Impact of COVID-19 Pandemic on Education using Sentiment Analysis

[1]Keerthi Sourav Namani, [2]Sahithya Namani, [3]Mounika Gurumurthy, [4]Sowjanya Jindam

[1,2] Student, [3]MT-Operations, [4]Assistant Professor

[1,4] Department of Information Technology at Maturi Venkata Subba Rao Engineering College, Osmania University, Hyderabad, Telangana, India

[2] Department of Computer Science and Engineering at Maturi Venkata Subba Rao Engineering College, Osmania University, Hyderabad, Telangana, India

[3]MT – Logistics, Vedanta Resources Ltd., Aluminium & Power – Jharsuguda, Odisha, India

Abstract : The COVID-19 has resulted in schools, colleges and universities shut all across the world. Globally, over 1.2 billion children are out of the classroom. In response to schools, colleges and universities closures, UNESCO recommended the use of distance learning programmes and open educational applications and platforms that schools and teachers can use to reach learners remotely and limit the disruption of education. By this e-learning there is a significant gap between those from privileged and disadvantaged backgrounds. Some students without reliable internet access and/or technology struggle to participate in digital learning. In our proposed system we collect the opinions of the e-learning from the students, teachers and parents to analyze the changes of educational systems during the COVID-19 pandemic. We use sentiment analysis to mine the polarity of the opinions of the students, teachers and parents. In this project, it is identified the impact of pandemic on education sector by categorizing the population set into students, teachers, and parents locating in urban and rural areas.

***IndexTerms* –** Sentiment Analysis, Impact, Education sector, Reviews, Sentiment, Logistic Regression, Natural Language Processing, Classification.

I. Introduction

With the developments of technology and automation in each and every sector, textual analysis and opinion mining is found to be challenging in developing insights from the big data. Among all the sectors, education is one of the major and important sector that contributes a lot to the growth of the economy. With this aspect, it is important to understand the impact of covid pandemic on the education sector by which technological advances can be made to avail education to all the required people. Thus a recommendation system which performs classification of the unstructured data, is developed that identifies the impact of pandemic on education based upon the reviews provided by teachers, students and parents belonging to urban and rural areas provided their further personal details.

Our aim is to build a recommendation system which provides impact of pandemic on education sector given the review of parent, teacher, and student belonging to either city, town or village using sentiment based recommendation approach.

Sentimentbasedrecommendationisbasedontextbasedrecommendationwhich is also known as Opinion mining.

Sentiment analysis is performed using Natural Language Processing which is a technique under Text mining or by either applying Machine Learning algorithms.

Proposed work focuses on extracting opinions of people on COVID-19 pandemic impact on education and then applying the Machine Learning algorithms to classify and to categorize the data and perform Natural Language Processing on sentiments of reviews.

This sentiment scores can be used for analyzing the people's opinions of education system in this pandemic situation. In this we are using Classification technique for evaluation and comparison of the results. In this mainly we are using Logistic Regression to classify the population and sample data set. Data Visualization tools are used to present the results.

