FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 4 No. 2, June, 2020, pp 704 – 711
DOI: https://doi.org/10.33003/fjs-2020-0402-331

A MULTINOMIAL NAÏVE BAYES DECISION SUPPORT SYSTEM FOR COVID-19 DETECTION

[*1]Jamilu Awwalu, [2]Nana Aisha Umar, [3]Mubaraka Sani Ibrahim, [1]Ogwueleka Francisca Nonyelum

[1]Department of Computer Science, Faculty of Military Science and Interdisciplinary Studies, Nigerian Defence Academy, Kaduna State, Nigeria
[2]Department of Information Technology, Faculty of Computing, Federal University Dutse, Jigawa State, Nigeria
[3]Department of Computer Science, Faculty of Computing, Baze University, Abuja, Nigeria

[*]Corresponding Author's Email: awachi.jami@nda.edu.ng

ABSTRACT

Coronavirus disease 2019 termed COVID-19 is a highly infectious and pathogenic illness caused by severe acute respiratory syndrome. Symptoms of COVID-19 range from mild to severe, in some cases leading to death. Early detection could help to monitor progression of the disease, mitigate spread of the disease and possibly reduce mortality rate. Computer-aided diagnosis systems are designed to complement health care systems and assist in the early detection of diseases. Currently, as it is not possible to test all citizens especially in developing countries with very large populations due to financial constraints and the standard of their healthcare facilities, the problem of identifying suspected cases and deciding laboratory test priority among citizens is evident and more pressing. Therefore in this study, we introduce an interactive Artificial Intelligent web system using the Multinomial Naïve Bayes algorithm with the aim of detecting warning COVID-19 symptoms and to provide fitting suggestions. Furthermore, the study also evaluates the performance of the Multinomial Naïve Bayes based on the different holdout approaches experimented. The experimental results are promising as the Multinomial Naïve Bayes is shown to achieve high accuracy detection thus providing a reliable method to identify warning symptoms of COVID-19.

Keywords: COVID-19, Multinomial Naïve Bayes, Detection, Decision Support System, Website.

INTRODUCTION

The 2019 Coronavirus (COVID-19) pandemic is an infectious respiratory illness caused by severe acute respiratory syndrome (SARS-COV-2). Symptoms produced by COVID-19 range from mild to severe and in some cases results in death. Common symptoms are fever, cough and tiredness. Other symptoms include shortness of breath, pneumonia, acute respiratory distress syndrome (ARDS), multi-organ failure, heart problems, acute kidney injury and blood clots (Mayo Clinic, 2020). The infectious nature of COVID-19 has caused widespread outbreak across the globe leading to large volume of confirmed cases and high number of mortality rate. As of June 10th 2020, 12,323,504 COVID-19 cases and 555,997 deaths were confirmed (John Hopkins University and Medicine, 2020). The high number of cases infection significantly burdened health care systems and practitioners, as it is a challenge for patients suffering from both communicable and non-communicable diseases to access healthcare services.

In recent times, Artificial intelligence plays a pivotal role in the growth of various sectors including the healthcare sector. Several studies employ the use of Artificial intelligence and machine learning technologies to complement health care facilities and lessen the impact of COVID-19 in people's lives. Vaishya et al. (2020) identified key applications of AI for COVID-19 pandemic such as early detection and diagnosis of the infection, contact tracing of the individuals, monitoring treatment, projection of cases and mortality, development of drugs and vaccine. Similarly a study by Xu et al. (2020), univariate and multivariate ordinal logistic regression model were introduced to identify the independent predictors of COVID-19 severity. In their work, they particularly aimed at identifying the determinants of illness severity and to classify patients into moderate, severe and critical illness group.

