PSYCHOLOGICAL EMOTION RECOGNITION OF STUDENTS USING MACHINE LEARNING BASED CHATBOT

Suha Khalil Assayed[1], Khaled Shaalan[1], Sana Alsayed[2], Manar Alkhatib[1]

[1]Faculty of Engineering and IT, The British University in Dubai, UAE
[2]Department of Counseling Psychology, Philadelphia University, Jordan

ABSTRACT

Anxiety and depression can have a significant impact on students' academic performance, however, these mental health impacts were increased during the Covid-19 pandemic, and accordingly students and parents need some people to share their feelings together; however, there are different types of social media apps and platforms such as Facebook, Twitter, Reddit, Instagram, and others. Twitter is one of the most popular social application that people prefer to share their emotional states. Interestingly, the psychologist and computer scientists are inspired to study these emotions. In this paper, we propose a chatbot for detecting the students feeling by using machine-learning algorithms. The authors used a dataset of tweets from Kaggle's paltform, and it includes 41157 tweets that are all related to the COVID-19. The tweets are classified into categories based on the feeling: Positive and negative. The authors applied Machine Learning algorithms, Support Vector Machines (SVM) and the Naïve Bayes (NB) and accordingly they compared the accuracy between them. In addition to that, the classifiers were evaluated and compared after changing the test split ratio. The result shows that the accuracy performance of SVM algorithm is better than Naïve Bayes algorithm, but the speed is extremely slow compared to Naive Bayes model. In future, other neural network algorithms such as the RNN, LSTM will be implemented, and Arabic tweets will be included in the future.

KEYWORDS

Psychological Emotion, NLP, Covid-19, Chatbot, Machine Learning, Students, SVM, Naïve Bayes

1. INTRODUCTION

Due to the growth of the internet and the telecommunication around the world, several applications are developed by using state-of-the- art models and algorithms. However, the chatbot is considered one of these smart technologies that inspired the researchers around the world to study it into different fields, including education, finance, healthcare and others, as a result it can improve the quality of services as well as solving different challenges and issues that might affect customers' satisfaction. Since the Covid-19 pandemic was one of these issues that preoccupied the world in the past couple of years, it had a significant impact from different dimensions including the education [1]. Accordingly, the social media interactions have increased with peoples' comments that reflected their feelings towards the Covid-19 pandemic and its impact. For example, during the pandemic, some students expressed their experiences and their panic when they got sick while they had exams, while others expressed their opinions toward learning from home. Moreover, many schools and decision makers attempted to close schools and colleges in order to contain the spread of the COVID-19 pandemic, and this closure affected above 60% of the worlds of students' population [2].Several studies revealed that students reported different psychologically impact of the Covid-19 on students and learning, such as

stress, anxiety and depression [3, 4]. According to the study of Pelucio et al. [5], which conducted in one of the universities in Brazil for evaluating the presence of depression, and anxiety in university students during Covid-19 pandemic, the results revealed that most of the students reported emotional impact with significant difference of depressive symptoms but no significant difference in anxiety. However, the computer scientist and psychologist can work together in detecting the students feeling by using the opinion mining or the sentiment analyses to study their attitudes and emotions [6]. Therefore, in this study the sentiment analysis of students will be detected and evaluated by developing a chatbot based on supervised learning algorithms, however the corpus in this study is English tweets which classified as positive or negative. Authors used the Support Vector Machines (SVM) and the Naïve Bayes (NB) algorithms to compare the accuracy of the classifier and the execution time between the two models with using different features. Figure 1 shows the diagram of processing the sentiment analysis.

