ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse148952020.pdf
https://doi.org/10.30534/ijatcse/2020/148952020

Towards Using Recurrent Neural Networks for Predicting Influenza-like Illness: Case Study of Covid-19 in Morocco

Rachida Moulay Taj[1], Zakariyaa Ait El Mouden[2], Abdeslam Jakimi[2], Moha Hajar[1]
[1]RO-I Team, Faculty of Sciences and Techniques Errachidia, Moulay Ismail University, Meknes, Morocco.
r.moulaytaj@edu.umi.ac.ma, moha_hajjar@yahoo.fr
[2]GL-ISI Team, Faculty of Sciences and Techniques Errachidia, Moulay Ismail University, Meknes, Morocco.
mouden.zakariyaa@outlook.com, ajakimi@yahoo.fr

ABSTRACT

The influenza does not affect people's health only, but it is also an essential topic of governments and health care facilities. Early analysis, prediction and response is the most effective control method for flu epidemics. The Artificial Intelligence (AI) scientists are conducting their efforts to develop supervised and unsupervised models in order to analyse epidemics. In this paper, we present the most used Machine Learning (ML) and Deep Learning (DL) models in order to understand Covid-19'sbehaviour by analysing time series data. Among several algorithms of ML, Recurrent Neural Network (RNN) was chosen for tracking this epidemic and predicting its future outbreak. Since the appearance of the first case of COVID-19 in Morocco, the cumulative number of reported infectious cases continues to increase, however this number varies according to the regions of the Kingdom. Also, in this paper, we propose an analysis and prediction model of influenza-like illness Covid-19 by regional distribution. The proposed model is further used to obtain statistical summaries.

Key words: RNN, LSTM, Influenza-like Illness, Covid-19, Coronavirus, Prediction.

1. INTRODUCTION

In late December 2019, a pneumonia with unknown causes was detected in Wuhan, China and then reported to the World Health Organization (WHO) Country Office in China on 31 December 2019 [1].The Centre of disease control experts declared that the pneumonia is a novel coronavirus, it was officially named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) by the International Committee on Taxonomy of Viruses based on phylogenetic analysis. On January, 30, 2020, the WHO declared the outbreak a Public Health Emergency of International Concern, and then it was declared as a pandemic on 11th March, 2020.

In the majority of people who carry the coronavirus, COVID-19 causes a mild respiratory illness similar to influenza. In other individuals, it can lead to a severe respiratory condition that requires hospitalization. People can be without symptoms, or asymptomatic, despite having a SARS-CoV-2 infection, this means that they can still spread the virus to others even though they do not feel unwell. This makes COVID-19 potentially dangerous, as it is highly infectious. However, the most common signs of COVID-19 infection include Fever, Cough and Breathing difficulties. The WHO suggest a preventive measures for Covid-19 include maintaining social distance, washing hands frequently, avoiding touching the mouth, the nose and the face [2].

The first case of Covid-19 was reported in Morocco on 2ndMarch, 2020 for a citizen traveling from Italy to Casablanca city [3] few weeks after the spread of the coronavirus in Italy. On March 19, the government decreed a state of health emergency and the compulsory containment of population has been declared. However, a part of the population continued to work to supply the necessities to the confined population. Every day, the Ministry of Health in Morocco reports infectious individuals and all individuals that have a close contact with them are identified and quarantined.

In 1st June, 2020, the total number of confirmed cases in Morocco was reported to be 7807 infectious and 205 deaths. To date (August, 20, 2020 at 16h00) the number of confirmed cases in Morocco has reached 46,313 including 743 deaths (1.6%) and 31,576 recovered (68.18%), while 13,994 (30.22%) cases are still in hospitalization.

Covid-19 growth data contains temporal information presenting dynamic number of confirmed cases, recovered cases and deaths over time. Due to its characteristics Covid-19 time serie data is complex, non linear, long interval(days, weeks, months) and uncertain, the traditional statisticl technique and analyse become powerless in front of this complixity.

The use of multiple open data (multivariante) as training set fortrainning the model is beneficial on acount of epidemic growth is affected by many factors, meanning that the growth of the number of confirmed cases due to a multiple parameters. In this case, for preliminary, number of

confermed cases, recovered cases, deaths, population, age and population density are used as parameters. There are a relationships between demography parameters like age with the number of confermed cases as proved by a previous study [4].