The Naive Bayes algorithm is usually used to solve text related classification problems. There are two models that are commonly used namely; multinomial and multivariate models. MNB being a variant of the Naïve Bayes classifier is mainly used for efficient text classification. According to Kibriya et al (2004), out of the two Naive Bayes algorithm, it is generally agreed that the multinomial Naïve Bayes supersedes the multivariate Bernoulli event model, and they found it favourable to work with. Similarly, when it comes to computational efficiency and excellent predictive performance, Frank and Bouckaert (2006) states that, the Multinomial Naïve Bayes (MNB) is considered among the popular methods for document classification. In this work, web based Decision Support Systems (DSS) is implemented. The system adopts the Multinomial Naïve Bayes classifier to detect warning COVID-19 symptoms and provide clarification on whether patient should seek COVID-19 appropriate medical care.

The underlying concept that necessitated this study is that it is currently not possible to test all citizens especially in countries with very large populations; as such the decision provided by implementation of this study would (a) solve the problem of identifying suspected cases and deciding laboratory test priority among citizens (b) enlightening the public about the common symptoms of Covid-19. The remaining of this paper is organized as follows: (I) review of related work, (II) description of design approach and methodology, (III) evaluation process, experimental results and general discussion, (IV) the implementation of the web based systems, (V) and conclusions and future work.

RELATED WORK

Information systems, e-health and other computer based health systems have played a major role in transforming the health care system in the world. The several systems developed and deployed in the health care industry have played a successful role in transforming health care. These systems have impacted the health care system greatly as they provide support and new insight on how to tackle problems arising in the deployed area. The role Artificial Intelligence played include autonomous prediction, diagnosis and lots more, key areas applied include cardiac diseases, cancer, tuberculosis, diabetes, Parkinson disease and others.

The deployment of an autonomous AI system at 10 primary care centers by Abramoff et al. (2018) was used to detect Diabetes Retinopathy (DR) and Diabetic Macular Edema (DME) on subjects with diabetes that have not been diagnosed with DR or DME. The tests that were carried out on these patients were successful that the FDA authorized the deployment of the first autonomous AI diagnostic system. A study by Krishnaiah et al. (2013) used multiple data mining classification techniques to predict lung cancer diagnosis. They used rule set classifiers, neural networks architecture, decision tree algorithm, and Bayesian network structure discoveries. After all the applications, they found that Naïve Bayes produced the best result for diagnosing lung cancer.

With the growing number of COVID-19 cases around the world, many technological researches have been conducted to find ways to curb the pandemic. A study by Hemdan et al. (2020) introduced COVID X-Net to assist radiologist to instantly diagnose COVID-19 from x-ray images. They used deep learning framework incorporated with multiple conventional neural networks models, like Google MobileNet and six others to analyze and classify X-ray images to determine if the patient is positive or negative. VGG19 and Dense Convolutional Network (Dense Net) models gave better results than the InceptionV3 model.

Just as Hemdan et al. (2020) suggested, Shi, et al., (2020) also reviewed the various artificial intelligence techniques used to diagnose Covid-19 using images from computed tomography (CT) and X-ray. They came to a conclusion that, imaging data for Covid-19 can have inadequate and incomplete labels which poses threat to accurate diagnosis of the virus. They recommended the fusion of these image results and other deep learning techniques to make accurate and effective diagnosis.

Due to the high transmission of Covid-19, a rapid and accurate detection of the positive patients are required to assist health workers in distinguishing patients with Covid-19 and those with other respiratory diseases. According to Ying et al. (2020) they used deep learning models to diagnose Covid-19 patients quickly and to accurately differentiate them form patients showing similar symptoms with other respiratory diseases. These models adopted had high accuracy rate of quick diagnoses and efficient differentiation method with other diseases such as pneumonia, tuberculosis and so on.

Using a web-based system, Freitas et al. (2020) developed a system that used Random Forest that helps in decision making on Covid-19 patients with regard to hospitalization. The system helps health workers to determine if a patient's condition will require hospitalization in a regular ward, semi-ICU or ICU. The use of classic Decision Tree in the system has made it possible for them to achieve high diagnosis performance, handle testing availability issues and in return save lives.

METHODOLOGY

In this section, the methodology for this study is discussed based on the dataset, the algorithm, experimental settings, and system architecture.

