How to Cite:
Ajaz, F., Prince, P., & Seniaray, S. (2022). Sentiment classification and analysis of twitter data on distance learning using ML classifiers. *International Journal of Health Sciences*, *6*(S3), 5750–5759. https://doi.org/10.53730/ijhs.v6nS3.7231

Sentiment classification and analysis of twitter data on distance learning using ML classifiers

Farheen Ajaz
Department of Applied Mathematics, Delhi Technological University, New Delhi 110042, India
*Corresponding author email: farheenhanna92@gmail.com

Prince
Department of Applied Mathematics, Delhi Technological University, New Delhi 110042, India
Email: sharma18.prince@gmail.com

Sumedha Seniaray
Department of Applied Mathematics, Delhi Technological University, New Delhi 110042, India
Email: sumedhaeniaray@dtu.ac.in

***Abstract*---**Online education comes in shades of grey. We analyze public view on the pandemic in regards with mental stress, interrupted power supply, affordability and access to internet, flexibility of schedule, reduction of long-distance commute, risk of covid and other such consequences of online learning and Government's take on the need for inclusive education policies. In this paper, we qualitatively inspect the consequences of COVID-19 pandemic on education of the students. This study primarily focuses on the response of students of all age groups, educators, college professors, school teachers and also parents of young students towards the approach of distance learning or Online education in the past two years. We have taken two datasets, first being the Twitter dataset comprising of tweets from around the whole world and second, dataset which is specific to tweets from India. The data has been extracted from twitter with the aid of twitter API and then two sentiment analysis approaches have been implemented, first Machine learning classifiers namely, Naïve Bayes, SVM, Random Forest, Logistic Regression, KNN, XG-Boost and secondly, Lexicon Based algorithms, VADER and TEXTBLOB. Upon performing the said approaches, the maximum accuracy achieved is 94%.

***Keywords*---**COVID-19, distance learning, online education, twitter.

Introduction

The novel Coronavirus (COVID-19) is a communicable ailment which is caused by the SARS-CoV-2 virus. Mild to moderate symptoms are experienced by people who fall victim to COVID-19, and they usually recover without any kind of special treatment. Although, for some people it becomes fatal and leads to the requirement of immediate medical attention and therefore this disease should not be taken lightly by any means. The first case of novel corona virus was encountered in Wuhan, China and within a span of three months it spread all across the world. COVID-19 is not only a global pandemic and public-health emergency; It has also had a significant impact on the global economy and financial industry. Significant reductions in income, an upsurge in unemployment, and disturbances in the transportation services and manufacturing industries are among the consequences of the disease vindication measures that have been executed in many countries.[1]

Along with these impacts, one of the biggest hit sectors is the sector of education. The Government all over the world has been encouraging online mode to impart education ever since the beginning of COVID-19, in order to achieve academic steadiness. These measures need to be taken to protect children from COVID-19 as well as to ensure that their studies remain unhampered. While few educational establishments find online education a staggering task, many have already made the switch from offline to online mode through platforms like Microsoft teams, Google meet, Zoom etc. There are various challenges of online education which occur at different levels. The first and the foremost being reliable internet access. According to a study conducted by the National Center for Education Statistics that over twenty three percent Indian students, including those from urban areas, do not have access to the internet for online studies while 6 percent of them have access only through smartphone.

The next prime challenge is the consistency of the students. Change has always been tough to embrace, let alone adapting oneself amidst a pandemic. It is hard to move from an organized day at school to a completely unstructured routine at home. It mostly affects younger children since for them, it is challenging to sit still in front of a camera to watch a lesson their teacher recorded or taught live online. It makes them miss important concepts from the class, mainly the reason being distraction. Through this research we aim to establish what is the common man's opinion concerning the online mode of education. Although Online education has been present in the form of distance learning (which is offered by institutes like IGNOU) as well for past few years but now it has become predominant and essential since the beginning of the pandemic. There are various machine learning algorithms generally used for sentiment analysis namely Convolutional Neural Network (CNN), Naïve Bayes, Support Vector Machine (SVM), Recurrent Neural Network (RNN), etc.

Naive Bayes is an elementary collection of algorithms of probabilistic kind which, for sentiment analysis classification, assigns to every word, a probability that the given word or phrase should be accepted as positive or negative. Sentiment analysis, also popularly known as opinion mining, is an approach to natural language processing that identifies the emotional tone behind a given sentence or

collection of words. It is a common technique utilized by various businesses to determine peoples' sentiments regarding any new product, service or concept. This method makes use of artificial intelligence and machine learning simultaneously to extract subjective information for text mining. As discussed in [2], E-Learning is an approach which combines distance learning and education within the classroom. Using modern means of communication, like smartphones, internet, graphics, and digital libraries, the motive of E-learning is to reach the student in the shortest time, with minimum effort and producing maximum benefit. It is expected that this pattern of imparting education will prevail in most teaching institutions around the world in near future.