II. Figures

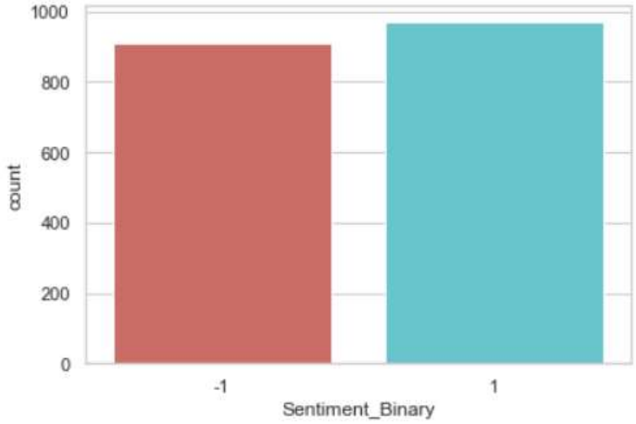


Fig-1. Frequency of Positive & Negative Impact of Covid on Education

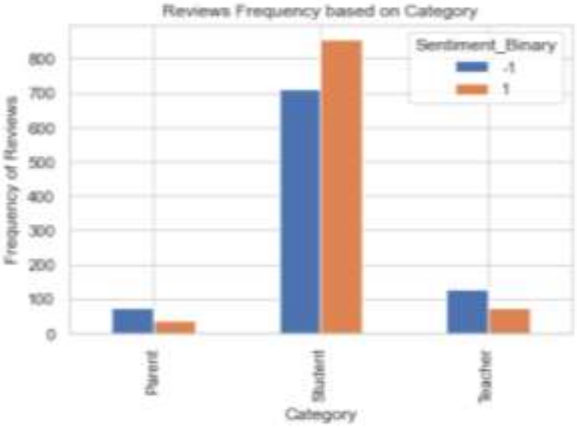
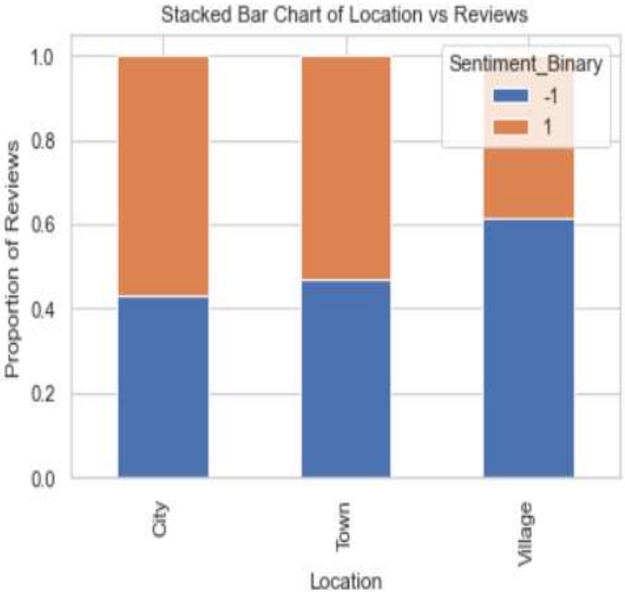