Dataset

The dataset used is a collection of the reported symptoms of Covid-19 cases that are put together by different contributors. Obtained from the global symptoms records of Covid-19 patients, the repository on GitHub is accessible through https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data. The description of the dataset is shown in table 1.

Table 1: Dataset Description

Attributes Characteristics	Multivariate	**Attributes**	Symptom 1	Nominal
Attributes Type	Nominal, Continuous		Symptom 2	Nominal
Year	2020		Symptom 3	Nominal
Missing Values	None		Symptom 4	Nominal
No. of Records	51		Symptom 5	Nominal
No. of Attributes	7		Contact	Nominal
Name	nCoV2019		Class	Continuous

The Multinomial Naïve Bayes Algorithm

The multinomial Naive Bayes or multinomial NB according to McCallum and Nigam (1998) is a probabilistic supervised learning method. In text detection problem we aim to find the probability of a document *d* being in class *c* computed as shown in equation 1.

$$P(c|d)\alpha P(c)\prod_{1\leq k\leq n_d} P(w_k\,|c) \qquad \text{(Eq1)}$$

Multinomial Naïve Bayes begins by calculating *P(c)* which is the prior probability that a document *d* belongs to a class *c, P(w_(k)|c)* is a measure of contribution for each word *wk* in determining the correct class *c* for the document. *w1, w2, . . ., wnd* are the words in *d* that are part of the vocabulary we use for detection and *nd* is the number of words in *d*. Since this is a multinomial Naïve Bayes each term *P(w_1|c)* through *P(w_(k)|c)* is a multinomial distribution. The best class in Naïve Bayes classification is the most likely or the maximum a posteriori (MAP) class given by the result from equation 2.

$$c = \underset{c\in C}{argmax} P(c|d) = \underset{c\in C}{argmax} P(c) \prod_{1\leq k\leq n_d} P(w_k \,|c) \qquad \text{(Eq2)}$$

Experimental Settings

The experiments conducted in this study are based on four different holdout approaches on the Multinomial Naïve Bayes algorithm as shown in figure 1. This is in order to observe the differences in terms of performance and to ensure that the holdout split with the best performance is adopted as the model for building the proposed detection system.

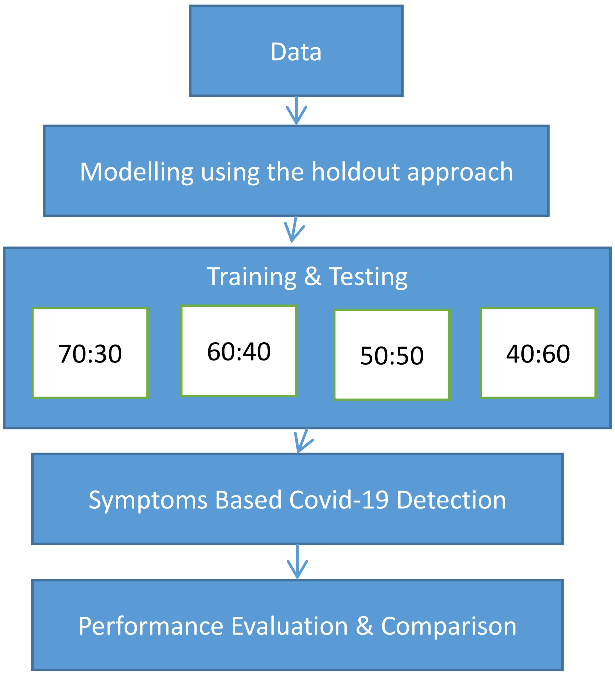

Data
Modelling using the holdout approach
Training & Testing
70:30
60:40
50:50
40:60
Symptoms Based Covid-19 Detection
Performance Evaluation & Comparison

Figure 1: Experimental Setting

System Architecture

The model with the best performance (i.e. which is the 50:50 split as obtained from the experiment results) based on the experimental settings presented in figure 1 is used in implementing the detection as shown in the system architecture in figure 2.

