JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 8(2), JUNE, 2020
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

Email Urgency Classifier Using Natural Language Processing and Naïve Bayes

[1]Ogude, U. C., [2]Olaniyan O. M., [1]Oni, Adeola C., [1]Uwadia, C. O.
[1]Dept. of Computer Sciences
University of Lagos, Lagos, Nigeria
[2]Dept. of Computer Engineering
Federal University,Oye Ekiti, Nigeria *Corresponding Author:olatayo.olaniyan@fuoye.edu.ng*

ABSTRACT
Emails are today's most commonly used means of communication. It is one of the Internet's greatest inventions. Billions of emails are transferred daily for various purposes. The Harvard Business Review states that after an interruption (e-mail or other), it takes 20 minutes to regain full focus. This is backed by the argument of Loughborough University that when interrupted by phone, tasks take 33 percent longer to complete. Hence, people are constrained on how to manage their time effectively; this is because a piece of time-critical information may be received and needed to be worked upon urgently. Large organizations receive thousands of emails every day; they cannot decide on which email to respond to first because the emails are already in a hierarchy, based on the time they were received. Hence, emails that are time-critical or require urgent response may be pushed to the bottom of the list. The aim of this paper is to present a model that can interpret human language and efficiently discern between urgent and not urgent emails. When emails are sent, the model prioritizes which emails should be responded to by the receiver based on the level of urgency and importance of the content of the email. The tools used to classify the emails into urgent and not urgent are Natural Language processing and Naïve Bayes for text classification. The model is created using Python programming language on Spyder (a Python Integrated Development Environment (IDE)) and Scikit-learn (an efficient Python data mining and analysis tool). The result obtained is the percentage (polarity) of the urgency of the email, hence urgent emails are placed on top of the inbox list. This paper contributes to knowledge by the creation of a dataset on Kaggle. There is unavailable dataset for this project, I had to create a new one

ARTICLE INFO
Article History
Received: January, 2020
Received in revised form: April, 2020
Accepted: May, 2020
Published online: June, 2020

KEYWORDS
Emails, Human Language, Natural Language processing, Naïve Bayes, Text classification, Urgency.

INTRODUCTION

Communication is the act of exchanging of information between two or more people at different places, via speaking, writing or any other medium. The importance of communication cannot be overemphasized or underemphasized. Written communication is the most effective way of relaying information, intent or ideas between groups of people. This could be

Corresponding author: olaniyan, O. M. ✉ olatayo.olaniyan@fuoye.edu.ng ✉ *Department of Computer Sciences, University of Lagos, Nigeria.*

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 8(2), JUNE, 2020
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

through SMS (short message service), email (electronic mail), social media channels or paper written messages. Communication through several miles and time difference is a luxury that SMS, email and social media provide [1]. Information management is very crucial in large organizations because time-critical information may be sent through email, such an email should not go unattended to within reasonable time. Due to the vast amount of emails that organizations receive daily, they cannot decide on which email to respond to first because the emails are already in a hierarchy based on the time it was received. An average person checks their emails about 15 times a day. Researchers at the University of British Columbia found that there was substantially less stress on people who decreased the amount of times to 3 times a day. There is a huge difference between urgent and important emails, urgent emails require immediate attention and they have repercussions if they are not attended to immediately. Urgent emails cannot be put on hold because they are short-termed.

RELATED WORKS

One of the first concerns about the growth of computers is that 'will a computer program ever be able to translate a text in such a way that it will describe the natural meaning of the text?' Natural Language Processing (NLP) is a field of Artificial Intelligence that is focused on processing and understanding unstructured data. It is targeted at making machines understand and gain from raw text.

Classification is the mechanism whereby the class of data points is expected. Often, groups are referred to as targets/labels or divisions. Predictive modeling classification is the job of approximating a mapping function (f) from variables of input (x) to discrete variables of output (y) [2].

Naïve Bayes was developed based on the Bayes Theorem named after the statistician and philosopher Thomas Bayes. The theorem is the base for the Naïve Bayes Algorithm. Bayes theorem provides the method to calculate conditional probability. This means that the probability of an event is based on previous knowledge available.

