Health Informatics and Biomedical Engineering Applications, Vol. 78, 2023, 88–94
https://doi.org/10.54941/ahfe1003459

VIS-NLP: Vaccination Inventory System for Justified User Using Natural Language Processing

Vu Minh Phuc, Satyam Mishra, Oni Damilola Igbagbo, and Le Trung Thanh

International School, Vietnam National University, Hanoi (VNU-IS), Vietnam

ABSTRACT

In the healthcare industry, especially the Covid-19 pandemic in 2020, produced huge problems with isolate patient and patient heath. Thus, created large amount of data that has been generated every day for the patient heath, in this case is to justify the vaccination of users from social network Twitter. Processing such large volume of the data involves high computation overhead. Good health and well-being; to ensure healthy lives and promote well-being for all at all ages is United Nations 3rd Sustainable Development Goal and we want to align our study with it as well. It is crucial to create an application that is beneficial for humanity health. When we get large datasets from pandemics like Covid-19, for large scale datasets, we presented a solution to verify the user if they are vaccinated or not vaccinated by using Natural Language Processing methods to build an accuracy result, we tried to reduce the computation overhead by storing the data in distributed environment. After processing data, training the data, used pad_sequences, Keras, NLP to build the model. Through multiple epochs we have got an accuracy towards 90 to 91% (which is closer to state-of-the-art methods i.e., 95%). And since our accuracy is higher, we can further utilize it to increase for higher number of epochs. We hope scientists can further develop it and use it in real world applications so that more precious human being lives can be saved. By implementation of its successful results, it also aligns with one of the United Nations Sustainable Development Goals i.e., 3rd: Good Health and Well-Being.

Keywords: Natural language processing, Epochs, UNSDG, Jupyter, Keras, Tokenizer, Pad_sequences

INTRODUCTION

Sustainable development depends on encouraging well-being at all ages and ensuring healthy lives. The COVID-19 (Nambiar et al., 2013) virus, which is causing billions of people's lives to be turned upside down and spreading human agony around the world, is currently causing a global health crisis unlike any other. The covid pandemic has been causing large amount of struggling in livelihoods, health economic, education, travel, etc. (*Impact of COVID-19 on People's Livelihoods, Their Health and Our Food Systems*, n.d.) And during also after the pandemic, people were tend to use social media frequently to update the situation at their location all around the world (in this case authors gather information from Twitter). Also, humanity should get the vaccinations in order to get back to their lives, but they also need to aware of knowing who get affected and who get vaccinated. In this case, we used database of a medical system, so there will be a lot of details regarding to a person such as person's age can affect the delivery of the vaccine (elder persons may require higher vaccination or in a fast rate, while compared to youngsters). There are some works being done for image processing as well using Canny Edge Detection Algorithm (Mishra & Thanh, 2022). Also, some neural network approach training for object detection etc. for driverless vehicles to take covid patients (Mishra et al., 2022). On investigating our research domain's algorithm's performance on large dataset used in this work, the experimental results show that the Natural Language Processing is better than other algorithms in terms of accuracy, execution time and error rate (Abu Ghosh & Maghari, 2017). We used Jupyter notebook to run the program because each time authors run the code; our dataset will be going to update it will be easier to get acquainted with current situation of our dataset. Good health and well-being; to ensure healthy lives and promote well-being for all at all ages is United Nations 3rd Sustainable Development Goal and we want to align our study with it as well. It is crucial to create an application that is beneficial for humanity health. When we get large datasets from pandemics like Covid-19, for large scale datasets, we presented a solution to verify the user if they are vaccinated or not vaccinated by using Natural Language Processing methods to build an accuracy result, we tried to reduce the computation overhead by storing the data in distributed environment. After processing data, training the data, used pad_sequences, Keras (Team, n.d.), NLP ("What Is NLP?," n.d.) to build the model. Through multiple epochs we have got an accuracy towards 90 to 91% (which is closer to state-of-the-art methods i.e., 95%). And since our accuracy is higher, we can further utilize it to increase for higher number of epochs. We hope scientists can further develop it and use it in real world applications so that more precious human being lives can be saved. By implementation of its successful results, it also aligns with one of the United Nations Sustainable Development Goals i.e., 3rd: Good Health And Well-Being (*THE 17 GOALS | Sustainable Development*, n.d.).

