

An Optimal Machine Learning Approach for Interpretation of Emotions and Prediction of School Students' Attitude towards Online Education as a Post-Effect of Covid-19 in India

Kavita Pabreja¹, Kanika Mittal² and Himanshu Jinjharwal³

¹Associate Professor, Department of Information Technology and Systems, Lal Bahadur Shastri Institute of Management, New Delhi, India

Email: kavitapabreja@lbsim.ac.in

²⁻³Department of Computer Applications, Maharaja Surajmal Institute, an affiliate of GGSIPU, New Delhi, India

Email: {kanika00121202021, himanshu00221202021}@msijanakupuri.com

Abstract—The Indian education system was severely impacted during COVID-19 because of the unpreparedness of the entire education arrangements to manage the studies in an online mode. The students particularly senior secondary was deeply shocked when the schools closed down unprecedentedly and efforts were made by the schools to conduct classes and examinations in an online mode. This study aims to analyze the students' attitude towards the teaching-learning process considering the availability of digital devices and Internet facilities, before, during, and post covid times. Data from 113 school students have been collected and it comprises twenty independent features and one dependent feature. Exploratory data analysis followed by the application of a Support Vector Machine has been implemented to predict the dependent feature viz. the choice to adopt the online mode of learning. A grid search has been used to tune the threshold and locate the optimal value of the threshold for the ROC Curve. Python scikit-learn's univariate feature selection methods to select the top five features has resulted in the selection of just five independent features that are capable enough to predict the mentioned dependent features with an accuracy of 0.74 to predict the choice to adapt the online mode of learning with Support Vector machine learning classifier. The sentiments and emotions of the students pre-COVID, during-COVID and post-COVID have also been extracted.

Index Terms— COVID-19, Support Vector Classifier, Receiver Operating Characteristic Curve, Feature Selection, Natural language processing and teaching-learning process.

I. INTRODUCTION

Globally, as of November 2023, COVID-19 has impacted over 772 million individuals, including 6.98 million deaths, as reported by WHO [1]. This pandemic has harmed every industry in India, including corporate and educational sectors. Following the discovery of a growing number of Corona cases throughout India, the central and state governments decided to close colleges, schools, and universities. All exams, including entrance exams, for schools, colleges, and universities were postponed indefinitely, and classes were suspended. Every student's schedule was ruined as a result of the lockdown. In India, all academic pursuits came to an end

when about 32 crore students dropped out of their existing institutions[2]. Despite being a unique scenario in educational history, COVID-19 has made it possible to transition from the traditional, demanding classroom teaching style to a new, digital one [3].

The COVID-19 pandemic has demonstrated to us the inevitable nature of change. It has encouraged academic institutions to create and select innovative platforms and techniques. The education industry has been striving to overcome the pandemic threat by adopting a new strategy and digitizing the challenges.

This study aims to understand the attitude of school students towards teaching-learning pre-COVID, during COVID, and post-COVID. It also analyzes the students' emotions and feelings towards online education. There is objective as well as subjective questions being asked as part of the survey where data has been collected from 113 students studying in classes 9 to 12. Artificial Intelligence and Machine Learning techniques have been applied for exploratory and predictive modeling to get deeper insights into students' emotions and feelings towards online education.

The paper is structured as follows. The next section explains the background study followed by methodology which includes Data collection and data pre-processing, Exploratory Data Analysis, and Machine Learning algorithm application. Following this, the Results of the application of the Linear SVC model on the whole dataset are discussed along with the Role of Feature Selection. After this section, the Analysis of subjective answers is mentioned, and at the end, the Conclusions are written.

II. RELATED WORK

The coronavirus disease COVID-19 was reported to be spreading in Wuhan, China, towards the end of 2019[4]. The World Health Organisation (WHO) designated it a pandemic on March 11, 2020, due to the rapid spread of COVID-19[5]. The majority of governments imposed lockdowns, enforced physical and social seclusion, and imposed immigration restrictions in an effort to stem the COVID-19 pandemic. Almost 600 million students worldwide were impacted by the closure of educational institutions[6].

