7].

The SHAP approach was applied to identify the keywords with the highest importance in different forecast horizons. Some keywords such as ‘bars and pubs’, ‘apply for unemployment’, ‘bank rates’, ‘bank jobs’, ‘welfare’, etc. showed high SHAP values across different prediction horizons. Fig. 9-Fig. 13 show the 10 keywords with the highest SHAP values across different forecast horizons.

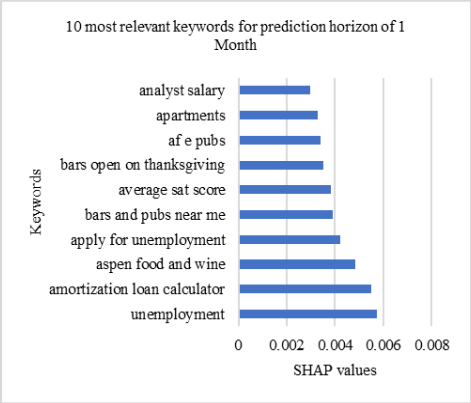


Fig. 9. Top 10 keywords with the highest SHAP values for a prediction horizon of 1 month.

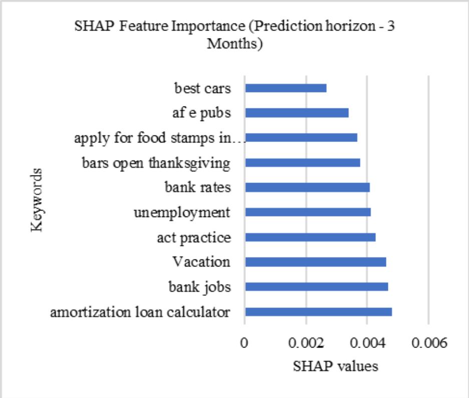


Fig. 10. Top 10 keywords with the highest SHAP values for a prediction horizon of 3 months.

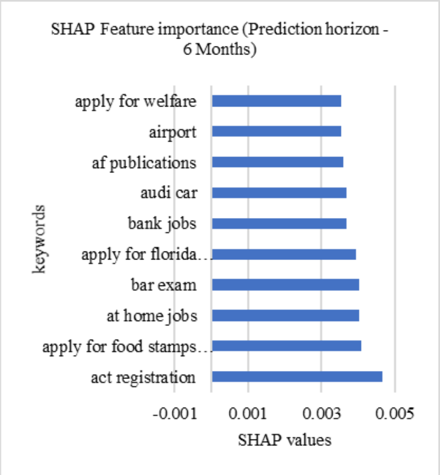


Fig. 11. Top 10 keywords with the highest SHAP values for a prediction horizon of 6 months.

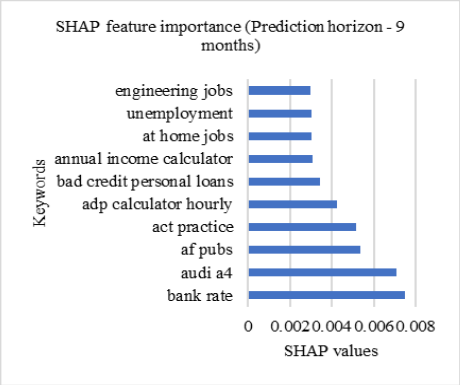


Fig. 12. Top 10 keywords with the highest SHAP values for a prediction horizon of 9 months.

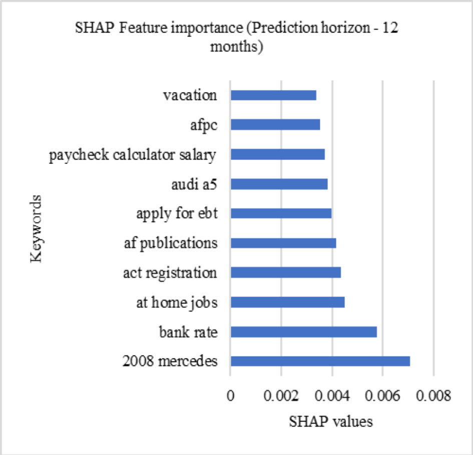


Fig. 13. Top 10 keywords with the highest SHAP values for a prediction horizon of 12 months.

VI. Conclusion

This study focuses on the use of Recurrent Neural Networks, specifically, LSTMs for the prediction of unemployment claims in the US using online search data. The keywords used are chosen from 5 broad categories related to the lifestyle of the people. The results show that the LSTM model applied solely on the online search data is able to perform significantly better than the VAR model applied on the official unemployment claims data and the google trends data for the keyword ‘job offers’. The reason for this can be attributed to the fact that an increase or decrease in unemployment affects the overall lifestyle of the people. The VAR model can only be used to capture the relationship between search trends of a limited number of keywords. The LSTM model is able to overcome this problem and output better results. This implies that the use of keyword related to the overall lifestyle of the people can provide a better forecast of the unemployment claims in the US.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

All authors had equal contribution in the research carried out.

Acknowledgment

The authors would like to thank Prof. Sujoy Bhattacharya of the Vinod Gupta School of Management, Indian Institute of Technology Kharagpur for his valuable guidance and support in publishing this work.

References

[1] A. Naccarato, S. Falorsi, S. Loriga, and A. Pierini, “Combining official and Google trends data to forecast the Italian youth unemployment rate,” *Technol. Forecast. Soc. Change*, 2018.

[2] F. D’Amuri and J. Marcucci, “The predictive power of Google searches in forecasting US unemployment,” *Int. J. Forecast.*, 2017.

[3] J. DiGrazia, “Using internet search data to produce state-level measures: The case of tea party mobilization,” *Sociol. Methods Res.*, 2017.

[4] N. McLaren and R. Shanbhogue, “Using internet search data as economic indicators,” *SSRN Electron. J.*, 2012.

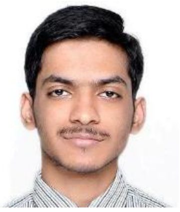
[5] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, 1974.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, 1997.

[7] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.

Sourav Kundu is a final year undergraduate student at the Indian Institute of Technology Kharagpur, India. He is pursuing a dual degree course with bachelor of technology and master of technology in chemical engineering.

He did a summer research internship at the Department of Applied Mathematics and Theoretical Physics, University of Cambridge and a winter internship at the Indian Institute of Sciences, Bangalore in 2018. He also pursued a summer internship at the Citibank in Mumbai, India in 2019.



Rajshekhar Singhania is a final year undergraduate student at the Indian Institute of Technology Kharagpur, India. He is pursuing a dual degree course with bachelor of technology in industrial and systems engineering and master of technology in financial engineering.

He did a summer research internship at the University of Alberta, Canada in 2018, and a winter internship at the Indian School of Business, Hyderabad in 2018. He also pursued a summer internship at the Capital One Financial Corporation in Bengaluru, India in 2019.