Related work

Malak Aljabri et al. [3] proposed a model to analyze the distance learning in Saudi Arabia with the help of Twitter Dataset during COVID-19 pandemic period. All the tweets collected are written in Arabic language and geographic location is set to Saudi Arabia. The tweets are classified into the educational stages i.e., primary school and kindergarten, intermediate and high schools, or university. Then 6 Machine Learning Algorithms (LR, NB, KNN, XGB, SVM, RF) done on the dataset. The Logistic Regression ML technique has achieved the best accuracy of 89.9%. Filiz Angay Kutluk et al. [4] has proposed a model to analyze the student's satisfaction on distance education. In the model, Dataset is collected with the help of first degree data collection method from the two universities of Turkey, and with the help of T-test and One Way Anova the analysis is performed on it.

Usame Omer OSMANOGLU et al. [5] has proposed a model to measure the satisfaction for distance education course materials of Anadolu University. Overall, 6059 feedbacks were received, scaling was processed with the help of the triple Likert method and finally, Machine learning techniques were performed on the dataset. In which, 77.5% accuracy was achieved using Logistic Regression algorithm. Nimasha Arambepola [6] has proposed a model to analyse the effectiveness of the distance learning. Twitter Dataset is collected consisted 202,645 tweets during the lockdown period during the pandemic. The model gives out the result of 54% sentiments as positive, while 30% tweets are negative and 16% tweets comes out to be neutral. Imatitikua D. Aiyanyo et al. [7] has proposed a model to analyze the impact of COVID-19 pandemic on the educational sector of South Korea. Twitter dataset is collected during the COVID-19 period with the geographical boundaries set to South Korea. Pretrained sentiment Textblob and VADER were used and the result are analysed comparatively.

Methodology

The methodology can be broadly classified in three phases, namely Data collection and preprocessing, Tweet Classification and sentiment analysis using machine learning classifiers.

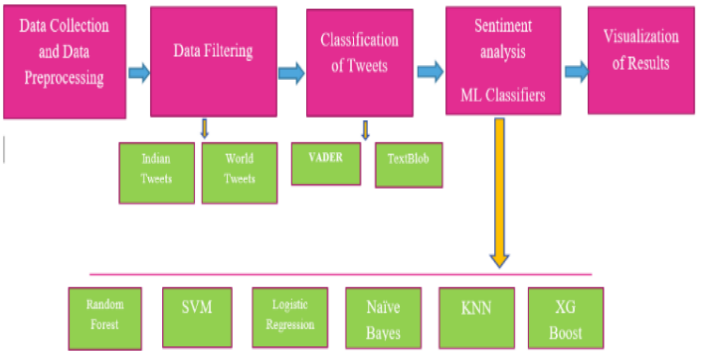


Figure 1. Represents the logical flow of our approach

Phase I
Data Collection

The twitter API allows a user to stream and download real time tweets. The process involves setting up a twitter developer platform. A user may apply for the same and the confirmation comes within a week. After that the user is provided with two keys including Consumer key and a secret key which has to be incorporated in the python program in order to stream and download real time tweets. In this paper we have extracted tweets in two phases:

- Extraction of tweets using hashtags
  The first phase of data collection was carried out by inputting particular hashtags in the code. These hashtags include 'Elearning', 'distanceeducation', 'googleclassroom', 'onlineclass', 'onlineexam', 'openschool', 'onlinelearning', 'eschool' and so on.
- Extraction of tweets using keywords
  The next phase is carried out in the same way as the phase 1 but the only difference is that the tweets are extracted by inserting a list of keywords in the search query instead of particular hashtags. These keywords are similar to the hashtags used in phase 1 of data collection. The tweets post extraction are represented in the Figure below