Fig-2. Frequency of Review based upon Student, Parent & Teacher



3. Frequency of Review based upon City, Town & Village

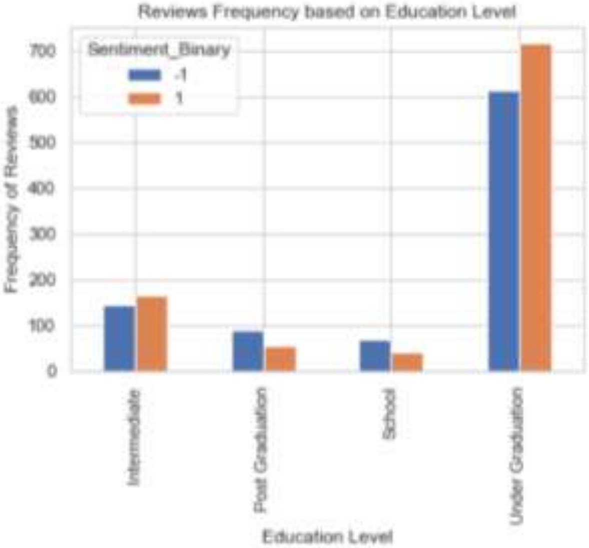


Fig-4. Frequency of reviews Based upon Education level

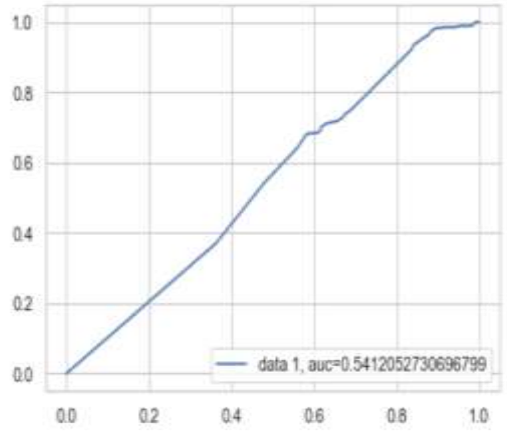
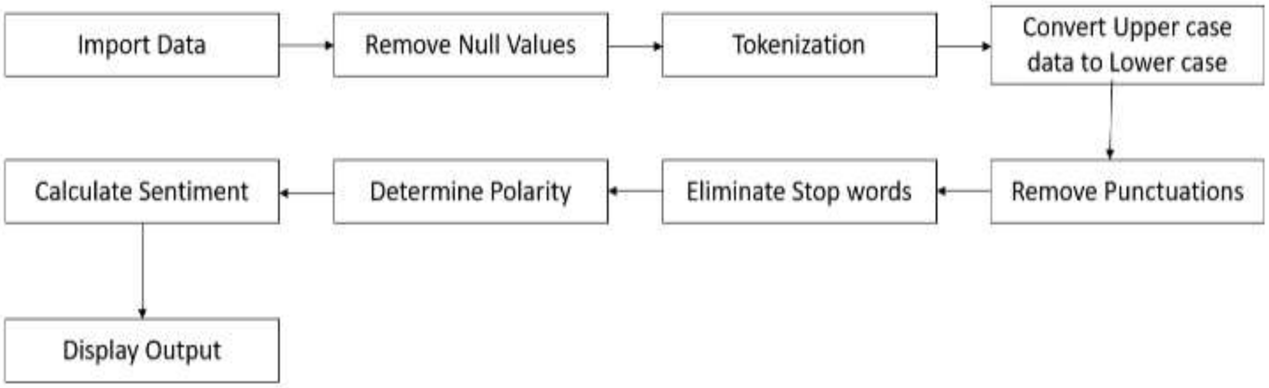


Fig -5. ROC curve of the model

III. SECTIONS

Totally, we have 4 sections as follows: Data Collection, Data Cleaning, Sentiment Analysis and Logistic Regression.

3.1 Block Diagram of the system:



3.2 Data Collection*:*

Data collection, or data acquisition, is the very first step. In our project, we used primary data collection method. Apart from Surveys, the most commonly used primary data collection methods are interviews and experiments. Primary data are usually collected from the source- where the data originally originates from and are regarded as the best kind of data in research.

The sources of primary data are usually chosen and tailored specifically to meet the demands or requirements of a particular research. Also, before choosing a data collection source, things like the aim of the research and target population need to be identified.

Few advantages of using primary data over secondary data are that Primary data is specific, accurate Up to data information, Control of the ownership to the researcher. But limitations of the primary data collection is that it is time consuming, expensive, and not always feasible subject to time, volume and cost constraints.

Procedure that we had chosen to collect data is through Google Forms. The responses are recorded and stored in Spreadsheet or Excel. The audience that our project considered is subjected to the categories of Students, teachers, and parents. The survey recorded approximately 3000 observations.

3.3 Data Cleaning*:*

Data cleaning is hard to perform, maintain and hard to know where to start. There seem to alwaysbe errors, mistakes, dupes, or format inconsistencies. One of the most challenging aspects of datacleaning has got to be maintaining a clean list of data, whether it's sourced from numerous resources or physically / manually entered by our hard work, or a combination of both.

In our problem, we remove the data points for which data is missing in various features. Forinstance, initially, we had 4,789 observations or data points after data collection stage and after removingthe data points for which the certain features are NULL (missing), we now have only 3,245 observations (data points). We brought down the number of data points from 4K to 3K approximately. After performing this step, we store the data in a separate excel file. This file is storedandloadeditlater with further filtering based on the required objective.

In data cleaning, we also remove duplicate observations for numerical data by replacing the missing value with the mean value of that particular variable. Few outliers are identified by calculating descriptive statistics for all the quantitative variables. Box plots are drawn. The data points that fall outside the box plot are considered to be outliers. We have two options, one to remove the outliers, or either replace the outliers values with the mean value of the variable. Since the population is only 3K, it is not feasible to remove the outliers which will further reduce the size of the data set. Hence, mean value is calculated and replaced it in the place of missing values of numerical data. Also the outlier points are detected and performed the same operation.

Not only mean, but the other methods to replace missing values is to calculate median or mode. But most commonly used technique is to replace it with mean, in order to obtain proper normal distribution of the data set.

Data cleaning procedure also involves removal of insignificant variables. Descriptive Statistics for Categorical variables is calculated and Frequency is plotted for all the variables. If any variable is found with single category, then the variable is considered to be insignificant. The entire variable is deleted from the data set, since the single category in the variable will be of no use to the research study.

3.4 Sentiment Analysis

Sentiment Analysis is also referred to as Opinion Mining. It is a Natural Language Processing technique used to determine whether the data or the observation is positive, negative, or neutral. It is often applied for textual data to help the users or businesses analyze, understand whether the data is positively impacted or negatively impacted, or else biased (neutral) based on the objective or the need of the target audience.