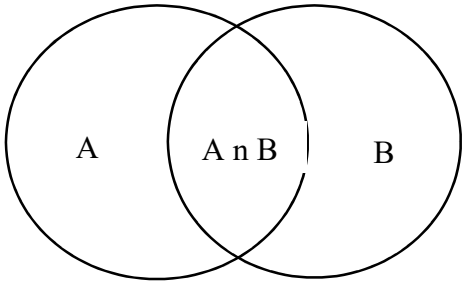

Figure 1.0: Venn diagram of two sets A and B

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (1.0)$$

P (A|B): the probability of occurrence of event A given the event B is true. It is called the evidence

P (A): the probability of occurrence of event A. Event A is called the proposition

P (B): the probability of occurrence of event B. Event B is called the evidence

P (B|A): the probability of occurrence of event B given the event A is true. It is called the posterior

$$\text{Posterior} = \frac{(\text{Likelihood})\ (\text{Proposition prior probability})}{\text{Evidence prior probability}} \qquad (2.0)$$

Our system uses Natural Language Processing to do the following:

a. **Data cleaning:** The email content is transformed into lowercase, all non-alphabets are removed and each sting is split for easy removal of stopwords and stemming.
b. **Stop Words Removal:** Stop words are removed from the body of the email to reduce irregularities. This is done by checking the stopwords library to see if a word is contained

Corresponding author: olaniyan, O. M. olatayo.olaniyan@fuoye.edu.ng *Department of Computer Sciences, University of Lagos, Nigeria.*

in it, if it is, the word is not added to the corpus to be trained.

c. **Stemming:** Each word is reduced to its base form or stem. The words are stRipped of their prefixes and suffixes. This is done by an inbuilt library known as potterStemmer.

There are a number of applications built using either or both Natural Language Processing and Naïve Bayes Classifier. They are as follows:

Sentiment Analysis*:* Sentiment Analysis (or opinion mining) refers to the use of Natural Language Processing, text analysis, computational linguistics and biometrics to define, detect, measure and systematically analyze affective states and subjective information. It refers to mining of feelings, sorting of feelings, etc. It is a way of determining a writer's mood or point of view. In other words, extracting the emotions from the text is the aim [3].

Spam/Ham Filter*:* A predictive model is built to classify if a message is spam or ham. After using Natural Language Processing to understand the text. The classification task is to classify into the target variable which is distinct, either 'ham' or 'spam' [4].

Spell check*:* A spell checker system uses NLP to compare potentially misspelled words in the corpus as a domain-specific source for dictionary terms. Spell check is an NLP method that is now used by everyone. It is unobtrusive and easy to use [5].

Autocomplete: An autocomplete system gives you advice on how to complete sentence you are writing. It gives options of words gotten from the stem of what you are typing. Language modeling is the best tool to be used in this case. It builds a statistical language model using NLP to find a distribution probability over word sequences [6].

Virtual Assistants: In this new era of the 21st century, virtual assistant is a boon for everyone. It has paved the way for new age in which we can ask computer questions and get answers. Whether you are riding, have your hands full or just on the go. These assistants will make calls, receive messages, provide constructive ideas and alternative solutions to problems. Examples are Siri, Google Assistant, and Alexa. A big part of their make-up relies on NLP [7].

EXISTING SYSTEM

Email classification Using Back Propagation: This is an automated email classification system based on a neural network divided into user-defined "word groups." This is an experiment focused on topic field data in a mailbox and quality of email messages. The classes are words with meaning (Critical, Urgent, Very important and others). The learning technique used is an associative training that trains the network by supplying it with patterns of input and output matching. An external instructor or the device that comprises the neural network (self-supervised) provides these input-output pairs [8].

PROPOSED SYSTEM AND TOOLS USED

After looking for datasets for the problem statement and finding none, we decided to create one on Kaggle. Naïve Bayes is the best option when it comes to getting good results using very little data. Traditional text classification algorithms are the best when building a model with very limited amount of data.

TEXT CLASSIFICATION ALGORITHMS

Traditional Text Classification Algorithms

a. Naïve Bayes: Naïve Bayes classification algorithm is a probabilistic classifier that learns from the probability of an object with certain features belonging to a particular group and class. It is useful when we have scarce computational resources or data. The Naïve Bayes algorithm is called 'naïve' because it makes assumptions that the

Corresponding author: olaniyan, O. M. olatayo.olaniyan@fuoye.edu.ng *Department of Computer Sciences, University of Lagos, Nigeria.*

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 8(2), JUNE, 2020
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

occurrence of a feature is independent of the occurrence of other features**.**

b. Support Vector Machine (SVM): Support vector machine is just like Naïve Bayes in terms of the number of computational resources or data it requires. SVM requires larger computational resources and can produce better results [9].