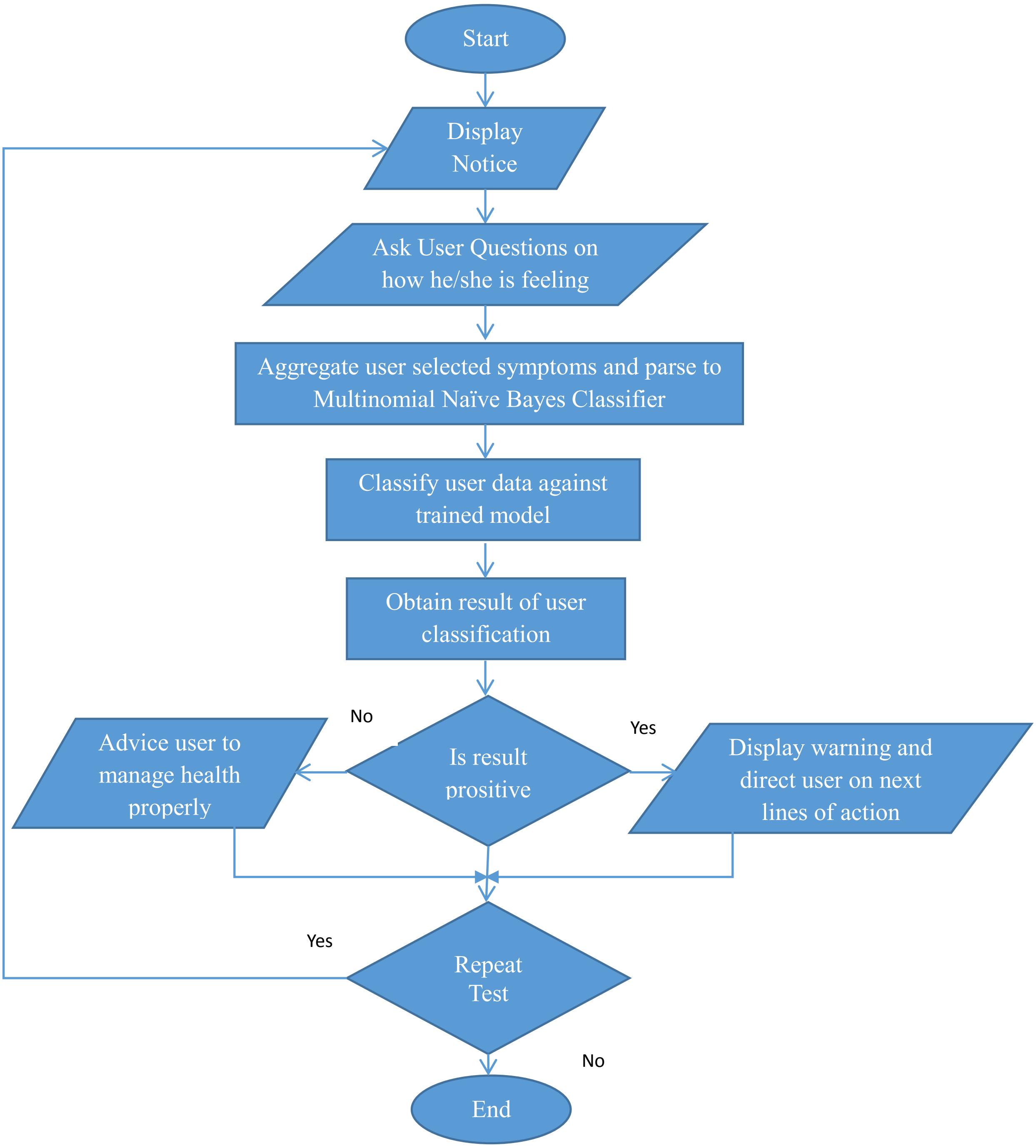
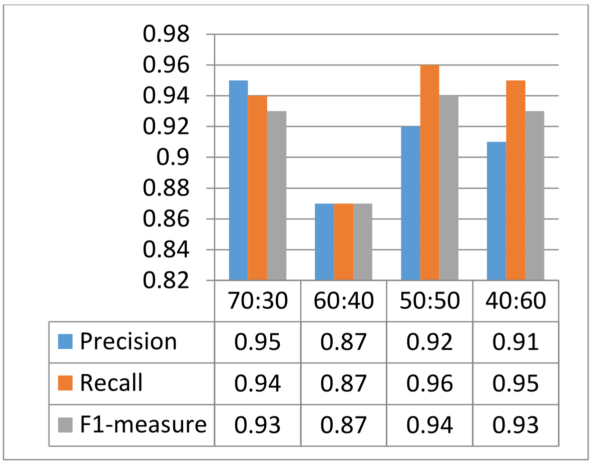