{"created_at": "Sat Nov 20 11:40:52 +0000 2021", "id": 1462023113985075208, "id_str": "1462023113985075208", "full_text": "RT @NehaSahni810: @PriyaMaryMathew @IvyRoseMathew Change is the END Result of all true learning.\n\n#Amityglobalconference #AmityOnline #AGOE\u2026", "truncated": false, "display_text_range": [0, 140], "entities": {"hashtags": [{"text": "Amityglobalconference", "indices": [98, 120]}, {"text": "AmityOnline", "indices": [121, 133]}], "symbols": [], "user_mentions": [{"screen_name": "NehaSahni810", "name": "Neha Sharma Sahni", "id": 2408521946, "id_str": "2408521946", "indices": [3, 16]}, {"screen_name": "PriyaMaryMathew", "name": "Dr Priya Mary Mathew", "id": 72035535, "id_str": "72035535", "indices": [18, 34]}, {"screen_name": "IvyRoseMathew", "name": "Ivy Rose Mathew", "id": 2406850579, "id_str": "2406850579", "indices": [35, 49]}], "urls": []}, "metadata": {"iso_language_code": "en", "result_type": "recent"}, "source": "Twitter Web App", "in_reply_to_status_id": null, "in_reply_to_status_id_str": null, "in_reply_to_user_id": null, "in_reply_to_user_id_str": null, "in_reply_to_screen_name": null,

Figure 2

Segregation of Indian tweets from original dataset

In this proposed approach, we have focused on the sentiment analysis of twitter data from India. Since most of the tweets have location mentioned in the description, we have made use of the location to filter the Indian tweets from the

world tweets dataset. Apart from this method, Geolocation has also been incorporated to directly extract the tweets specific to India. The geolocation aids in sieving tweets as per the wish of the user. It works by adding the 'geocode' parameter in the search query. The geocode parameter expects three values namely the latitude, longitude and the radius (which can either be in kilometers or miles). For example, if a user wants tweets from Delhi within 10 miles the code is illustrated below

```
new_tweets = api.search(q=search_query,
    count=tweetsPerQuery,
geocode=['28.5120' '77.3290' '10mi'],
lang="en",tweet_mode='extended')
```

Data preprocessing

The extracted tweets contain a lot of gibberish which is of no utility to the user. Further the tweets are in the form of a large paragraph filled with URLs and other such attributes which need to be eliminated for efficient of the text analysis. Following are the attributes of the extracted tweets.

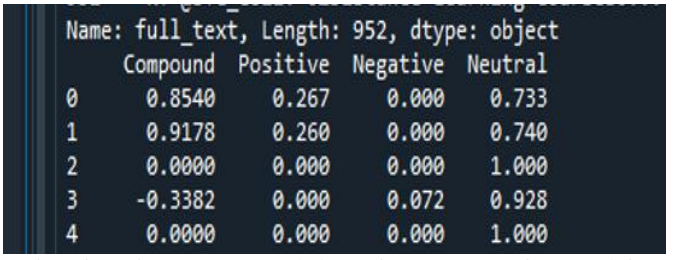
Here, created at: determines the date and time of the tweet creation.
Data obtained from twitter usually contains a lot of HTML entities, for example < > & which get embedded in the original data. As a result, it is vital to eliminate these entities. This process is collectively known as Data cleaning or Data preprocessing. Significant steps for Data cleaning:

- Removal of duplicate tweets: For this, the tweets were extracted at the gap of at least 7-10 days.
- Removing unwanted URLs
- Inserting the key attributes of the tweet in a CSV file
- Conversion of the CSV file to a dataframe in python

Phase II
Sentiment classification
VADER

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based tool for performing sentiment analysis that is specifically dedicated to sentiments expressed on social media. For the implementation of VADER (lexicon-based model for performing sentiment analysis), which has already been discussed formerly, we added classification attributes to the dataframe comprising the tweets. These attributes determine the polarity of the tweet and

assign sentiment score to every tweet. The columns are 'positive', 'negative', 'neutral' and 'compound' which signify the score as the name of the column suggests. VADER is implemented by incorporating the SentimentIntensityAnalyzer from the class Vadersentiment. Following the prescribed method, we have obtained positive, negative, neutral and compound score of corresponding to every tweet. Based on the compound score the tweet can be classified as positive, negative or neutral.



Name: full_text, Length: 952, dtype: object

	Compound	Positive	Negative	Neutral
0	0.8540	0.267	0.000	0.733
1	0.9178	0.260	0.000	0.740
2	0.0000	0.000	0.000	1.000
3	-0.3382	0.000	0.072	0.928
4	0.0000	0.000	0.000	1.000

Fig 3. Illustrates the dataframe comprising the tweets along with sentiment score.

Textblob

TextBlob is a very useful Python library which is used to process textual data. Textblob helps API to do the common natural processing tasks. With the help of Textblob, the following tasks were performed: Noun phrase extraction, Part of speech tagging, Sentiment Analysis, Classification, Tokenization.