Deep Learning

Deep learning is used for computation through multilayers to represent data after abstraction [10]. Deep learning is inspired by the human brain; each layer is computed based on representations from the previous layer. Deep learning uncovers hidden structure in large data sets of images, videos, speech, and audio through backpropagation algorithm. There are two major deep learning architectures used for text classification. They are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

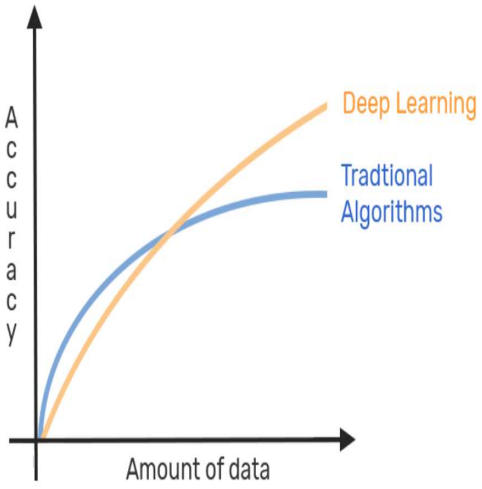

Figure 2.0: Graph showing the relationship between the amount of data and the prediction accuracy [9].

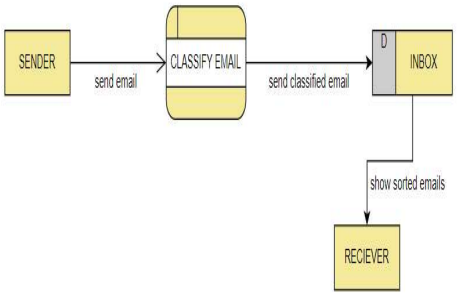

Figure 3.0: Data flow diagram of the email urgency classifier

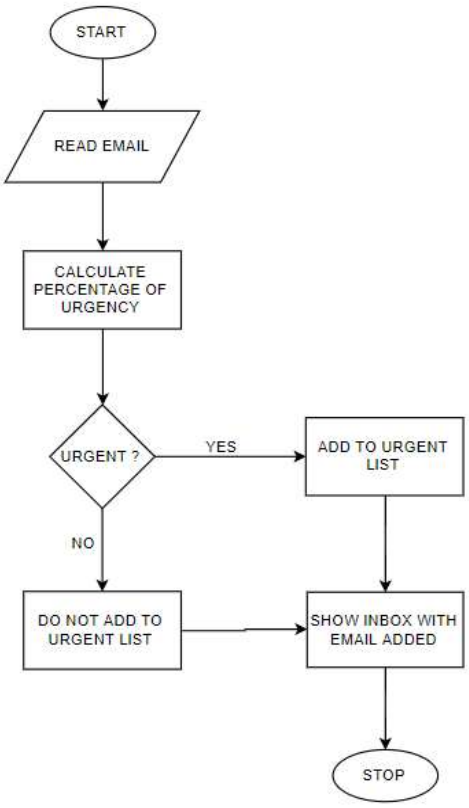

Figure 4.0: Flow chart of the processes involved in the classification of an email

Our model accepts user input (a received email) and returns the prediction after running the email through the model. The data in the email body is cleaned by removing all non-alphabets, changing the case to lower case, splitting the sentence into individual words, removal of stopwords,

Corresponding author: olaniyan, O. M. olatayo.olaniyan@fuoye.edu.ng *Department of Computer Sciences, University of Lagos, Nigeria.*

and reducing each word to its stem form. A bag-of-words model is created by making a text unigram model and maintaining the track of each word's occurrences. Individual phrases in the bag-of-words model are considered and each word is offered a particular subjectivity score.