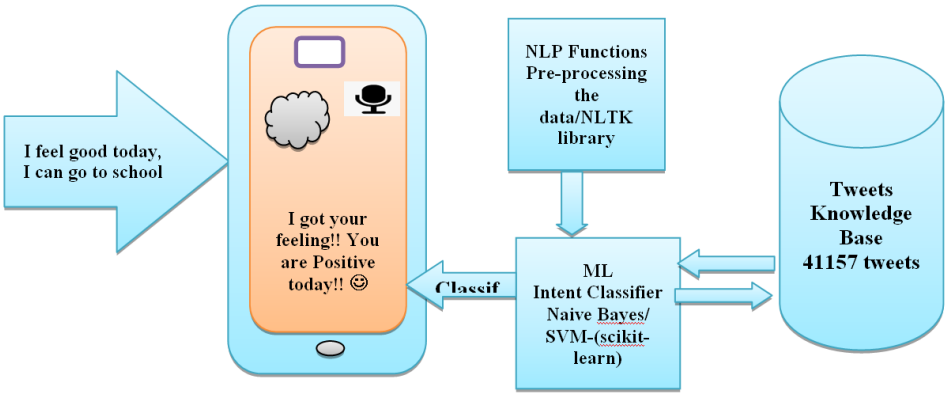

Figure1. The Framework of the chatbot – detection emotions in this study

2. RELATED WORK

Researchers around the world are inspired to develop the state of-the-art chatbot by using different machine learning algorithms such as naïve Bayes algorithm and support vector machine (SVM) [7], [8].

Sentiment analysis is an approach to analyze people' feeling and opinion from texts of contents [9]. Moreover, some authors perform mathematical operations to examine people's feelings about specific events [10]. Rani & Singh [11] study the sentiment analysis for Twitter data which that collected by using the Twitter (API). Then they preprocessed the data, afterward they used SVM for mining the people's feeling with applying the following features: TF-IDF, Linear, and Kernel. Moreover, they used the F-score, recall, accuracy, and precision to measure the performance of the model. The results show that the SVM outperform the Naïve Bayes algorithm.

Moreover, Alabid & Katheeth [12] used the twitter data in the SVM to predict the sentiment analysis during the COVID-19 pandemic. They used recall, F1, precision, and confusion matrix in order to evaluate the performance of the SVM algorithm. The authors applied in their study 629 tweet texts and divided it as the following: 40% of tweets show neutral sentiments, 25% of tweets show positive, while 35% of tweets show negative. Then the dataset divided to 80% training and 20% testing data, the result of the performance reached to 71%. Furthermore, Naw [13] used SVM and K-Nearest Neighbor (KNN) algorithm to conduct sentiment analysis on dataset collected by Twitter API. The author used different features including the term frequency and inverse document frequency (TF-IDF) the data were classified as negative, neutral, and

positive [13]. Other algorithms used as well by Alabid & Katheeth [12], they applied both the SVM and Naïve Bayes algorithms to conduct a sentiment analysis of the twitter texts related to the COVID-19 vaccines. The ratio of training data was 80% and the ratio of testing data was 20%. They preprocessed the dataset by removed stop words, punctuation and they applied the part -of-speech (PoS) tag. Subsequently, they selected the adjectives sentences for clearing the ambiguous words. the results revealed that SVM was better than NB with test ratio .01 while the stop words was removed from the texts. Moreover, the results showed that the performance of NB was better than SVM with ratio .06, when they used the PoS tag and removing stop words. In general, sentiment analysis attracted a lot of researchers to pay more attention to this field and to use several algorithms to improve the classifiers. Though, social media play an important role during the Covid-19 particularly in driving researchers around the world to use several techniques in Natural Language Processing (NLP) to analyze people's perspectives and opinions during this pandemic.

2.1. Emotions detections Based on Social Media during Covid-19

Ouerhani et al. [14] developed a novel chatbot, called COVID-Chatbot for communicating with people during Covid-19 to increase their awareness towards the risk of this pandemic. Liu et al. [9] conducted other research paper to study how people think and act during this pandemic from the lens of social media posts and blogs by using the Bidirectional Encoder Representations from Transformers (BERT), as well as other layers from neural networks techniques.

In general, most studies were intended to study the people's feeling to measure and detect the level of their anxieties and depression. Fauziah et al. [15] developed two machine language models, the random forest and xgboost to detect the anxiety feeling and emotions during the pandemic, the author used 4862 records from a dataset that collected from the YouTube comments. Furthermore, Ryu et al. [16] used the Machine Learning models to detect the patients' anxiety during the Covid-19 pandemic by using data from two different types of social media apps namely a communication app as well as a social networking app. On the other hand, Vahedi et al. [17] used data from Facebook's platform to predict the spreading of new cases of Covid-19.

Chin et al. [18] analyzed 19,782 conversation utterances that related to COVID-19 which cover multiple countries, the authors used different techniques from NLP and ML methods to analyze the emotional sentiments. Yao et al. [19] and other authors used machine language algorithms to identify the peoples' feedback from the vaccinations [20, 21].