Phase III
Machine Learning Techniques and Sentimental Analysis

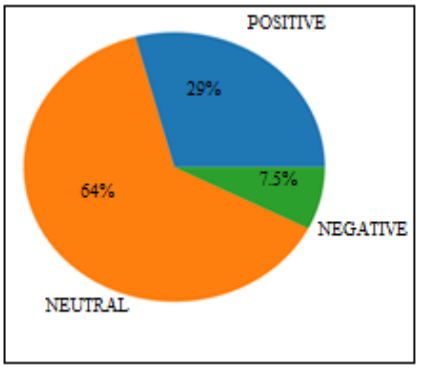
Machine learning is a data analytic technique which teaches computers to think like humans, means to approach to a problem from its experience. Machine learning algorithms uses computational methods to learn from data.Usually there are 3 basic approaches of Machine learning: Supervised learning, Unsupervised learning and Semi-supervised learning. In this model, we apply supervised machine learning algorithms for the sentiment analysis i.e., Naïve bayes, Support Vector Machine, Logistic Regression, XG-Boost, Random Forest.

Results

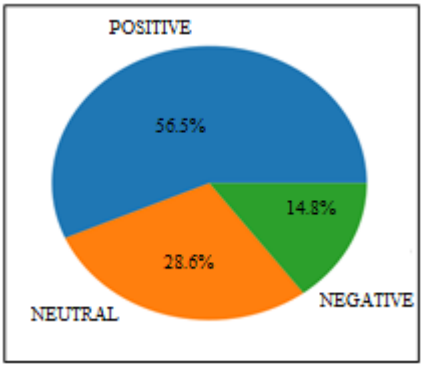
Sentiment Classification
Vader

For world tweets, we obtain that 29% are positive, 7.5% negative and 64% neutral tweets.

	Positive	***Negative***	***Neutral***
World	29%	7.5%	64%
India	56.5%	14.8%	28.6%

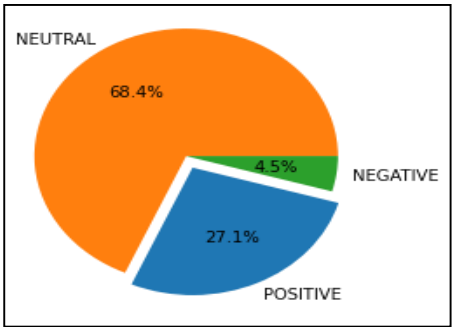


For Indian tweets we obtain 56% positive, 14.8% negative and 28.6% are neutral tweets.

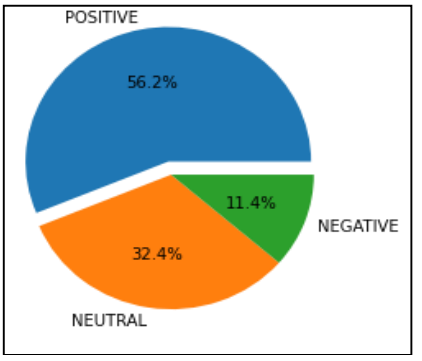


Textblob

With the help of pretrained sentiment Textblob. For world tweets, we obtain that 27.1% are positive, 4.5% negative and 68.4% neutral tweets.



For Indian tweets, we obtain 56.2% positive, 11.4% negative and 32.4% neutral tweets.



	Positive	Negative	Neutral
World	27.1%	4.5%	68.4%
India	56.2%	11.4%	32.4%

With the help of TextBlob, figure 4 illustrated below are the Top 10 positive hashtags

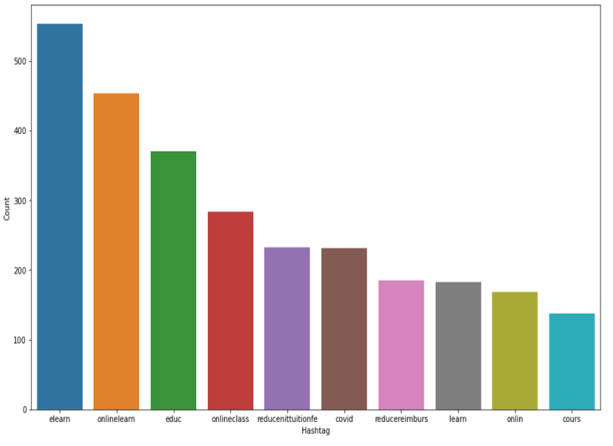


Figure 4

A Wordcloud is illustrated in figure 5 which displays the most prominent words in the dataset.