Machine learning algorithms play a critical role in the analysis of epidemics, espicialy for predictions [5, 6, 23]. With the presence of massive data descibing this epidemic situation, the machine learning techniques help to find patterns so that timely action can be planned to stop the spread of the virus [24]. In this search, the machine learning and deep learning models are used to observe daily behaviour with the prediction of the future cross-country accessibility of Covid-2019 using real-time information provided by the official open data source.These models can forecast the near future and help to reduce the negative effects of the Covid-19.

In this work, first, we propose the use of an LSTM-based approach to learn patterns in COVID-19 data. Since, machine-learning approaches involve allowing the model to automatically learn complex models from data based on the model constructed and fitted the hyper-parameters. This is particularly useful for processing epidemiological data from time series such as that of COVID-19. Then, following validation of the model by parameter adjustment. We will analyze how demographic conditions such as population density can affect the propagation of Covid-19 in different regions of Morocco.

The rest of this paper is organized as follows; Section 2 presents some related works. Section 3 gives a background review of Recurrent Neural Networks (RNN) and Long Short-Term Memory Network (LSTM). Then, section 4describes the main steps of the proposed approach. Finally, Section 5 closes the paper with a conclusion and some perspectives.

2. Related Work

The spread of Covid-19 has led to global research fields to try to understand its different facets. In addition to virology, artificial intelligence is playing a critical role in forecasting the spread of Covid-19 with numerous applications such as computer vision, graph analytics, geographic information systems (SIG), machine/deep learning …etc.

The authors of [7] proposes an LSTM based neural network for real-time influenza-like illness rate (ILI) forecasting in Guangzhou, China, a multi-channel mechanism was added to support the heterogeneity of data collected from different sources and with different formats.

Another work [8] proposed a deep learning model based on LSTM to predict Influenza-like illness using multiple open data sources in Taiwan centers for disease control, the study gave important results for predicting the disease outbreak in Taiwan. . In relation with ILI, the authors of [9] proposes their model DEFSI (Deep Learning Based Epidemic Forecasting with Synthetic Information) which is a short-term based LSTM deep neural network for monitoring ILI in different centers of disease control and prevention especially in USA.

According to [10], existing ILI prediction models are not sufficient, the authors proposed a deep learning based model for prediction and linked their model to a client/server web application developed with Django for visualization, the main advantage of this work is the implementation of an algorithm for the automatic selection of data from the website of the Department of Health and Welfare's Disease Control Agency, the data processing was developed using Python.

Time series are also widely used in prediction; in [11], the authors showed the experimental results of forecasting using *i)* ILI data alone, *ii)* Feature vectors such as week number and *iii)* Time delay embedding, the results showed a high correlation and a low RMSE (root-mean-square error) in comparison the well known state-of-art algorithms. In [12], the authors also proposed a time series model to analyze the trend pattern of the incidence of COVID-19 outbreak in six countries in order to highlight the current epidemiological stage in the World.

A combined model consisting of LSTM and Gated Recurrent Unit (GRU) [13] is applied to prediction of confirmed, negative cases, recovered cases, and death; the model is capable of generating an automated way of confirming, estimating the current position of the pandemic.

In the other hand, numerous works were published in relation with Covid-19 and computer vision; The main objective of those approaches is to classify chest X-ray images as Covid-19 and non Covid-19 using deep neural network [14]. Other works [15, 16] propose multi-class classifiers to predict normal, bacteria and Covid-19 from chest X-ray images.

Recent works are considering graph analytics as a powerful tool for tracking Covid-19 as a complex network, using graph-based machine learning algorithms such as spectral clustering [17] and Graph neural networks [19].

3. Background Review of RNNs and LSTM

Recurrent neural network (RNN) [19] is a neural network model used for processing sequence data. Topically, a neural network contains an input layer, one or several hidden layers and an output layer. The output is controlled by the activation function, and the layers are connected by weight. For making a decision, RNN considers the current input and the output that it has learned from the previous input (see Figure 1).