3. EXPERIMENT AND DEVELOPMENT

This study adopted a machine learning approach by using different NLP techniques as well as supervised learning algorithms SVM and Naïve Bayes classifier. However, the methodology is mainly divided into four sections, 1- Data Collection, 2- Preprocessing the Data, 3- Building the model and Training the Data 4- Testing the Machine Learning Algorithms. The authors used the Anaconda platform with different packages from Python libraries such as the Scikit-learn, due to the fact that it's considered as one of the most powerful text processing tools that support and provide tokenization, filtration of tokens, and stemming.

3.1. Data Collection (Corpus)

The Corpus is always the starting point for any preprocessing function, as it has the main texts and sentences; each sentence is divided into words, which called a token.

This study used a Tweets corpus as a collection of text tweets that were collected from the Twitter platform, which retrieved from Kaggle's website in CSV, and it includes 41157 tweets, and all are taged and classified based on the sentiment of the tweet (Extremely Positive, Positive, Neutral, Negative, Extremely Negative). Moreover, the testing data split it into different ratios 10%, 20%, and 30% in order to compare the performance of SVM and Naïve Bayes models. Figure 2 shows the description of the tweets corpus.

```
[3]: train_dataset.describe()
```

[3]:

	UserName	ScreenName
count	41157.000000	41157.000000
mean	24377.000000	69329.000000
std	11881.146851	11881.146851
min	3799.000000	48751.000000
25%	14088.000000	59040.000000
50%	24377.000000	69329.000000
75%	34666.000000	79618.000000
max	44955.000000	89907.000000

Figure.2 The characteristics of the corpus.

3.2. Data Preprocessing

The Pre-processing is a necessary step for any Natural Language Processing system since most elements of the texts including characters, words, and sentences are important through the entire stages of the text processing. The purpose of all these stages is to make the text more understandable for the machine learning models. Thus, in simple words, we can define it as a technique for converting the raw data into understandable tokens for ML. In general, preprocessing the text includes four processes: 1-Text Tokenization, 2- Removing stop words, 3- Normalization, and 4- Stemming & Lemmatization. These four processes are utilized in order to simplify the text to a new format that can be understood by NLP and ML models. Figure 3 shows the main steps of preprocessing the data.

NLTK (Natural Language Toolkit) in Python, is the most important component for preprocessing the text, and authors used this library for most of techniques in preprocessing the data.

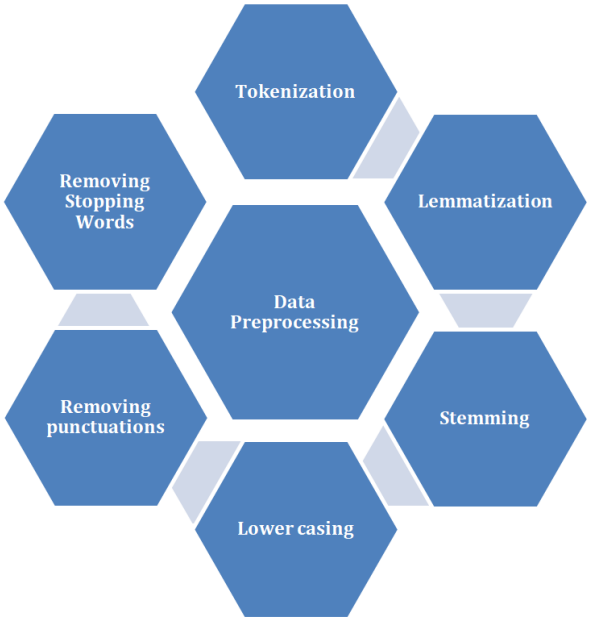

Figure 3. The main steps of preprocessing the data

3.2.1. Text Tokenization

The Tokenization is the first step for NLP and it's defined as splitting the text into characters, words, sub-words as a token by using different methods. However, the sub-words known as n-grams and (n), are considered as number of tokens, since some words can be more understandable when combined together. The Tokenization has an important impact on analyzing and processing the data as these tokens become as an input to other functions such as parsing and data mining.

3.2.2. Removing Stop words

The datasets for both the Training and the Testing are cleaned from the Stop words by importing the module by using functions from NLTK library, which can maximize the efficiency of the Dataset.