Start
Display Notice
Ask User Questions on how he/she is feeling
Aggregate user selected symptoms and parse to Multinomial Naïve Bayes Classifier
Classify user data against trained model
Obtain result of user classification
No
Yes
Advice user to manage health properly
Is result prositive
Display warning and direct user on next lines of action
Yes
Repeat Test
No
End

Figure 2: Logical System Architecture

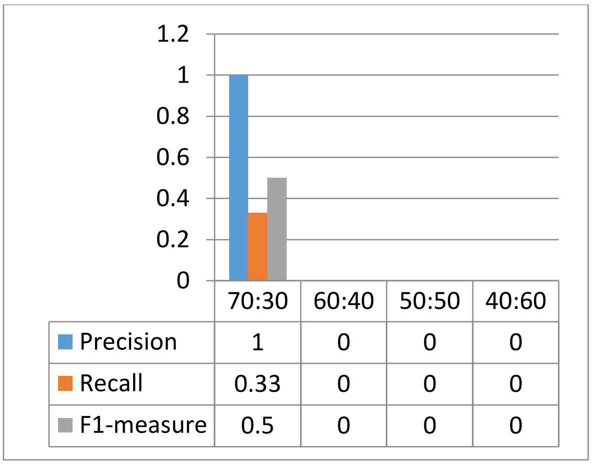
RESULTS AND DISCUSSION

Results from the different experiments conducted based on specifications from figure 1 are presented and discussed in this section. Performance in terms of precision, recall, and f1-measure are shown in figures 3, 4, and 5 respectively.



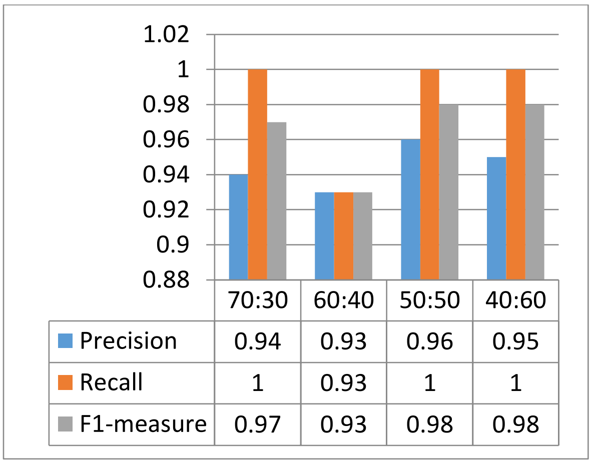
	70:30	60:40	50:50	40:60
🟦 Precision	0.95	0.87	0.92	0.91
🟧 Recall	0.94	0.87	0.96	0.95
■ F1-measure	0.93	0.87	0.94	0.93

Figure 3: Weighted performance average score on all holdout splits



	70:30	60:40	50:50	40:60
🟦 Precision	1	0	0	0
🟧 Recall	0.33	0	0	0
■ F1-measure	0.5	0	0	0

Figure 4: Performance score on negative cases on all holdout splits.



	70:30	60:40	50:50	40:60
🟦 Precision	0.94	0.93	0.96	0.95
🟧 Recall	1	0.93	1	1
■ F1-measure	0.97	0.93	0.98	0.98

Figure 5: Performance score on positive cases on all holdout splits.