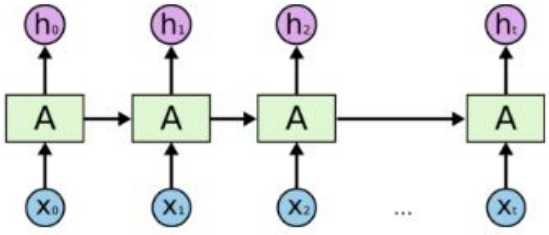

Figure 1: RNN architecture.

RNN is a powerful tool for processing data and predicting time series, since it is capable of managing a sequence[20] by

storing large historical information in its intermediate state, in order to surpass vanish gradient problems.
The gradients transport information used in the RNN parameter update and when the gradient becomes smaller and smaller, the parameter updates become insignificant which means no real learning is done, which leads to the large training time or training does not work at all.

$$\check{C}_t = tanh\ (W_c \times [h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \check{C}_t \quad (4)$$

Output Gate decides what will be the output of the cell based on cell state at instant t as well as the freshest added data.

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (5)$$

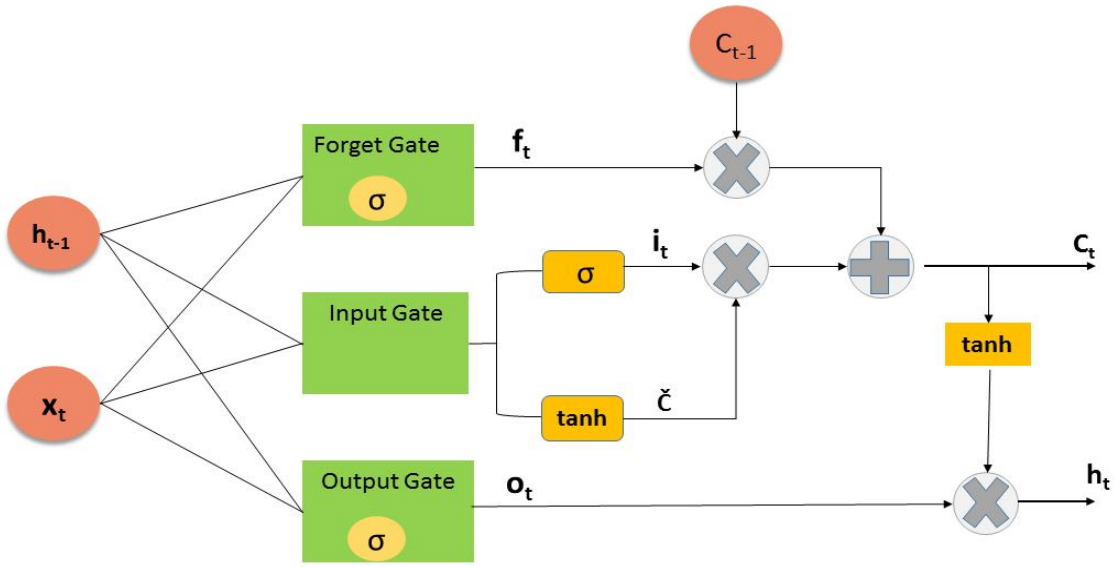

Figure 2: Basic structure of LSTM

To remove the shortcoming of RNN, Hochreiter and Schmidhuber propose Long Short-Term Memory Network (LSTM) [21], it's a special kind of Recurrent Neural Network. Therefore, LSTM is more suitable for processing important events with longer intervals or delay time units in the time series than the RNN. To have the ability of capturing long-term dependencies, LSTM uses memory cell which is specifically designed to store information over long times.
Furthermore, the forget input and output gates, in each memory block can control the flow of information inside memory. The structure of LSTM consists of three gates [22] i.e. input gate, forget gate, and output gate as shown in figure2.
Initially, LSTM initiates with a forgot gate is defined as:

$$f_t = \sigma(W_i \times [h_{t-1}, x_t] + b_f) \quad (1)$$

That uses a sigmoid function combined with previous hidden layer (h_{t-1}) and current input (x_t) to determine the information to forget in the memories in the previous state.
The input gate decides which new data will be added in the cell; first, a sigmoid layer chooses which values will be changed, it is defined as:

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \quad (2)$$

Next, a *tanh* layer proposes a vector of new information that could be added to the state.