3.2.3. Normalizing and Stemming/ Lemmatization the Data

Stemming and Lemmatization is basically for simplifying the words to a unique meaningful word, since one word can turn into different forms of the word, but all can be shared by the same meaning. For instance, "study", "studies", "studying", "studies", etc. without stemming and lemmatization the corpus will be tokenize as 4 different tokens, but after preprocessing it will be counted only one token" study".

3.3. Training the Machine Learning Model

In this study, we selected Python for deploying two selected machine language algorithms, the Naïve Bayes algorithm and the Support Vector Machine (SVM), due to the fact that Python includes several effective Machine Learning libraries such as scikit-learn, TensorFlow, etc. Besides that, Python is the most preferred language for data science and ML due to the low-level libraries and clean high-level APIs. Furthermore, we used the Jupiter Notebook 3.0.14 from the Anaconda platform, since it's considered as one of the most powerful environments for data scientists.

3.3.1. Naïve Bayes Machine Learning

Naïve Bayes algorithms was used for classification, at it is one of the supervised learning algorithms. This classifier works by training the data with labeled categorical feeling inputs. This classifier works based on Bayes theorem by calculating the probabilities for each tag. For example, in our dataset, we have positive and negative tweets. First, we need to classify whether each word in the tweet is Negative or Positive, then will calculate the frequency in each one. This would be followed by creating the probability for each tag. Figure 4 shows a sample of students' positive and negative feeling of tweets and Table 1 explain how the Naïve Bayes algorithms work.

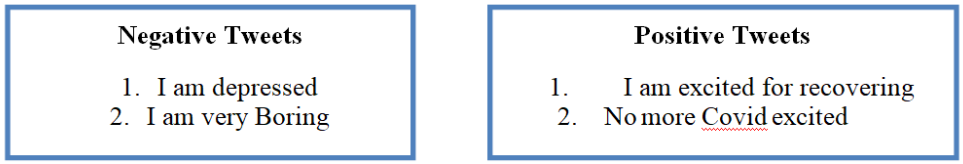

Figure 4. Sample of The Tweets Corpus

.

Table 1. The frequency table of Naïve Bayes algorithm

Frequency Table		
Words	**Positive**	**Negative**
I	1	2
am	1	2
excited	2	0
For	1	0
Recovering	1	0
Depressed	0	1
No	1	0
More	1	0
Covid	1	0
Very	0	1
Boring	0	1
Total	**9**	**7**

Naïve Bayes classifier is solved by using Bayes theorem:

P(A|B) = (P(B|A) * P(A)) / P(B)
P(A|B): The probability of event A given B (called posterior)
P(B|A): The probability of event B given A (called likelihood)
P(A): The probability of event A (called prior)
P(B): The probability of event B (called evidence)

We can apply it into the tweets predictions as the following:

P (Pos | "Recovering") = P ("Recovering" | Pos) * P(Pos) / P (Recovering)

The word "Recovering" is a positive sentiment? Is this statement correct?
= (1/9 * 9/16) / (1/16)
= (.11 *.56) / .062 = .99

Naive Bayes uses a similar method to predict the probability of different class (Negative, Neutral, Extremely Negative, Extremely Positive).

In this study we used the below code as shown in Figure 5 for defining the Naive Bayes classifier and fitting the training dataset as the below code:

```
[83]: from sklearn.naive_bayes import MultinomialNB
      classifier = MultinomialNB()

[84]: from sklearn.pipeline import Pipeline

      pipeline = Pipeline([
          ('bow', bow),  # strings to token integer counts
          ('tfidf', tfidf),  # integer counts to weighted TF-IDF scores
          ('classifier', classifier),  # train on TF-IDF vectors w/ Naive Bayes classifier
      ])

[86]: pipeline.fit(train_dataset['OriginalTweet'],train_dataset['Sentiment'])

[86]: Pipeline(steps=[('bow',
                       CountVectorizer(analyzer=<function preprocess at 0x000001FC33DE6C10>)),
                      ('tfidf', TfidfTransformer()),
                      ('classifier', MultinomialNB())])
```

Figure 5. Fitting the training dataset by using SVM

3.3.2. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm that could be applied in both the classification and regression analysis, however in this study we will use it in the classification model for predicting the students' feeling. The idea behind SVM is finding a hyper plane that can best divide our training dataset into different classes, which is known as (multiclass classification); Figure 6 illustrated the SVM technique.