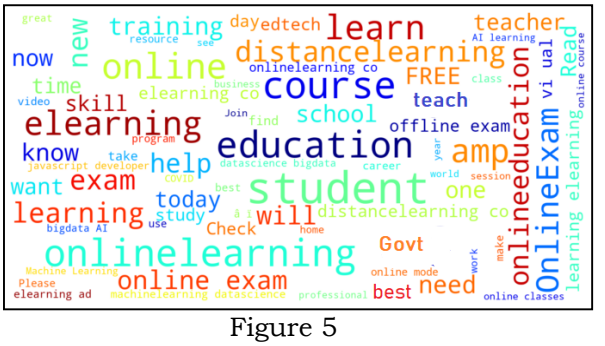


Figure 5

Experimental Setup

All experiments were conducting using Windows 10 operating machine having 2.1 GHz 12 CPUs AMD Ryzen 5 5500U processor with 8192 MB 3200Mhz DDR5 memory. With the help of 6 Machine Algorithms (NB, SVM, LR, XGB, RF, KNN), We get the following result where maximum accuracy achieved using XGBOOST to be 0.94

Classifier	Accuracy	F-Score	Precision	Recall
NB	0.92	0.96	0.92	1.00
SVM	0.93	0.96	0.94	0.99
LR	0.93	0.96	0.93	1.00
XGB	0.94	0.97	0.95	0.99
RF	0.92	0.96	0.92	1.00
KNN	0.92	0.96	0.93	0.99

Figure 6. Result

Conclusion

As the pandemic takes over, Distance learning comes into the play as students are not able to go school physically, classrooms are now taken to the home of students. In this work, we have examined how the distance learning plays a vital role in building the future of students in India and globally. It is observed that there are some areas of improvement to make distance learning more efficient and accessible to the students of the country. There are many plus points of distance learning in India, Now, Students are now getting quality education also in remote areas and they also do not have to travel to attend school far from their home. The biggest problem in distance learning is the screen time facing by the students which is affecting the mental and physical health of student. One way to reduce screen time is to have limited number of classes in a day and there should be break between each class which may help the students to understand easily and they will also refresh after taking their classes. Since the cases are now declining in India, schools are now open in offline mode but we can apply the positive outcomes of distance learning in offline mode like submitting the assignments online or having quizzes from the home which will help students as they don't have to bring load kgs of school bag daily. Finally, With the help of distance learning we can make the education for students more accessible which help them to make their future bright and make themselves assets in the development of the country.

Future scope

In this study, we analyze the effect of COVID-19 pandemic on the life of students and educators in India by summarizing the dataset collected from Twitter API and analyze it with the help of pretrained python package (i.e., TextBlob, Vader) and Machine learning techniques (i.e., Naïve bayes, Support Vector Machine, Logistic

Regression, XG-Boost, Random Forest). There are also some extra scopes of research is to analyze how distance learning playing role specifically in school going children and college going children and how the difference is there in between two. The study can also be further expanded and should compare the distance learning before the Covid-19 period and during the Covid-19 period. The results should also be done on the developing countries and the results should compare with the developed countries so the required steps should be taken to make distance learning better across the world.

References

1. AntonPak, Oyelola A.Adegboye,,"Economic consequences of the covid-19 outbreak: The need for epidemic preparedness"
2. Yahya Almurtadha , Sentiment Analysis to Measure Public Response to Online Education During Coronavirus Pandemic
3. Aljabri, M.; Chrouf, S.M.B.; Alzahrani, N.A.; Alghamdi, L.; Alfehaid, R.; Alqarawi, R.; Alhuthayfi, J.; Alduhailan, N. Sentiment Analysis of Arabic Tweets Regarding Distance Learning in Saudi Arabia during the COVID-19 Pandemic. Sensors 2021, 21, 5431. Doi: 10.3390/s21165431
4. Filiz Angay Kutluk, Mustafa Gulmez, "A research about distance education students' satisfaction with education quality at an accounting program" Published by Elsevier Ltd. Selection and/or peer review under responsibility of Prof. Dr. Hüseyin Uzunboylu
5. doi: 10.1016/j.sbspro.2012.05.556
6. Osmanoglu,U.O., Atak,O.N., Caglar,K., Kayhan, H. &Can, T.C. (2020). Sentiment Analysis for Distance Education, Doi: 10.31681/jetol.663733
7. Nimasha Arambepola, Analysing the Tweets about Distance Learning during COVID-19 Pandemic using Sentiment Analysis, ISSN 2756-9160 / November 2020.
8. Aiyanyo, I.D.; Samuel, H.; Lim, H. Effects of the COVID-19 Pandemic on Classrooms: A Case Study on Foreigners in South Korea Using Applied Machine Learning. Sustainability 2021, 13, 4986. Doi:10.3390/su13094986