The memory stat o_t is the output gate that will be used the determine the quantity of memory output:

$$h_t = o_t * tanh\ (C_t) \quad (6)$$

In the above equations, *tanh* is used to scale the values into range−1 to 1, σ is the activation function which is taken as sigmoid and W are the corresponding weight matrices.

4. PROPOSED APPROACH

In this work, we propose a model that allows us to predict the number of confirmed, recovered and temporarily dead cases in order to analyze the influence of the population density factor on the growth of the number of cases in the different regions of Morocco. The issue is a time series problem because the current number of cases (confirmed, recovered, death) changes in a way that depends on the previous sate, the flowchart of the proposed model is shown in figure3.

4.1 Data collection

The Covid-19 data of different regions of Morocco (12 region) is collected each day at 11 pm from the official Coronavirus Portal of Morocco [3], in this work we will use the chronological data from the appearance of the first case of Covid-19 in Morocco to July 31, 2020 .

To determine the target variables, a statistical method is used to select the variables from the original database (Dataset0), based on regression weight. The selected variables for both the training database and the validation database are presented in table 1.

Table 1: Main parameters

Dataset parameter	*Description*
Confirmed cases	Daily confirmed cases by regional distribution
Recovered cases	Daily recovered cases by regional distribution
Death cases	Daily deaths by regional distribution
Population	The population of each region
Area	Area of each region in km^2
Density	Population per km^2 for each region

4.2 Problem description

We define a Covid-19 prediction model as a supervision machine-learning task; Given a time series contain *N* daily data point for *M* step ahead prediction. In other words, the model is training to be able to predict the number of confirmed cases at time *t + 1* by giving the observation at time *t* and possibly other time steps before that.

The input *X* of supervised ML model is $x_{t-N+1}, \ldots, x_{t-m}$ and the output *Y* is $y_{t-N+1}, y_{t-N+2}, \ldots, y_t$.

The learning and evaluation Model is made up of input elements, which are integrated into the model, and output elements. Note that input and output are both in the form of daily cases, the model has a many to many LSTM architecture (see Figure 4).