Since, the users express their thoughts and feelings more openly than ever before, Sentiment Analysis is becoming an essential tool to monitor and understand that sentiment. Automatically analyzing user reviews, such as opinions in survey responses and social media conversations, allows users (Educational Institutions, Government etc.) to learn what makes them feel good or bad about Online Education especially in the pandemic situation, so that they can tailor technological advancements and facilities to all the students, as well teachers to meet their needs.

Sentiment Analysis models focus on polarity (positive, negative, or neutral) based on the feelings and emotions (happy, sad etc.). Sometimes urgency and intentions can also be considered based on the user need or objective of the research.

This study is oriented towards considering the emotions of the audience. Their reviews depict whether there is positive, or negative, or neutral impact of covid pandemic on the online education. Therefore, the polarity is calculated for all the reviews and the sentiment is determined based on the range of polarity. If the polarity is greater than 1, the review is considered to be positive. If the polarity is less than 1, the review is considered to be negative. And in case if the polarity is observed to be 0, it is said that the review is neutral. For the entire dataset, polarity is calculated based on the criteria of categorizing the users. The average of the polarity determines the complete impact of pandemic on the online education. It also gives insights to government as well as educational institutions to take respective actions to ensure that all the users are in avail of proper quality education.

IV. METHODOLOGY

4.1 The Training and Prediction Processes

In the training process (a), our model learns to associate a particular input (i.e. a text) to the corresponding output (tag) based on the test samples used for training. The feature extractor transfers the text input into a feature vector. Pairs of feature vectors and tags (e.g. *positive*, *negative*, or *neutral*) are fed into the machine learning algorithm to generate a model.

In the prediction process (b), the feature extractor is used to transform unseen text inputs into feature vectors. These feature vectors are then fed into the model, which generates predicted tags (again, *positive*, *negative*, or *neutral*).

1. Break each text document down into its component parts (sentences, phrases, tokens and parts of speech)
2. Identify each sentiment-bearing phrase and component
3. Assign a sentiment score to each phrase and component (+1 or -1 or 0)
4. Optional: Combine scores for multi-layered sentiment analysis

4.2 Feature Extraction from Text

The package that is used to perform Sentiment Analysis for our project is “textblob”. The sentiment function of textblob returns two properties- Polarity and Subjectivity. Polarity is a float value that lies in the range [-1,1] where 1 means positive statement, and -1 means negative statement. Subjectivity refers to the personal opinion, emotion or judgement whereas Objectivity refers to factual information. Subjectivity is also a float value that lies in the range [0,1].

4.3 Logistic Regression

The coefficients (Beta values b) of the logistic regression algorithm must be estimated from your training data. This is done using maximum-likelihood estimation. Maximum-likelihood estimation is a common learning algorithm used by a variety of machine learning algorithms, although it does make assumptions about the distribution of your data. The intuition for maximum-likelihood for logistic regression is that a search procedure seeks values for the coefficients (Beta values) that minimize the error in the probabilities predicted by the model to those in the data (e.g. probability of 1 if the data is the primary class).

4.4 Preparing Data for Logistic Regression:

The assumptions made by logistic regression about the distribution and relationships in the data are much the same as the assumptions made in linear regression. Ultimately in predictive modeling machine learning project it is focused on making accurate predictions rather than interpreting the results.

- **Binary Output Variable**: Logistic regression is intended for binary (two-class) classification problems. It will predict the probability of an instance belonging to the default class, which can be snapped into a 0 or 1 classification.

- **Remove Noise:** Logistic regression assumes no error in the output variable (y), consider removing outliers and possibly misclassified instances from your training data.

- **Gaussian distribution:** Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output. Data transforms of the input variables that better expose this linear relationship can result in a more accurate model.

- **Remove Correlated Inputs:** Like linear regression, the model can overfit if it has multiple highly-correlated inputs. Consider calculating the pairwise correlations between all inputs and removing highly correlated inputs.

- **Fail to Converge:** It is possible for the expected likelihood estimation process that learns the coefficients to fail to converge. This can happen if there are many highly correlated inputs in the data or the data is very sparse (e.g. lots of zeros in the input data).