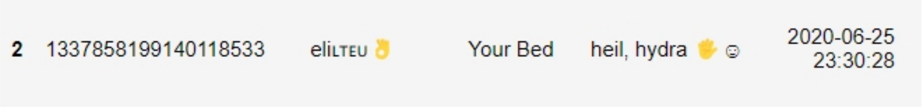
METHODOLOGY

In order to get the successful results, we created and followed these steps:

- Data Explanation
- Processing Data
- Data Visualization
- Building Model

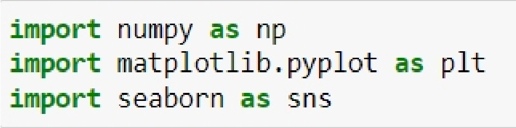
Data Explanation

In order to build a model with an accuracy, some of the data is not helpful for prediction such as: username, user id, user followers, user create account time etc. Figure 1 below shows a little sample from dataset.



2	1337858199140118533	eliʟᴛᴇᴜ 👌	Your Bed	heil, hydra 🖐 ☺	2020-06-25 23:30:28

Figure 1: Dataset (Vaccination data on Data.World | 28 datasets available, n.d.).



```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Figure 2: Libraries used for cleaning data, processing data, data visualization.

We need to find those null values, delete unhelpful variables/data columns, and add value to it (in dataset), which is helpful for building model to make the better prediction, then plotted it into graph to get the visualization. We used Pandas, Numpy, Matplotlib, Seaborn library as we can see in figure 2.

Processing Data

Basically, the task is to create a verified user for a vaccination or not. We eliminated the data such as user followers, username, user followers due to the purpose of building a model for prediction. Eliminate the particular values (cleaning data) that are not required for the prediction of a model: UserID, Username, User's invalid description, User followers, User friends, User favourites, User's created, Date, User favourites, etc.

Add particular values in null value due to build the model with better prediction. Data visualization, create data scatter to check the outliers, which of those values are biased or overshoots from the original lines. Preprocessing values by LabelEncoder, assign the values with some fixed values or it will assign some randoms values but also fixed each value (Suppose a name of a user is repeated 4 times, then it will assign that value 4 times but it can be any value).

Check the user verified by using sns.countplot(x='user_verified',data=data). Create the plotting data to build model for a better and accurate vaccination process, show the distributions of user location and their follower on Twitter as we can see in the figure 3 earlier.

Data Visualization

In this step, we separated the verified users and non-verified users and calculated the percentage of them in this figure 4.

It can be seen that after the calculation, the percentage of verified users approximate to 14.13% less than non-verified users which is approximate to 85.86%.

Building Model

We use get the TensorFlow from Keras, Tokenizer library, Pad_sequences library to build the model prediction as we can see in figure 5.

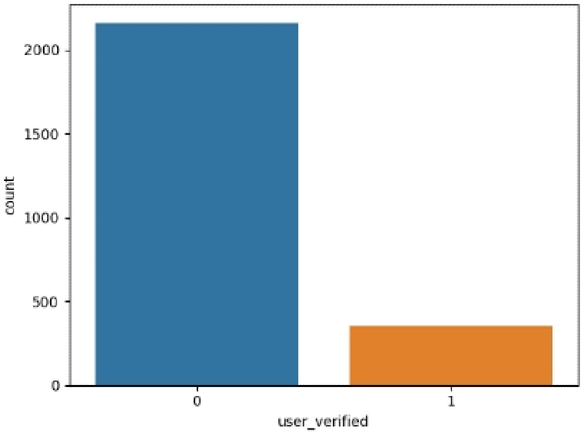
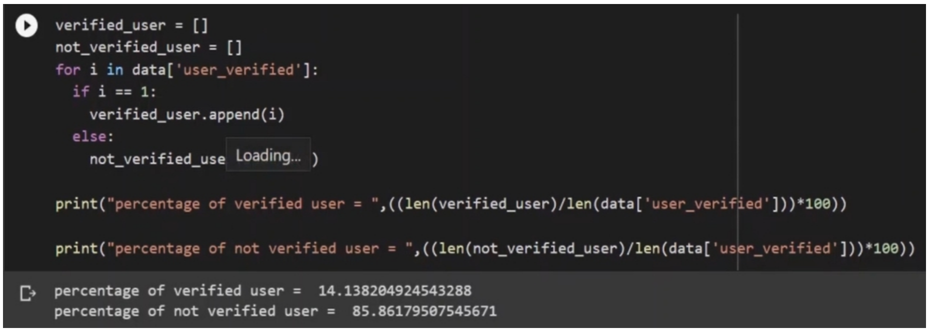


Figure 3: Graph of verified users and unverified users.



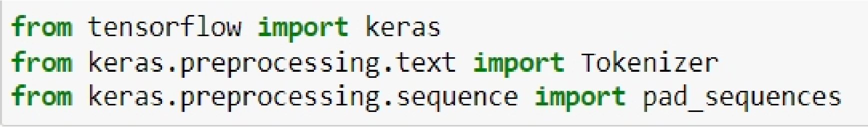
```
verified_user = []
not_verified_user = []
for i in data['user_verified']:
  if i == 1:
    verified_user.append(i)
  else:
    not_verified_use Loading... )