TOOLS

The languages, frameworks, and libraries used for the creation of the urgency classifier are python, flask, HTML, CSS, pickle, Javascript, pandas, numpy, matplotlib, seaborn, re, nltk and sklearn. The functions of these are

a. Python: Python is a language of programming that is interpreted, high-level, general-purpose. Python was used in the development of the machine learning model.
b. Flask: This is a web application framework for a lightweight WSGI (Web Server Gateway Interface) application. It is designed with the ability to scale up to complex applications to make getting started quick and easy. This served as a bundler for the front end and the model.
c. HTML: Hypertext Markup Language (HTML) is the standard markup language used to view documents in a web browser. This was used in structuring the interface and creating the form for submitting user input.
d. CSS: CSS stands for Cascading Style Sheets, it defines how to view HTML elements on computer, paper or other media This was used to design the interface.
e. Pickle: Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it "serializes" the object first before writing it to file. Pickling is a way to convert a python object (list, dictionary, etc.) into a character stream. The idea is that this character stream contains all the information necessary for the reconstruction of the object in another python script
f. Javascript: JavaScript is a language for high-level scripting comprehension. This was used for action calls on the form and sending data to the flask module.
g. Pandas: This is an open-source library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language [11].
h. Numpy: NumPy is the fundamental package for scientific computing with Python. It is useful linear algebra, Fourier transform, and random number capabilities. NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases. [12].
i. Matplotlib: Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across the platform [13].
j. Seaborn: seaborn provides a high-level interface to draw statistical graphics. it is complimentary to Matplotlib and it specifically targets statistical data visualization [14].
k. Re: Regular Expression, is a sequence of characters that forms a search pattern. It is used to check if a string contains the specified search pattern. Python has a built-in package called re.
l. Nltk: NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic

Corresponding author: olaniyan, O. M. olatayo.olaniyan@fuoye.edu.ng Department of Computer Sciences, University of Lagos, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 8(2), JUNE, 2020
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

reasoning [15]. A word cloud is created using WordCloud library. This is an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance. So, the more often a specific word appears in your text, the bigger and bolder it appears in your word cloud.

RESULTS

Figures 5.1 to 5.4 show the output and workflow of the model before and after an email has been sent.

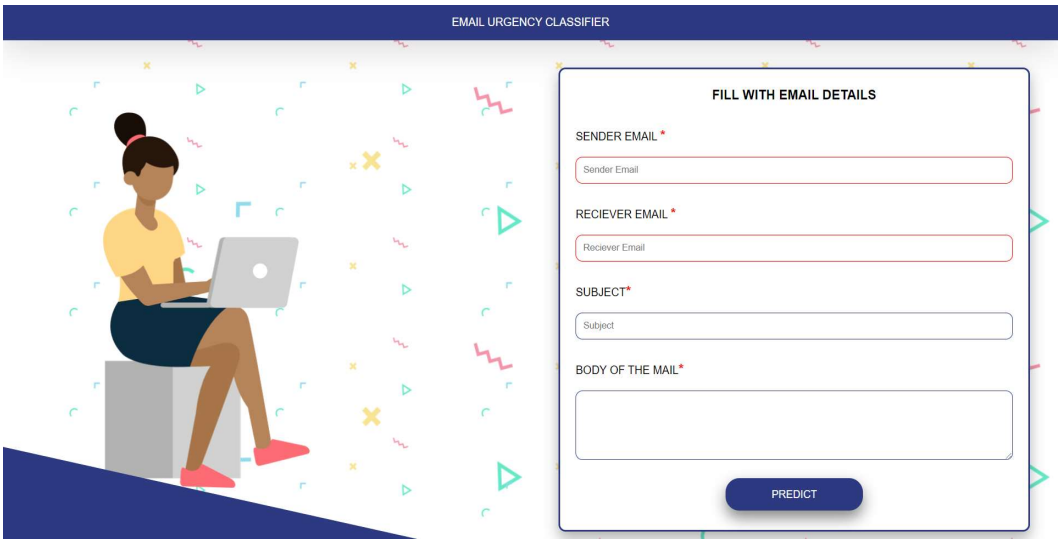

Figure 5.1: An image of the developed interface for testing the model.