The COVID-19 epidemic presents a variety of ecological, technological, and psychological challenges for schoolchildren. During the COVID-19 lockdown, the whole education system—from elementary to tertiary—collapsed in India and around the world[7]. At the same time[8], it was found that the students' expressed positive attitudes about online distance learning are evident, and this suggests that they are more inclined to accept it as a viable educational option. Most of them liked to actively use ICT in their studies in addition to using it in their daily lives. It is possible to draw the conclusion that the primary goals of students are associated with the use of an efficient virtual learning environment that incorporates integrated technology to facilitate online discussion amongst participants, online assignment submission, and online teacher help [9].

In another study on Bangladesh students, the findings indicate that the majority of students have a favourable attitude towards online learning. However, because of the Covid-19 epidemic, online learning is the most effective way to learn during a lockdown. More in-depth statistical research on the benefits of online learning under lockdown may be conducted [10].

The authors found that in industrialised nations like Indonesia, where the majority of students lack internet connection due to both financial and technological difficulties, online learning is unable to yield the intended effects. The usefulness of online learning compared to traditional classroom settings was the focus of this study, particularly for college students. According to their survey, 73% of students had access to a computer with internet, and 71.4% of students believed they were competent to operate a computer or laptop [11].

According to a study on undergraduate Indian students, Sixty-nine percent of the students believed that they learned better in actual classrooms and by participating in MOOCs (39.9%) as opposed to online learning. However, the students thought that the instructors' online teaching abilities had improved with time. of the pandemic (68.1%) and the current usefulness of online learning (77.9%). The pupils appreciated the internet resources and tools being used to support virtual learning. Nonetheless, the pupils believed that learning online is demanding and impacting their social and physical well-being. A widespread adoption has resulted from this pandemic. of online learning, and the knowledge we gain today will be useful later [12].

In Kurdish EFL university, the majority of students, have a bad opinion of online learning. In a similar vein, they thought that education received on campus was more beneficial. The results also showed that the vast majority of students said they were not satisfied with their online education. Last but not least, in addition to internal obstacles like time management, concentration problems, and opening cameras because of social issues, the students also mentioned external factors like electricity and inconsistent internet connections as challenges encountered throughout their online education [13].

These studies are in general addressing the online education but our research is based not only on the attitudes and emotions of students for online education but also about application of machine learning techniques for extraction of important factors.

III. METHODOLOGY

A large number of relevant research papers have been studied and an appropriate questionnaire was designed that comprises the following categories: -

1. Demographics - Three questions
2. Teaching-related questions -Four questions
3. Learning-related questions – Eight questions
4. Feeling and emotions towards Online Education – Four questions
5. Resource Availability – Four questions
6. Subjective questions - Six questions
7. Output column- Would like to continue in online or offline mode

The methodology comprises data collection, data pre-processing, exploratory data analysis, application of machine learning algorithm, optimization of Machine Learning model performance by tuning the AUC (Area under ROC curve), dimensionality reduction by feature selection and finally applying ML techniques on a reduced dataset for prediction of output class (whether students would like to continue in online or offline mode in future). The answers to six subjective questions have been analyzed for their polarity (positive, negative, or neutral) and emotions for these answers have been plotted.

We have used Python's libraries viz. pandas, numpy, seaborn, matplotlib, nltk.corpus, nltk.stem.porter, vaderSentiment.vaderSentiment, transformers, sklearn and hugging face library for this research paper.

A. Data collection and data pre-processing

Non-probability convenience sampling technique has been used for selecting the respondents from a pool of school students of Delhi state. We collected data from 113 students from three different schools who were studying in classes 9 to 12 during April 2023. The demographics have been explored in the next section. The data was complete as we had all questions marked as compulsory. We have asked closed-ended questions for the categories viz. Teaching-related, Learning-related, Feeling and emotions towards Online Education, Resource Availability, and Output column (Would like to continue in online or offline mode). The responses to these are on a Likert scale and for our machine learning algorithms to work on these responses, we have done label encoding.