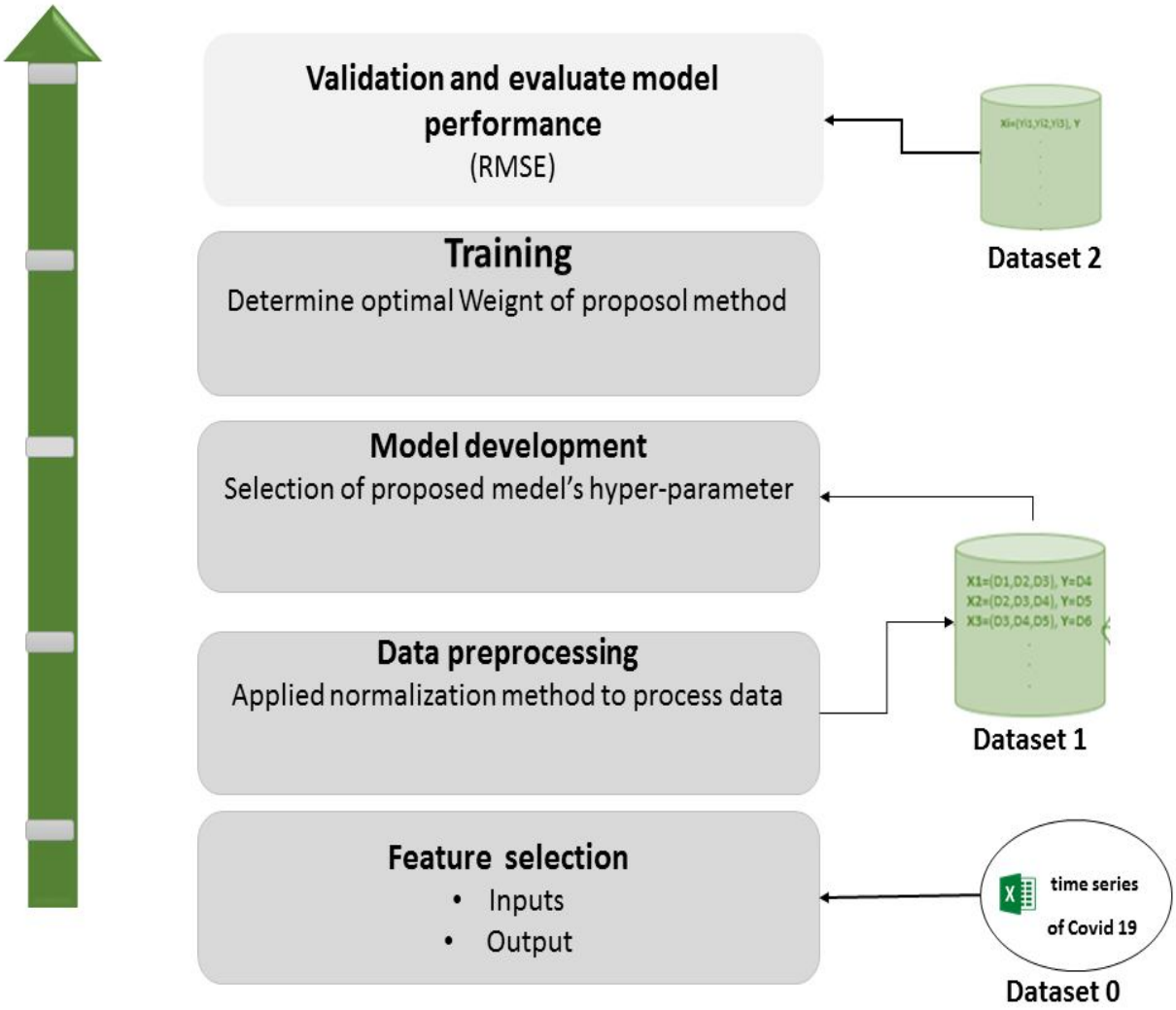

Figure 3: Flowchart of the proposed model.

4.3 Data processing

The preparation of the data before training aims at normalizing the data in order to make it more consistent as LSTM is very sensitive to normalization, especially for

capturing time series data. Before applying the LSTM method to predict the number of cases, we apply min-max scaling of all data with the maximum and minimum values of the data set. The normalization allows us to transform the data to the same scale and avoid bias in training and validation steps.

$$X_{normalized} = \frac{X - X_{minimum}}{X_{maximum} - X_{minimum}}$$

4.4 Training and Performance evaluation

In the training phase of the model, 80% of data are used (dataset1) and the rest 20% (dataset2) for validation and evaluation of the proposal. We choose 8 states as training, 2 states as validation and 2 states as test data.

The input data is first placed in the LSTM layer, in the input gate of the LSTM layer. Moreover, recompose the input data and decide which input data is important. As well, the LSTM layer can retain previous information that can help improve the model's ability to learn from time series data. In phase of training of our model 88 sequence input is split into 1st to 64th day as input neural and 65thto 100th day as label.

To measure the performance of the model we used Mean Absolute Percent Error (MAPE) to measure the difference between the model output (prediction) and the actual data (observation).

$$MAPE = 100 * \sum \frac{\left[\frac{obs_t - pred_t}{obs_t}\right]}{n}$$

where *obs* is the observation value, *pred* is the forecast value and *n* is total number of observation. The validation of this model we offer the best architecture of LSTM as number of layer and hidden layer and hyper parameters.

5. CONCLUSION

After the hard situation of Covid-19 in the last months, no one can deny the critical role of scientific research in the present. Covid-19 has launched many researches in different fields, starting from biology and bio-informatics to artificial intelligence and geographic information systems in order to find solution and control the fast outbreak on the coronavirus. This paper proposed the use of machine learning and deep learning models for epidemic prediction and analysis using LSTM model to predict Covid-19 pandemic growth in Morocco. This prediction model can be useful for making important decisions in achieving a faster response and control of the situation.

Future extensions of this work are in the progress to implement this model using Tensorflow environment, and to analyze the relationship between the different parameters of the dataset. Another extension is a Graph-based convolutional neural network model to combine between the prediction task and the complex networks data structure.