4.5 Natural Language Processing to perform Sentiment Analysis:

In our study, to perform Sentiment Analysis the procedure used was Natural Language Processing. NLP in Python is implemented using NLTK library. NLTK is a perfect library for education and research. Within the package, TextBlob is used to perform Sentiment Analysis. VADER is used to calculate the polarity, sentiment score for the reviews. Since TextBlob can’t be used for performing textual analysis such as Tokenization, Lemmatization, and Stemming, nltk library is used.

V. EQUATIONS

Logistic regression equation:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Eachcolumn in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

Final equation for Logistic Regression could be written as follows, for more than one independent variable:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

The above equation is referred to as Logit Function which is used in expressing the equation of logistic regression.

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

VI. Algorithms

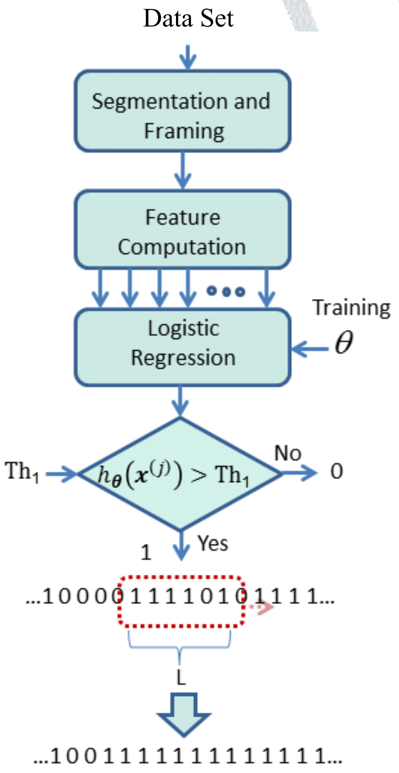
# Algorithm to Calculate Polarity and Subjectivity of each review

```
subjectivity = [ ]
polarity = [ ]

while index in "length of cleansed_dataframe" do
    review = TextBlob("text")
    subjectivity.append(review.sentiment.subjectivity)
    polarity.append(review.sentiment.polarity)
end while

cleansed_dataframe['Subjectivity'] = subjectivity
cleansed_dataframe['Polarity'] = polarity
```

# Flow Chart for Logistic Regression



# Algorithm to remove null values

```
Node deleteNode(Node root, int valueToDelete) {
  if root = null
    return node
  if root.value < valueToDelete
    deleteNode(root.right, valueToDelete)
  if root.value > valueToDelete
    deleteNode(root.left, valueToDelete)
         else
    if (isLeafNode(root))
      return null
  if (root.right == null)
    return root.left
  if (root.left == null)
    return root.right
  else
    minValue = findMinInRightSubtree(root)
    root.value = minValue
    removeDuplicateNode(root)
return root
```

VII. Suggestions

1) This study can be a recommendation tool to state and central governments to take corresponding actions based upon the impact in various scenarios and provide facilities across the country so that each and every child is made beneficial of the online education system

2) It provides relevant suggestions to the users (student, teacher, and parent) to effectively use best of technology so that overall impact increase with a positive slope

3) It can also offer relevant recommendations to Ed-Tech companies to enhance technology and automation based on the impact of the pandemic.

4) Based upon the impact, the growth or decline of the economy is observed and provides the government to facilitate and formulate respective strategies. It also increases awareness of the new-normal condition to the users who are facing negative impact.

VIII. Conclusion

In our study, to perform Sentiment Analysis the procedure used was Natural Language Processing. NLP in Python is implemented using NLTK library. NLTK is a perfect library for education and research. Within the package, TextBlob is used to perform Sentiment Analysis. VADER is used to calculate the polarity, sentiment score for the reviews. Since TextBlob can’t be used for performing textual analysis such as Tokenization, Lemmatization, and Stemming, nltk library is used.

IX. References

[1] Dr. Pravat Kumar Jena, Assistant Regional Director, IGNOU Regional Centre, Bhubaneswar, “Impact of Pandemic Covid-19 on Education in India”, International Journal of Current Research Vol.12 Issue 07, July 2020.

[2] Shamantha Rai, Sweekrithi and Prakyath Rai. “Sentiment analysis using machine learning classifiers: Evaluation of Performance “ IEEE 2019 4th International conference on computer and communication systems.

[3] Singh, Prabhsimran, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government." ICT Express (2017).

[4] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.

[5] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.