B. Exploratory Data Analysis

The descriptive data analysis has been performed to understand the distribution of data corresponding to various response categories and also to find the patterns that exist within our collected data. Below are the findings: -

- The focus was on the students of standard 9 to 12, i.e., the age group of 13 to 19 years. The dataset reveals that there is almost equal participation of males and females. Also, 37.2% of students are studying in class 9, 22.0% in class 10, 17.7% in class 11 and 23.0% in class 12.
- The majority of the respondents, approximately 50%, belong to the 14 to 15 years of age group.
- Corresponding to the Teaching related questions, 72.6% of the respondents have the viewpoint that during pre-COVID faculty was teaching more effectively; 61% felt they had good interaction with the teachers and 55% felt that Overall teaching was effective. However, approximately 40% of the students were not happy with the resources offered by the school to learn from home.
- Corresponding to Learning related questions, it is observed that students were happier with the learning methodology during offline mode (followed during the pre-COVID time). Before COVID time; 55% to 80% of the respondents felt that their "Understanding level was high", "Ability to attend classes was high", "Higher interest in studies", "Ability to study more attentively".
- Also, 84% felt that the Offline mode of learning is more effective and they gained more knowledge in the Offline mode.
- In approximately 50% of cases, time management was difficult and approximately 70% of the respondents spent 1 to 5 hours per day in an online mode of education.
- For the emotions and feelings of the students towards online education, it was found that more than 60% of the students felt that face-to-face contact with the teacher is necessary for effective learning to

take place, there is a lot of difference in feeling and efforts required to produce same results in online mode compared as it was in offline mode. The same is reflected in the graphs.

- Since the students being surveyed are from private schools and belong to middle or upper-middle-income groups, 73.5% had access to resources needed for online study at home, and used mobile phones, laptops, desktops, and tablets. Though in 56% of the cases, these devices were shared with siblings/ parents they were available for use. In more than 50% of the cases, there were issues related to Internet services.
- It was found that 70.8% of the students did not like to continue with the online mode of learning and the rest were willing to continue in the online digital mode only.

C. Machine Learning Algorithm Application For Prediction Of Students' Choice To Continue In Online Or Offline Mode Of Education

A Support Vector Classifier (SVC), also known as a Support Vector Machine (SVM) for classification, is a supervised machine learning algorithm that is used for binary classification tasks[14], [15]. Its primary goal is to find the optimal hyperplane that best separates data points belonging to two different classes in a way that maximizes the margin between the classes. In this context, a hyperplane is a decision boundary that helps classify data points into one of two classes.

We have performed label encoding corresponding to all input features and target columns to convert all categorical variables into numerical format. It is particularly required when working with an SVC algorithm as it can only operate on numerical data. All of the respondents correspond to the data points that are to be classified into either of the two output classes viz. "Want to continue learning in online mode" or "Do not want to continue learning in online mode". There are 22 questions and have answers in different types of Likert scales. We could gather responses from 113 school students. The reason for selecting SVC is that it is designed to work well in high-dimensional feature spaces. SVC can find complex decision boundaries, even in situations where the number of features is much larger than the number of records.

We have used the Radial basis function (RBF) kernel to transform the training set of data so that a non-linear decision surface can transform to a linear equation in a higher number of dimension spaces. RBF is more complex and efficient at the same time that it can combine multiple polynomial kernels multiple times of different degrees to project the non-linearly separable data into higher dimensional space so that it can be separable using a hyperplane.

IV. RESULTS OF APPLICATION OF LINEAR SVC MODEL ON WHOLE DATASET

We have predicted the choice of continuing in an online mode based on students' responses to 20 questions. The output classes are NO and YES. The model has demonstrated a reasonable accuracy of 0.85 but with Specificity 0.96 and Sensitivity 0.43; which means that the model's performance for predicting the negative class is poor. The Area under the ROC curve is 0.69 as shown in Fig.1, which demonstrates that the model is good enough at discriminating positive and negative classes.

To improve the sensitivity, we have done the tuning of the Area under ROC curve (AUC) to select the best threshold value where sensitivity and specificity both are at their best. We found that the best threshold value is 0.287813 where sensitivity and specificity – both are best i.e Accuracy 0.62, Specificity 0.59, and Sensitivity 0.71. The graph showing the tuning of AUC is shown in Fig. 2.