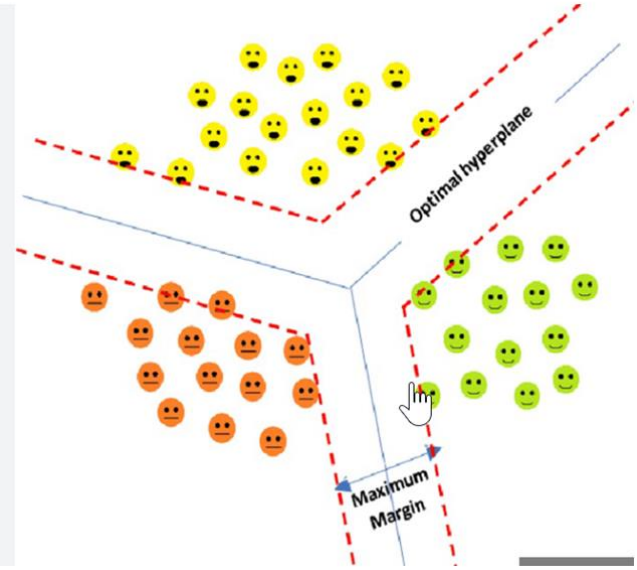

Figure 6. The SVM techniques

In this study we used the below code in Figure 7 for defining the SVM classifier and fitting the training dataset as the below code:

```
[23]:
from sklearn import svm
clf = svm.SVC()
classifier = clf

[24]:
from sklearn.pipeline import Pipeline

pipeline = Pipeline([
    ('bow', bow),  # strings to token integer counts
    ('tfidf', tfidf),  # integer counts to weighted TF-IDF scores
    ('classifier', classifier),  # train on TF-IDF vectors w/ Naive Bayes classifier
])

[25]:
pipeline.fit(train_dataset['OriginalTweet'],train_dataset['Sentiment'])

[25]:
Pipeline(steps=[('bow',
                 CountVectorizer(analyzer=<function preprocess at 0x000001A929AC5700>)),
                ('tfidf', TfidfTransformer()), ('classifier', SVC())])
```

Figure 7. Fitting the training dataset by using the models

3.4. Testing the Machine Learning Algorithms

There are four effective measures applied in this study, which all are based on confusion matrix output which are (True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN)). Machine learning prediction depends on the following formulas of the prediction scores:

Precision(P) = TP/(TP+FP)
Recall(R) = TP/(TP+FN)
Accuracy(A) = (TP+TN)/ (TP + TN + FP + FN)
F-Measure (Micro-averaging) = 2. (P.R)/(P+R)

4. RESULTS AND DISCUSSIONS

The results reveal high performance in the Support Vector Machine (SVM) model accuracy compared to Naïve Bayes algorithm, as it shown in Table 2 and Figure 8. though, the accuracy factor is very vital in terms of evaluating the Machine Learning model and it can increase the credibility to any algorithm.

In this study, we tried to improve the performance by changing the test split ratio as the following 10%, 20% and 30%. However, table 2 shows the results of the accuracy into the two algorithms.

Table 2. The accuracy results of SVM and Naïve Bayes models in changing the test split ratios

Machine Learning Model	**Test Split Ratio Accuracy**		
	10%	20%	30%
Naïve Bayes	37%	36%	35%
SVM	97%	97%	97%