In terms of the score for negative case detection, the 70:30 holdout split achieved the highest score on all three evaluation metrics of precision 1.0, recall 0.33, and f1-measure 0.5 while the others (i.e. 60:40, 50:50, and 40:60 scored 0 on precision, recall, and f1-measure. On the other hand, the performance on positive case detection for the four holdout approaches had the 50:50 and 40:60 split scoring the highest f1-measure of 0.98 followed by 70:30 split scoring 0.97, and 60:40 achieving the least score of 0.93. On the recall evaluation metric, all the splits except the 60:40 achieved same score of 1.0 while the 60:40 split achieved 0.93. Also, in terms of precision, the 50:50 achieved the highest score of 0.96 followed by 40:60 scoring 0.95, 70:30 scoring 0.94, and lastly by 60:40 scoring 0.93.

The weighted average performance of the four holdout approaches (i.e. 70:30, 60:40, 50:50, 40:60) as shown in figure 3 provides a balanced view of performance between the performance of the two (negative and positive cases) results shown in figures 4 and 5 experimented in this study. From the weighted average results as shown in figure 3 50:50 split obtained the best f1-measure of 0.94, followed by 70:30 and

40:60 both scoring 0.93, and lastly by 60:40 which scored 0.87. This is however focused only on f1-measure as it provides the harmonic mean score that balances performance between precision and recall. On precision and recall, the 50:50 split also performed well by scoring the highest precision and second to the 70:30 split in terms of recall. This on a general comparison clearly shows that the 50:50 carried the majority of highest scores on f1-measure and precision except for recall where it came second to the 70:30 split.

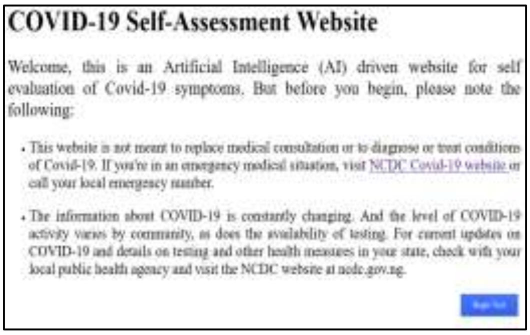
As the 50:50 holdout split performed better from the results obtained on weighted average on two out of the three evaluation metrics, it was selected to build the model for the implementation of this study. The logical design for the implementation is shown in figure 1 (i.e. system architecture), while the physical design is shown and discussed in the next section i.e. implementation.

IMPLEMENTATION

The implementation of this study based on the selected model which is the 50:50 holdout split is illustrated in this section.

a. Landing Page: This page is the default when the user visits the website. It includes an introduction, disclaimers, and links to NCDC website for guidelines, and a link to begin the symptoms test. The landing page is shown in figure 6.

b.



COVID-19 Self-Assessment Website

Welcome, this is an Artificial Intelligence (AI) driven website for self evaluation of Covid-19 symptoms. But before you begin, please note the following:

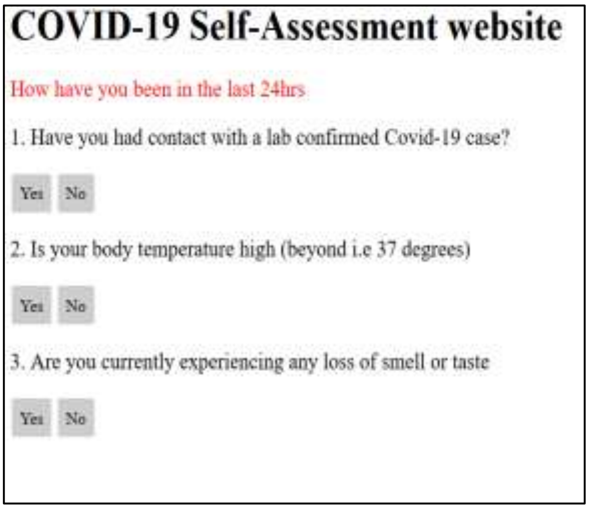
- This website is not meant to replace medical consultation or to diagnose or treat conditions of Covid-19. If you're in an emergency medical situation, visit NCDC Covid-19 website or call your local emergency number.
- The information about COVID-19 is constantly changing. And the level of COVID-19 activity varies by community, as does the availability of testing. For current updates on COVID-19 and details on testing and other health measures in your state, check with your local public health agency and visit the NCDC website at ncdc.gov.ng.