A. Role of Feature Selection

Eliminating redundant and irrelevant variables while keeping the most important ones boosts the prediction capacity of the algorithms. So, we have applied two popular feature selection techniques viz. mutual_info_classif and SelectKBest. Finally, five important features that are common in the results of these two techniques have been selected for further application of SVC again on the reduced dimensions dataset. These five features are as follows: - KnowledgeGain (Do you think you gained more knowledge because of the freedom of just joining class in online and utilizing your time to study in your own way?), effective_teaching (Faculty was teaching more effectively :Before corona/During corona/Aftercorona),

FaceToFace (I feel that face-to-face contact with my instructor is necessary for learning to occur), EffectiveModeOfLearning (For me the effective mode of learning was online or offline) and understandingLevel (My understanding level was high: Before corona/During corona/After corona)

After applying the SVC model with the reduced feature set, it was found that the AUC is 0.97 which is better than the earlier score with all twenty input features (as shown in Fig. 3). The Accuracy was 0.82, Specificity was

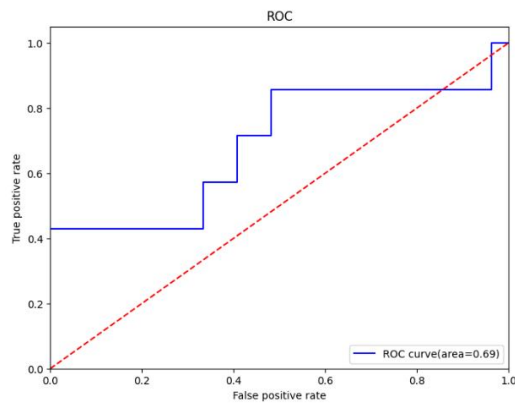


Figure 1. Receiver Operating Curve for test data with all twenty features

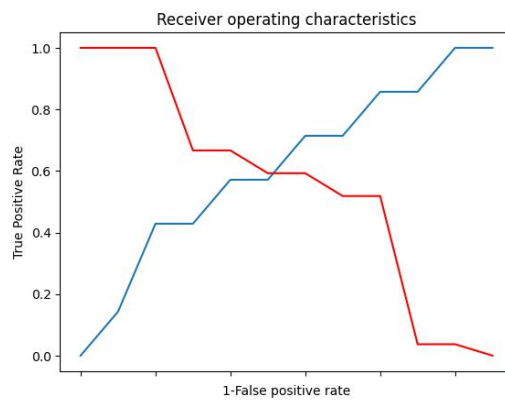


Figure 2. Tuning of ROC curve

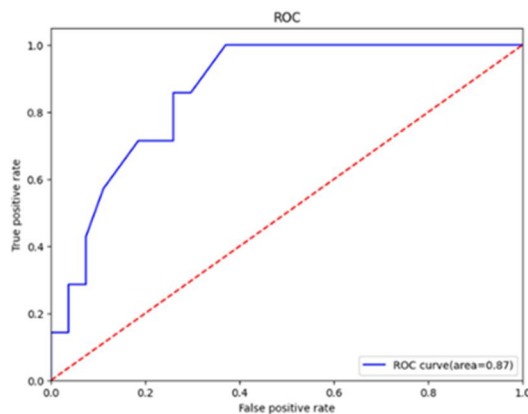


Figure 3. Receiver Operating Curve for test data with reduced set of features

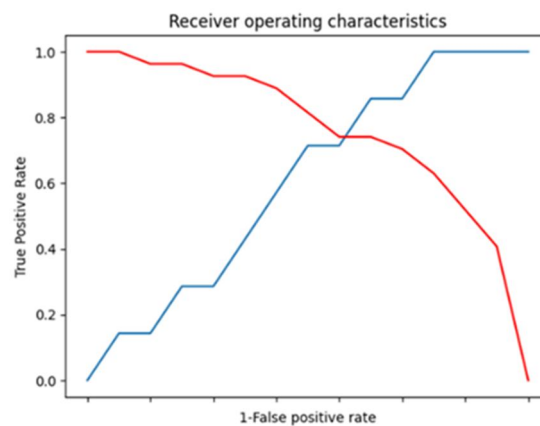


Figure 4. Tuning of ROC curve

0.96 and Sensitivity was 0.29. Again, the model was not performing well in predicting the negative class so AUC tuning was done, as shown in Fig. 4. It was found that the best threshold value is 0.335458 where sensitivity and specificity – both are the best. The performance metrics values are Accuracy 0.74, Specificity 0.74, and Sensitivity 0.71.

B. Analysis of subjective answers

We have used the Python toolkit, NLTK[16] designed for Natural language processing (NLP) operations. It gives us access to several test datasets through different text-processing frameworks. NLTK can be used for a wide range of operations, including tokenizing, parse tree visualization, and more.