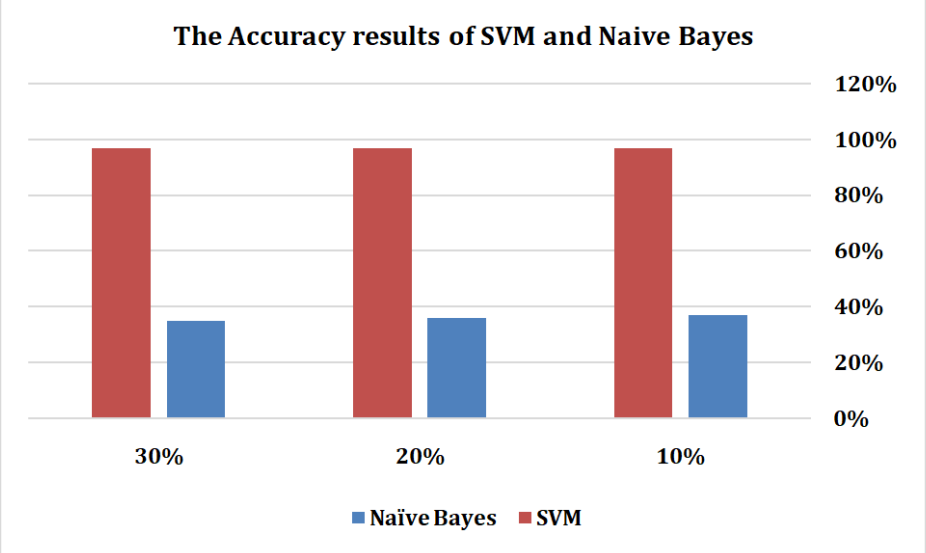

Figure 8. The Accuracy results of SVM and naïve Bayes

Furthermore, the models show more accurate when decreasing the test-split ratios in training the datasets as it shows in Table 3.

Table 3. The training speed per minutes in SVM and Naïve Bayes classifier

Machine Learning Model	The speed per minutes / Split Ratios		
	10%	20%	30%
Naïve Bayes	3 Minutes	2 Minutes	2 Minutes
SVM	40 Minutes	35 Minutes	35 Minutes

Moreover, the study reveals that training speed in the SVM classifier is relatively slow comparing to Naïve Bayes classifier. Table 3 shows the training speed per minutes for both classifiers and by using different test split ratios. Figure 9 shows the chatbot predicting the students' emotions from the texts.

```
Chatbot: Hello ,,I can predict your emotions in Covid! How do you feel Today?
 I like working from home
I can understand your question, you feel  'Positive']

Chatbot: Hello ,,I can predict your emotions in Covid! How do you feel Today?
 Feel Sick and having Fever
I can understand your question, you feel  ['Negative']

Chatbot: Hello ,,I can predict your emotions in Covid! How do you feel Today?
 Today I am much better than Yesterday
        Great! I can understand your question, you feel  ['Positive']

from sklearn.metrics import classification_report
all_predictions = pipeline.predict(test_dataset['OriginalTweet'])
print(classification_report(test_dataset['Sentiment'], all_predictions))
```

Figure 9. Chatbot -predicting the students'emotions from the texts

5. CONCLUSIONS AND FUTURE WORK

Anxiety and depression are among the main symptoms that have impacts on students' achievement and success, however, computer scientist and psychologist are inspired to work together in order to study the social media interactions from students and parents' comments that reflected their feelings. For example, during the Covid-19 pandemic, some students expressed their experiences during the pandemic as they depressed and worried to get sick, others had sleeping problems as well as mental health symptoms from the isolation and social restriction. Therefore, developing a chatbot that can detect students' emotions from the tweets messages by using the opinion mining or the sentiment analyses techniques can add value to the decision makers from educational and other sectors to enhance and improve the students wellbeing services. However, the SVM and Naïve Bayes are the two machine learning algorithms that are performed in this chatbot, TFIDF transformed is used in this model to improve the accuracy performance. The results revealed that the accuracy increased when decreasing test split ratios. In addition, the results showed a high performance in (SVM) model accuracy compared to NB model. Moreover, the study revealed that training speed varied in both models, since the speed of SVM classifier is extremely slow even though it is more accurate classifier.

In future, other neural network algorithms such as the CNN, LSTM will be implemented. In addition to that, we need to do further studies in order to study the factors that affect the speed of the SVM classifier without affecting the accuracy. Lastly, this study only focused on the English tweets, we will improve it by including other languages such as the Arabic language.

REFERENCES

[1] Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta bio-medica: Atenei Parmensis.2020; 91(1):157–160

[2] Akat, M., & Karataş, K. (2020). Psychological effects of COVID-19 pandemic on society and its reflections on education. *Electronic Turkish Studies*, *15*(4).

[3] Kabir, H., Nasrullah, S. M., Hasan, M. K., Ahmed, S., Hawlader, M. D. H., & Mitra, D. K. (2021). Perceived e-learning stress as an independent predictor of e-learning readiness: Results from a nationwide survey in Bangladesh. *PloS one*, *16*(10), e0259281.