print("percentage of verified user = ",((len(verified_user)/len(data['user_verified']))*100))

print("percentage of not verified user = ",((len(not_verified_user)/len(data['user_verified']))*100))
```

```
percentage of verified user =  14.138204924543288
percentage of not verified user =  85.86179507545671
```

Figure 4: Percentage of verified user and unverified user.



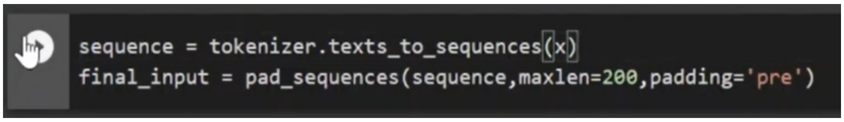
```
from tensorflow import keras
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
```

Figure 5: Libraries used for building model.

In *keras.preprocessing.text* from Tokenizer library, it will convert each word of a particular sentence to a particular numerical value and then add it in, so each time that value will be repeated, it will assign it to the value that is given by itself.

In *lera.preprocessing.sequence* from pad_sequences library, shows that in the particular sequence, whether you are applying an NLP project to any data, there might be some variations in lengths of different type sentences. The job of NLP does is using paired sequences, it needs similar line of sentences, itself add zeros to the end or end of that particular sentences, then send to the stack. Thus, the length of each sentence can become similar.

After pre-processing the data and created a biased dataset, we started to split the data and perform Natural Language Processing.



```
sequence = tokenizer.texts_to_sequences(x)
final_input = pad_sequences(sequence,maxlen=200,padding='pre')
```

Figure 6: Use pad_sequences with the max length as 200 and padding is three.

Set input as text into X, output as user verified into Y in order to calculate the percentage of verified users and non-verified users. pad_sequences libraries will create a stack in which it will convert each word of a particular sentence to a particular numerical value and then add it so that each particular time that value will be repeated, this library will assign it to the value that is given by itself.

According to the organizer, we choose to use participants, because participants in the particular fixed sequence, whether people are applying an NLP project to any data, there must be some variations in lengths of different types of sentences. Basically, an NLP does is in the using paired sequences, it needs similar line of sentences, it adds zeros to the end or to the end of that particular sentence you can set to the stack so that the length of each sentence becomes similar. So that converting the text into sequences and then saving it into a variable name as sequence.

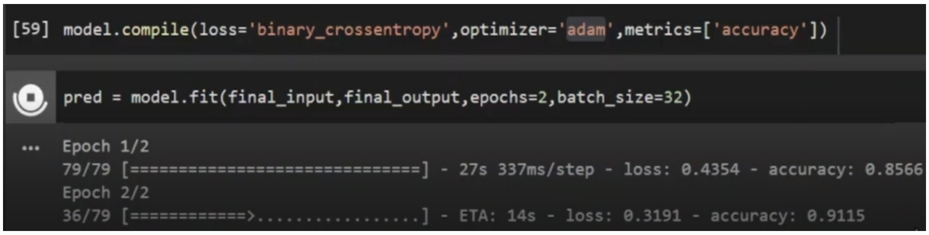
Applying the power sequence to the sequence, we gave the max length of the system as 200 and the padding is three as we can see in figure 6. That means the zeros will be added to the beginning of the sentences.

EXPERIMENTAL RESULTS

We used Optimizer as Adam and optimizer is used to decrease the laws biometrics of accuracy. At last, we fit the model and get the result, the epochs (number of persons) set for 2, and the test size is 32, achieve the successfully results. And if our accuracy is higher, we can then estimate that it will only increase for higher number of epochs.

So, the image shows on the end of towards our first epoch, it gave an accuracy somewhat around 86% proved that is a quite good value, and in the second epoch, you can see it has started the accuracy towards 90 to 91%.

Table 1 shows the comparison of our research finding with the state-of-the-art methods.