We have also used a vocabulary and rule-based sentiment analysis tool called VADER (Valence Aware Dictionary and Sentiment Reasoner)[17] is particularly adapted to the sentiments expressed in social media and performs well on texts from other domains.

Corresponding to the answers to subjective questions, **the analysis of each question is described below: -**

“What has been the hardest part about completing your coursework during online classes?”

From the sentiment polarity pie chart, it is evident that many hurdles and negative emotions correspond to the difficulties faced by students during online classes. It was found that the students wrote about the Lack of teachers, and face-to-face interaction.

“What do you not like about your remote learning classes?”

The polarity of words indicates that 62.7% of the words had negative sentiment and included words like teacher, friends, face, lack, problem, and interaction.

“What do you like about your remote learning classes?”

The students have found some positive aspects of the online classes though they are not very happy with it. The polarity of words indicates that 47.1% of the words were positive and included words like attend, sit, classes,

time, and don't. This indicates that the freedom of not sitting and attending classes was liked by the students and they could manage their time on their own.

“If you were teaching an online class yourself, what is the one thing you would do to make it more engaging?”

The answer's polarity is 52.9% positive sentiment and students have mentioned that they would make the classes filled with fun and play and have more interaction.

“How did you feel when your normal classes resumed?”

The answer to this is 80.4% positive and students felt happy, excited, and good.

“How do you think your future goals got affected due to the impacts of the pandemic?”

56.4% of the words have the negative polarity of the sentiment and the students felt that the pandemic has affected their time, future, and goals.

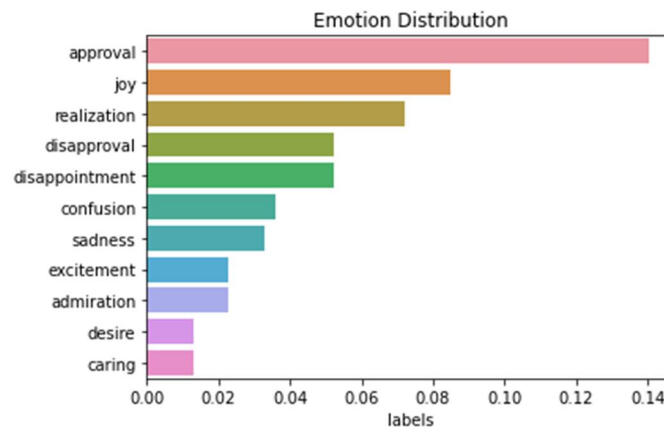


Figure 5. Emotions distribution corresponding to all subjective answers taken together

HuggingFace[18] is a huge open-source community that creates tools based on open-source code and technologies that let users create, train, and implement machine learning models. By using this library for emotion detection for all subjective questions' answers taken together, it was found that there is a significant presence of emotions of approval and joy as shown in Fig. 5.

V. CONCLUSION

The paper is based on real datasets collected through Google Forms. This paper is based on responses of the students to twenty objective questions related to Teaching and Learning, Feelings and emotions towards Online Education, and Resource Availability for online classes. The questions assess the respondents about their experience for the mentioned categories of questions pre-COVID, during-COVID, and post-COVID times. Based on responses to these questions, a Support Vector Classifier has been trained and tested to predict whether the students would like to continue studying in online or offline mode. The ML algorithm has established a reasonable accuracy of 0.85 but with a Specificity of 0.96 and Sensitivity of 0.43; which means that the model's performance for predicting the negative class is poor. So, its hyperparameters were tuned and the best value of the threshold value of the probability of a given class was obtained. We found that the best threshold value is 0.287813 where sensitivity and specificity – both are best i.e. Accuracy 0.62, Specificity 0.59, and Sensitivity 0.71. Next, the dimensionality reduction was addressed and the five most important features out of twenty features were selected that were able to perform classification well. Further, the AUC tuning was also done to derive the best results. It was found that finally, Accuracy was 0.74, Specificity was 0.74 and Sensitivity was 0.71. These results are better than those with all twenty features which means some redundant features caused the overfitting problem.

Another dimension of this paper is to extract the emotions and sentiments related to the teaching-learning process using sentiment analysis libraries by analyzing the subjective answers given by them. By using NLTK designed for Natural language processing (NLP), it was found that students found it hard to study in online mode and felt that their future goals were affected due to the impacts of the pandemic. At the same time, they had some innovative ideas in mind to make the online classes more interesting. Also, they felt excited and happy when the classes resumed in physical mode.