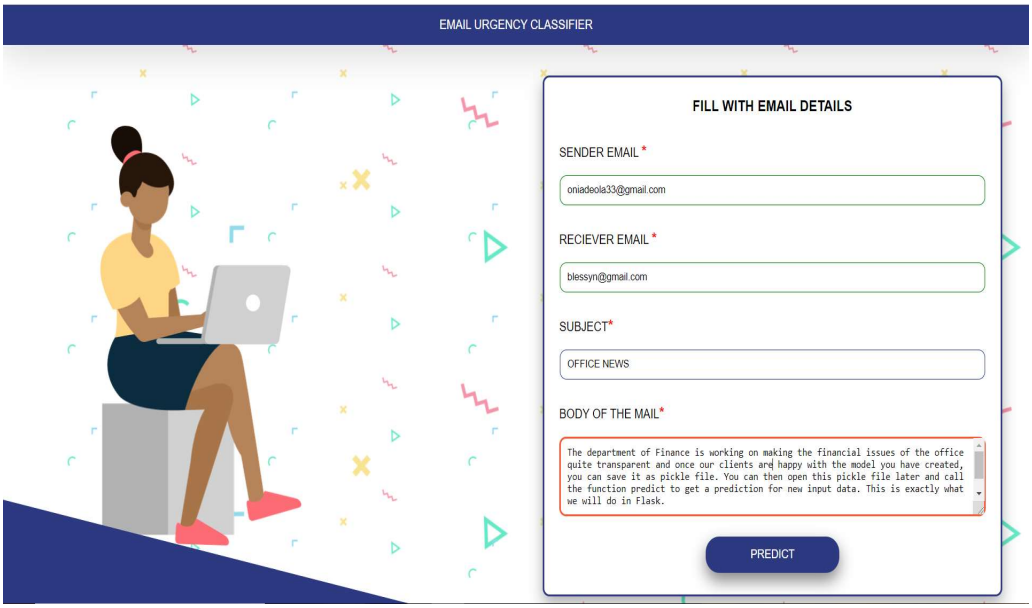

Figure 5.2: An image of the interface with an email to be sent

Corresponding author: olaniyan, O. M. olatayo.olaniyan@fuoye.edu.ng *Department of Computer Sciences, University of Lagos, Nigeria.*

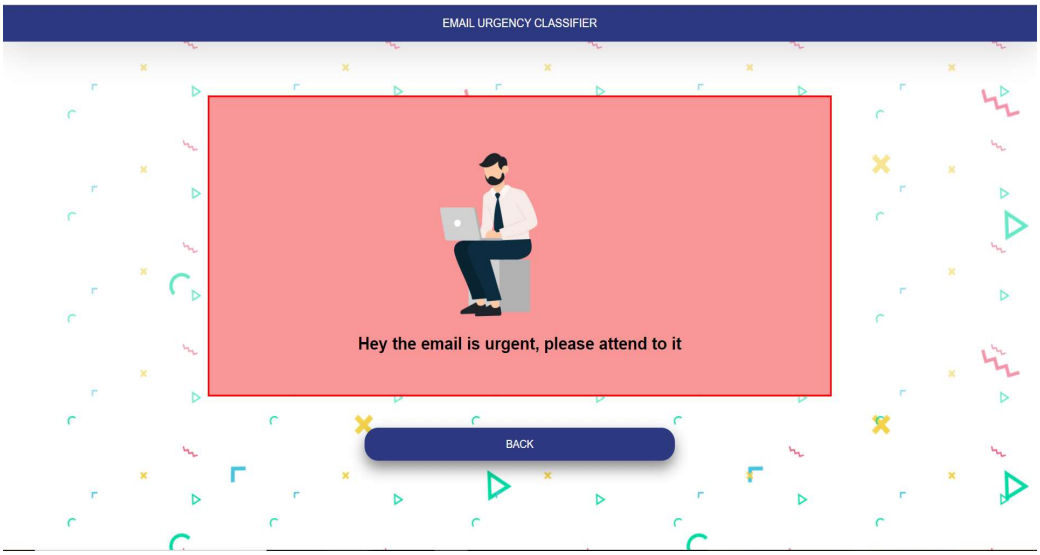

Figure 5.3: An image of the interface response when an urgent email is received.