Figure 6: Landing Page

c. Symptoms Test Page: In order to several navigations and user time consumption, this stage is implemented in two pages only as described below:

   i. Test Page 1: This page contains questions about basic symptoms that have been previously reported by Covid-19 patients. This page restricts the user to single choice response to each question as shown in figure 7.

   ii. Test page 2: This page contains all other symptoms that have been reported by Covid-19 patients. This page as shown in figure 8 allows the user to select as much symptoms as the patient is experiencing.



COVID-19 Self-Assessment website

How have you been in the last 24hrs

1. Have you had contact with a lab confirmed Covid-19 case?

Yes No

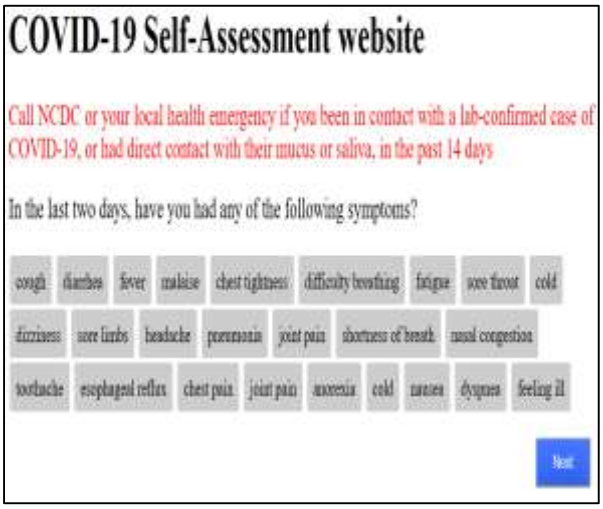
2. Is your body temperature high (beyond i.e 37 degrees)

Yes No

3. Are you currently experiencing any loss of smell or taste

Yes No

Figure 7: Test Page 1



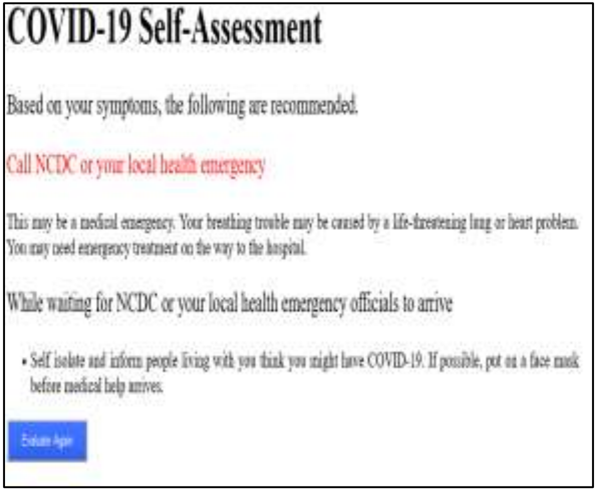
COVID-19 Self-Assessment website

Call NCDC or your local health emergency if you been in contact with a lab-confirmed case of COVID-19, or had direct contact with their mucus or saliva, in the past 14 days

In the last two days, have you had any of the following symptoms?

Figure 8: Test Page 2

d. Results Page: The two result pages i.e. one for detected Covid-19 case eligible for lab testing as shown in figure 9, and another for a not detected case that requires standard health management procedures as shown in figure 10. Based on the decision arrived at by the trained AI model on symptoms selection by the user, the appropriate result page is displayed.



COVID-19 Self-Assessment

Based on your symptoms, the following are recommended.