```
[59] model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])

pred = model.fit(final_input,final_output,epochs=2,batch_size=32)

...  Epoch 1/2
     79/79 [==============================] - 27s 337ms/step - loss: 0.4354 - accuracy: 0.8566
     Epoch 2/2
     36/79 [============>.................] - ETA: 14s - loss: 0.3191 - accuracy: 0.9115
```

Figure 7: Results with two epochs and 32 test size.

Table 1. Comparison of our research findings with state-of-the-art methods.

Our Findings	STATE OF THE ART (SOTA)
Accuracy of 91%	Accuracy of 95%
The use of Natural Language Processing (NLP) in a vaccination inventory system with an accuracy of 91% through multiple epochs is a good performance compared to traditional methods. We used multiple epochs, but it is proven that accuracy can be increased by training with higher epochs according to our research.	The state-of-the-art models also have a more robust architecture and use more advanced techniques, such as transfer learning and ensemble methods. The use of NLP in a vaccination inventory system can still provide significant improvements and make it easier for justified users to access the information they need (with high number of epochs as per findings of the research)

CONCLUSION

To conclude all, we developed a more accurate prediction based on scientific studies, the research's content has been centered on Keras, pad_consequence, Numpy, Pandas, Tokenizer, methods to increase vaccine users' understanding of the importance of vaccinations. The study has produced several novel findings that may be summed up as follows:

- The study has shown that a large amount of data can be processed and made into predictions precisely by this methodology, especially in social network through Covid-19 Pandemic.
- Using Jupyter notebook to code, build solutions using Natural Language Processing to make precise predictions shown efficiency, suites the research has provided an overview of how to be aware of users who get immunized or not from the society (in this case, Twitter).

After many months of testing, the research has found a solution to the issues raised by the initial objective and implemented the solution successfully. Through multiple epochs we have got an accuracy towards 90 to 91 % (which is closer to state-of-the-art methods i.e., 95%). And since our accuracy is higher, we can further utilize it to increase for higher number of epochs. We hope scientists can further develop it and use it in real world applications so that more precious human being lives can be saved. By implementation of its successful results, it also aligns with one of the United Nations Sustainable Development Goals i.e., 3rd: Good Health And Well-Being.

ACKNOWLEDGMENT

Foremost, we would like to express our sincere gratitude to our advisor Dr. Le Trung Thanh for the continuous support for our group's study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped us in all the time of research and writing of this research. We could not have imagined having a better advisor and mentor for our work.

REFERENCES

Abu Ghosh, M. M., & Maghari, A. Y. (2017). A Comparative Study on Handwriting Digit Recognition Using Neural Networks. *2017 International Conference on Promising Electronic Technologies (ICPET)*, 77–81. https://doi.org/10.1109/ICPET.2017.20

Impact of COVID-19 on people's livelihoods, their health and our food systems. (n.d.). Retrieved February 1, 2023, from https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems

Mishra, S., Minh, C. S., Thi Chuc, H., Long, T. V., & Nguyen, T. T. (2022). Automated Robot (Car) using Artificial Intelligence. *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 319–324. https://doi.org/10.1109/ISMODE53584.2022.9743130

Mishra, S., & Thanh, L. T. (2022). SATMeas - Object Detection and Measurement: Canny Edge Detection Algorithm. In X. Pan, T. Jin, & L.-J. Zhang (Eds.), *Artificial Intelligence and Mobile Services – AIMS 2022* (pp. 91–101). Springer International Publishing. https://doi.org/10.1007/978-3-031-23504-7_7

Nambiar, R., Bhardwaj, R., Sethi, A., & Vargheese, R. (2013). A look at challenges and opportunities of Big Data analytics in healthcare. *2013 IEEE International Conference on Big Data*, 17–22. https://doi.org/10.1109/BigData.2013.6691753

Team, K. (n.d.). *Keras documentation: Keras API reference*. Retrieved February 1, 2023, from https://keras.io/api/

THE 17 GOALS | Sustainable Development. (n.d.). Retrieved December 10, 2022, from https://sdgs.un.org/goals

Vaccination data on data.world | 28 datasets available. (n.d.). Data. World. Retrieved February 1, 2023, from https://data.world/datasets/vaccination

What is NLP? (n.d.). *NLP Training*. Retrieved February 1, 2023, from https://www.nlp.com/what-is-nlp/