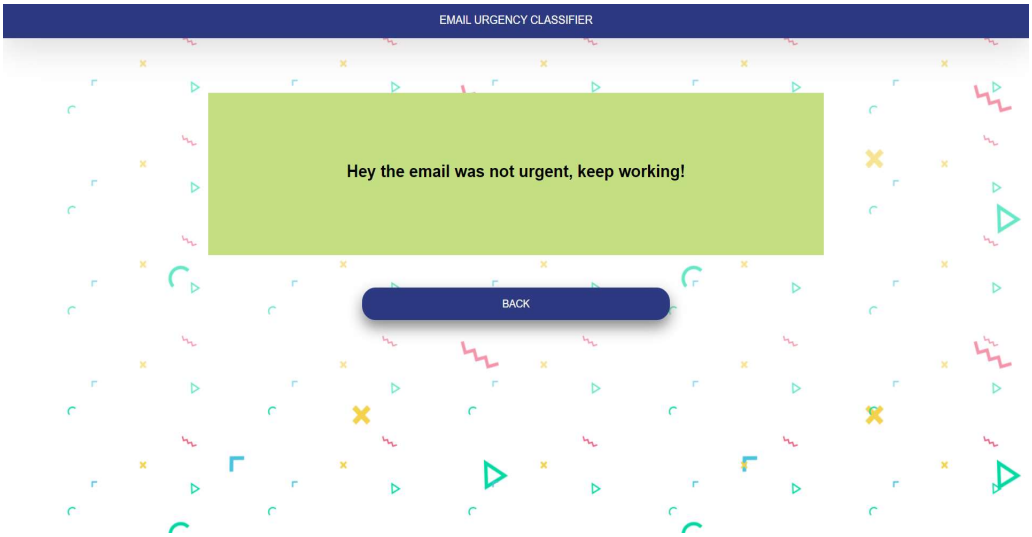

Figure 5.4: An image of the when an email that is not urgent is received.

CONCLUSION

In closing, this paper has been able to give a sense of orderliness to incoming emails. The model categorizes incoming pieces of text as urgent or not urgent based on if there is a request for immediate attention. This project will be useful in making sure that critical activities like a house on fire or rushing a pregnant woman going into labuor to the hospital take priority.

FUTURE WORKS: After gathering enough dataset, the model can be improved in the following ways:

1. The model can be trained using deep learning, to make it more diverse and futuristic.

Corresponding author: olaniyan, O. M. ✉ olatayo.olaniyan@fuoye.edu.ng ✉ *Department of Computer Sciences, University of Lagos, Nigeria.*

2. It can be added to email and messaging services as a feature.

REFERENCES

[1] J. Velentzas, and G. Broni, "Communication cycle: definition, process, models and examples", Technological institute of western Macedonia, Greece, 2004, pg.1-4.
[2] K. Pouria, and D. Sunita, "Short survey on Naive Bayes algorithm", International journal of advanced research in Computer Science and Management, 2004.
[3] M. El-Din, and Doaa,"Analyzing scientific papers based on sentiment analysis (first draft)", 2016.
[4] Geerthik & Prof. Sree, "Survey on internet spam: classification and analysis", International journal of computer technology and applications" 2013.
[5] V. V. Bhaire, A. A. Jadhav, P. A. Pashte, M.R. Mane, College of Engineering and Technology. International Journal of Scientific and Research Publications, Vol. 5, Issue 4, April 2015.
[6] Ward, David and Hahn, Jim and Feist, Kirsten, "Autocomplete as research tool: a study on providing search suggestions, information technology and libraries, 2012 P.g 31.
[7] T. Amrita, and D. Sudhir, "Survey on virtual assistant: Google assistant, siri", Cortana, Alexa: fourth International Symposium SIRS 2018, Bangalore, India, September, 2018.
[8] A. Taiwo, Z. Shikun, and K. Rinat. "Email classification using back propagation technique", International journal of intelligent computing research, Vol. 1. 2017.
[9] https://monkeylearn.com/text-classification (Last accessed July 2019).
[10] Y. LeCun, Y. Bengio, and G. Hinton,, "Deep learning", Vol. 521,, p.g 436, May 2015.
[11] https://pandas.pydata.org/ (Last accessed of August 2019).
[12] https://numpy.org (Last accessed of August 2019).
[13] https://matplotlib.org(Last accessed of August 2019).
[14] https://seborn.pydata.org/ (Last accessed of August 2019).
[15] https://www.nltk.org/ (Last accessed of August 2019).

Corresponding author: olaniyan, O. M. olatayo.olaniyan@fuoye.edu.ng Department of Computer Sciences, University of Lagos, Nigeria.