[4] Gavurova, B., Ivankova, V., & Rigelsky, M. (2020). Relationships between perceived stress, depression and alcohol use disorders in university students during the COVID-19 pandemic: a socio-economic dimension. *International journal of environmental research and public health*, *17*(23), 8853.

[5] Pelucio, L., Simões, P., Dourado, M. C. N., Quagliato, L. A., & Nardi, A. E. (2022). Depression and anxiety among online learning students during the COVID-19 pandemic: a cross-sectional survey in Rio de Janeiro, Brazil. *BMC psychology*, *10*(1), 1-8

[6] Huang, Z., Tay, E., Wee, D., Guo, H., Lim, H. Y. F., & Chow, A. (2022). Public Perception of the Use of Digital Contact-Tracing Tools After the COVID-19 Lockdown: Sentiment Analysis and Opinion Mining. *JMIR Formative Research*, *6*(3), e33314.

[7] Bird, J. J., Ekárt, A., & Faria, D. R. (2021). Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, 1-16.

[8] Assayed, S. K., Shaalan, K., & Alkhatib, M. (2022). A Chatbot Intent Classifier for Supporting High School Students. *EAI Endorsed Transactions on Scalable Information Systems*, *10*(3).

[9] Liu, Y., Whitfield, C., Zhang, T., Hauser, A., Reynolds, T. and Anwar, M., 2021. Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems*, *9*(1), pp.1-16.

[10] Medhat, W.; Hassan, A.; Korashy, H. Sentiment Analysis Algorithms and Applications: A Survey. Ain Shams Eng. J. 2014, 5,1093–1113. [CrossRef]

[11] Rani, S., & Singh, J. (2017). Sentiment analysis of Tweets using support vector machine. *Int. J. Comput. Sci. Mob. Appl*, *5*(10), 83-91.

[12] Alabid, N. N., & Katheeth, Z. D. (2021). Sentiment analysis of twitter posts related to the covid-19 vaccines. *Indonesian Journal of Electrical Engineering and Computer Science*, *24*(3), 1727-1734.

[13] Naw, N. (2018). Twitter sentiment analysis using support vector machine and K-NN classifiers. *IJSRP*, *8*, 407-411.

[14] Ouerhani, N., Maalel, A., Ghézala, H. B., & Chouri, S. (2020). Smart ubiquitous chatbot for COVID-19 assistance with deep learning sentiment analysis model during and after quarantine

[15] Fauziah, Y., Saifullah, S. and Aribowo, A.S., 2020, October. Design Text Mining for Anxiety Detection using Machine Learning based-on Social Media Data during COVID-19 pandemic. In *Proceeding of LPPM UPN "Veteran" Yogyakarta Conference Series 2020–Engineering and Science Series* (Vol. 1, No. 1, pp. 253-261).

[16] Ryu, J., Sükei, E., Norbury, A., Liu, S.H., Campaña-Montes, J.J., Baca-Garcia, E., Artés, A. and Perez-Rodriguez, M.M., 2021. Shift in Social Media App Usage During COVID-19 Lockdown and Clinical Anxiety Symptoms: Machine Learning–Based Ecological Momentary Assessment Study. *JMIR mental health*, *8*(9), p.e30833.

[17] Vahedi, B., Karimzadeh, M. and Zoraghein, H., 2021. Predicting county-level COVID-19 cases using spatiotemporal machine learning: Modeling human interactions using social media and cell-phone data.

[18] Chin, H., Lima, G., Shin, M., Zhunis, A., Cha, C., Choi, J., & Cha, M. (2022). User-Chatbot Conversations During the COVID-19 Pandemic: A Study Based on Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research*.

[19] Yao, Z., Yang, J., Liu, J., Keith, M., & Guan, C. (2021). Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19. *Cities*, *116*, 103273.

[20] Kwok, S.W.H., Vadde, S.K. and Wang, G., 2021. Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: Machine learning analysis. *Journal of medical Internet research*, *23*(5), p.e26953.

[21] Wibowo, D.A. and Musdholifah, A., 2021, December. Sentiments Analysis of Indonesian Tweet About Covid-19 Vaccine Using Support Vector Machine and Fasttext Embedding. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 184-188). IEEE.