REFERENCES

1. C. Huang et al., **Articles Clinical features of patients infected with 2019 novel coronavirus in Wuhan** , China, pp. 497–506, 2020.
2. Coronavirus disease 2019, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. visited on April, 4, 2020.
3. T. M. of H. of Morocco, **Communiqué N°10: Morocco announces 1st COVID-19 case**, 2020. [Online]. Available: www.covidmaroc.ma. visited on April, 4, 2020.
4. J. B. Dowd et al., **Demographic science aids in understanding the spread and fatality rates of COVID-19**, *Proc. Natl. Acad. Sci.*, vol. 117, no. 18, pp. 9696–9698, 2020.
5. G. Kalipe, V. Gautham, and R. K. Behera, **Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis**, in *2018 International Conference on Information Technology (ICIT)*, 2018, pp. 33–38.
6. R. Singh and A. Bhatia, **Sentiment analysis using machine learning techniques to predict outbreaks and epidemics**, *Int. J. Adv. Sci. Res.*, no. May, pp. 19–24, 2018.
7. C. T. Yang et al., **Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources**, *J. Supercomput.*, no. 123456789, 2020.
8. X. Zhu et al., **Attention-based recurrent neural network for influenza epidemic prediction**, *BMC Bioinformatics*, vol. 20, no. Suppl 18, pp. 1–10, 2019.
9. L. Wang, J. Chen, and M. Marathe, **TDEFSI: Theory-guided Deep Learning-based Epidemic Forecasting with Synthetic Information**, *ACM Trans. Spat. Algorithms Syst.*, vol. 6, no. 3, 2020.
10. C. T. Yang, L. Y. Lin, Y. T. Tsan, P. Y. Liu, and W. C. Chan, **The Implementation of a Real-time Monitoring and Prediction System of PM2.5 and Influenza-Like Illness Using Deep Learning**, *J. Internet Technol.*, vol. 20, no. 7, pp. 2237–2245, 2019.
11. N. Wu, B. Green, X. Ben, and S. O'Banion, **Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case**, *arXiv preprint arXiv:2001.08317*, 2020.
12. S. Deb and M. Majumdar, **A time series method to analyze incidence pattern and estimate reproduction number of COVID-19**, *arXiv preprint arXiv:2003.10655*, pp. 1–14, 2020.
13. S. Dutta and S. K. Bandyopadhyay, **Machine Learning Approach for Confirmation of COVID-19 Cases : Positive, Negative, Death and Release**, MedRvix no. Cdc, 2020.
14. J. Zhang et al., **Viral Pneumonia Screening on Chest X-ray Images Using Confidence-Aware Anomaly Detection**, no. July, arXiv: 2003.12338, pp. 1–11, 2020.
15. K. El Asnaoui and Y. Chawki, **Using X-ray images and deep learning for automated detection of coronavirus disease**, *J. Biomol. Struct. Dyn.*, vol. 0, no. 0, pp. 1–12, 2020.

16. L. Wang and A. Wong, **COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images**, pp. 1–12, *arXiv preprint arXiv:2003.09871*, 2020.
17. Z. A. El Mouden, A. Jakimi, and M. Hajar. **An application of spectral clustering approach to detect communities in data modeled by graphs**. In *the Proceedings of the 2nd International Conference on Networking, Information Systems & Security*. ACM International Conference Proceeding Series, Volume Part F148154, Article 4, 2019.
18. Q. Liu, M. Nickel, D. Kiela, **Hyperbolic graph convolutional neural networks**, *arXiv:1910.12892*, 2019.
19. M. Dinarelli and I. Tellier, **A study of Recurrent Neural Networks for Sequence Labelling**, In *Actes de la conférence conjointe JEP-TALN-RECITAL*, Vol. 2: TALN (Articles longs), pp. 98-111, 2016.
20. J. T. Connor, R. D. Martin, L. E. Atlas, and M. Ieee, **Recurrent neural networks and robust time series prediction - Neural Networks**, *IEEE Transactions on neural networks*, vol. 5, no. 2, pp. 240–254, 1994.
21. S. Hochreiter and J. Schmidhuber, **Long short-term memory**, *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
22. X. Li et al., **Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation**, *Environ. Pollut.*, vol. 231, pp. 997–1004, 2017.
23. S. Mifrah, E. H. Benlahmar, **Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9(4), pp. 5756 – 5761, 2020.
24. K. Dheeraj, **Analysing COVID-19 News Impact on Social Media Aggregation**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9(3), pp. 2848 – 2855, 2020.