Hence, this study is a practical application of advanced techniques of machine learning and sentiment analysis to provide a deep insight into the emotions of the students pre, during, and post covid times, and to predict their interest in online or offline classes.

REFERENCES

- [1] W. C. (COVID-19), "Americas," pp. 1–4, 2023.
- [2] S. Kritika, "Report by UNESCO," *The Print*, p. 2, 2020.
- [3] N. Waral and C. S. Sajan, "Impact of COVID-19 on Education System in India," *J. Inf. Manag.*, vol. 7, no. 2, pp. 165–169, 2020, doi: 10.5958/2348-1773.2020.00019.3.
- [4] S. K. Upadhyay, R. Singh, B. Malik, and G. Kumar, "the Outbreak and Challenges of Novel Coronavirus (Covid-19): the Global Pandemic Emergency of Early 2K20 and Indian Scenario," *Int. J. Biol. Pharm. Allied Sci.*, vol. 9, no. 5, 2020, doi: 10.31032/ijbpas/2020/9.5.5126.
- [5] V. M. Cucinotta D., "WHO Declares COVID-19 a Pandemic," *Scand. J. Immunol.*, vol. 92, no. 3, pp. 9–10, 2020, doi: 10.1111/sji.12918.
- [6] N. Kapasia *et al.*, "Impact of lockdown on learning status of undergraduate and postgraduate students during COVID-19 pandemic in West Bengal, India," *Child. Youth Serv. Rev.*, vol. 116, no. June, p. 105194, 2020, doi: 10.1016/j.childyouth.2020.105194.
- [7] S. Garg, D. Aggarwal, S. K. Upadhyay, G. Kumar, and G. Singh, "Effect of Covid-19 on School Education System: Challenges and Opportunities To Adopt Online Teaching and Learning," *Humanit. Soc. Sci. Rev.*, vol. 8, no. 6, pp. 10–17, 2020, doi: 10.18510/hssr.2020.862.
- [8] G. Kumar, "Effect of COVID-19 on Education , Economy , and Social Behaviour in Indian perspective," no. May, 2023.
- [9] S. Kadry, "Stress causes and outcomes statistical analysis," *Biometrics Biostat. Int. J.*, vol. 7, no. 4, 2018, doi: 10.15406/bbij.2018.07.00229.
- [10] K. M. Hossain, "STUDENTS ' ATTITUDE TOWARDS ONLINE TEACHING LEARNING DURING COVID-19 SITUATION IN WEST BENGAL," vol. 10, no. 9, pp. 343–351, 2022.
- [11] D. S. Bestiantono, P. Z. R. Agustina, and T.-H. Cheng, "How Students' Perspectives about Online Learning Amid the COVID-19 Pandemic?," *Stud. Learn. Teach.*, vol. 1, no. 3, pp. 133–139, 2020, doi: 10.46627/silet.v1i3.46.
- [12] P. Chakraborty, P. Mittal, M. S. Gupta, S. Yadav, and A. Arora, "Human Behav and Emerg Tech - 2020 - Chakraborty - Opinion of students on online education during the COVID-19 pandemic.pdf," *Hum Behav & Emerg Tech*, vol. 3, pp. 357–365, 2021.
- [13] S. Hussein, H. Barzani, and R. J. Jamil, "Students' Perceptions towards Online Education during COVID-19 Pandemic: An Empirical Study," *Int. J. Soc. Sci. Educ. Stud.*, vol. 8, no. 2, pp. 28–38, 2021, doi: 10.23918/ijsses.v8i2p28.
- [14] Javatpoint, "About Support Vector Classifier," 2023.
- A. Saini, "Guide on Support Vector Machine (SVM) Algorithm," *Anal. Vidhya*, pp. 1–35, 2024, [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/>
- [15] "NLTK Search Documentation Natural Language Toolkit".
- [16] S. S. Pragnya, "Valence Aware Dictionary and Sentiment Reasoner (VADER)," *Mediu. Dly. Dig.*, pp. 1–5, 2022.
- [17] Y. Tang, A. Wu, and D. Okhonko, "Using pretrained models - Huggingface," *Hugging Face community*.