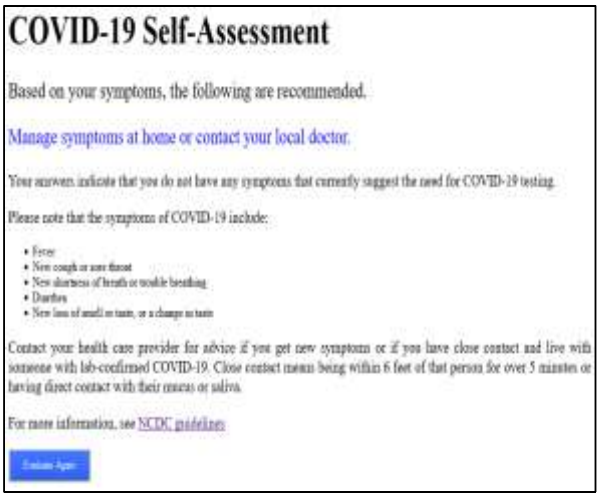
Call NCDC or your local health emergency

This may be a medical emergency. Your breathing trouble may be caused by a life-threatening lung or heart problem. You may need emergency treatment on the way to the hospital.

While waiting for NCDC or your local health emergency officials to arrive

- Self isolate and inform people living with you think you might have COVID-19. If possible, put on a face mask before medical help arrives.

Figure 9. Detected Case



COVID-19 Self-Assessment

Based on your symptoms, the following are recommended.

Manage symptoms at home or contact your local doctor.

Your answers indicate that you do not have any symptoms that currently suggest the need for COVID-19 testing.

Please note that the symptoms of COVID-19 include:

- Fever
- New cough or sore throat
- New shortness of breath or trouble breathing
- Diarrhea
- New loss of smell or taste, or a change in taste

Contact your health care provider for advice if you get new symptoms or if you have close contact and live with someone with lab-confirmed COVID-19. Close contact means being within 6 feet of that person for over 5 minutes or having direct contact with their mucus or saliva.

For more information, see NCDC guidelines

Figure 10 No Detection Case

CONCLUSION

In this study, the Multinomial Naïve Bayes algorithm was implemented on four holdout splits to detect Covid-19 based on presented symptoms. Although, the difference in terms of performance between the different holdout approaches was not significantly high, the 50:50 holdout split having scored the highest was selected for modelling the implementation of the web interface where the Covid-19 detection is carried out based on the presented symptoms.

REFERENCES

Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Med, 1*(39).

Frank , E., & Bouckaert, R. R. (2006). Naive Bayes for Text Classification with Unblanaced Classes. European Conference on Principles of Data Mining and Knowledge Discovery (pp. 503-510). Berlin: Springer.

Freitas, V. A., Gomes, J. C., Lima, C. L., Calado, R. B., Bertoldo, C. R., Albuquerque, J. E., . . . Santos, W. P. (2020). Covid-19 rapid test by combining a random forest based web system and blood tests. Brazil.

Hemdan, E.-D., Shouman, M. A., & Karar, M. E. (2020). COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose. arXiv preprint arXiv.

John Hopkins University and Medicine. (2010). *Covid-19 data in motion*. Retrieved July 10, 2020, from https://coronavirus.jhu.edu/map.html.

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinominal Naive Bayes for Text Categorization Revisited . Australasian Joint Conference on Artificial Intelligence (pp. 488-499). Berlin: Springer.

Krishnaiah, V., Narsimha, D., & Chandra, D. S. (2013). Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Science and Information Technologies, 4*(1), 39-45.

Mayo Clinic. (2020). *Covid-19 disease 2019 (COVID-19)*. Retrieved July 07, 2020, from https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963.

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 workshop on learning for text categorization, 752*(1), 41-48.

Shi, F., Wang, J., Shi, J., Wu, Z., Wang , Q., Tang, Z., . . . Shen, D. (2020). The Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19. IEEE Reviews in Biomedical Engineering.

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) application for COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14*(4), 337-339.

Xu, K., Zhou, M., Yang, D., Ling, Y., Liu, K., Bai, T., . . . Cheng, Z. (2020). Application of ordinal logistic regression analysis to identify the determinants of illness severity of COVID-19 in China. *Epideminology & Infection, 148*, 1-25.

Ying, S., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., . . . Yang, Y. (2020). Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. China.