Plagiarism Checker X - Report

Originality Assessment

12%

Overall Similarity

Date: Jan 24, 2023
Matches: 996 / 8346 words
Sources: 46

Remarks: Low similarity detected, check with your supervisor if changes are required.

Verify Report:
Scan this QR Code

Sentiment Analysis on Bangla OTT Platform Content Using Machine Learning and Deep Learning Approach Kutub Uddin ID: CSE 01806766 & Joy Dey ID: CSE 01806698 Department of Computer Science and Engineering Port City International University 7-14, Nikunja Housing Society, South Khulshi, Chattogram, Bangladesh January 2023

Sentiment Analysis on Bangla OTT Platform Content 24 Using Machine Learning and Deep Learning Approach Submitted by Kutub Uddin ID: CSE 01806766 & Joy Dey ID: CSE 01806698 Under The Supervision of Mrs. Farzina Akther Senior Lecturer Department of Computer Science and Engineering Port City International University For a portion of the requirements for the Bachelor of Science in Computer Science and Engineering degree, this thesis is submitted to the 3 department of computer science and engineering at Port City International University. January 2023

iii DEDICATION This thesis is dedicated to the All-Powerful Creator and our cherished parents, who serve as examples and sources of inspiration.

iv APPROVAL FOR SUBMISSION This thesis titled “Sentiment Analysis on Bangla OTT Platform Content using Machine Learning and Deep Learning Approach” by Kutub Uddin,Student ID: CSE 01806766 and Joy Dey Student ID: CSE 01806698. It has been authorized for Batch: 18 Day to be submitted 3 to Port City International University's Department of Computer Science and Engineering in partial fulfillment of the criteria for the Degree of Bachelor of Science. --------------------------- Mrs. Farzina Akther Senior Lecturer Department of Computer Science and Engineering Port City International University 7-14, Nikunja Housing Society, South Khulshi, Chattogram, Bangladesh.

v DECLARATION We respectfully affirm that the work for our undergraduate degree "Sentiment Analysis on Bangla OTT Platform Content using Machine Learning and Deep Learning Approach" is entirely original. This thesis includes correctly cited sections throughout. ---------------------------------- ----------------------------------- Kutub Uddin Joy Dey ID: CSE 01806766 ID: CSE 01806698 Batch: CSE 18 Day Batch: CSE 18 Day [3] Department of Computer Science and Engineering Port City International University

vi ACKNOWLEDGEMENT We start by giving thanks to our Creator for endowing us with the skills and endurance necessary to complete this work. Then, from the bottom of our hearts, we thank our respected supervisors Mrs. Farzina Akther, who provided us with a variety of support and guidance to help us do this task. Last but not least, we would want to express our gratitude to some of the teachers, big brothers and friends who have helped and motivated us since the beginning. ---------------------------------- ---------------------------------- Kutub Uddin Joy Dey ID: CSE 01806766 ID: CSE 01806698 Batch: CSE 18 Day Batch: CSE 18 Day [3] Department of Computer Science and Engineering Port City International University

vii ABSTRACT This research revealed a method for conducting sentiment analysis on evaluations of OTT (Over The Top) Platform Content which have been focused on [24] machine learning and deep learning algorithm to classify Bengali text reviews. This method can automatically analyze how viewers responded to a particular film, web series, song, etc. With more people openly expressing their opinions on social networking sites such as Facebook, Twitter, Instagram and YouTube analyzing the sentiment of comments made about a specific OTT content can indicate how well the content is accepted by the general public. The social media websites' publicly accessible comments and posts served as the source of the dataset for this experiment, which was manually gathered and labeled. This system is split into three classes (Positive, Neutral, Negative).In this system, we used machine learning algorithms such as Random Forest(RF),K-Nearest Neighbors(KNN),Decision Tree(DT), Support Vector Machine(SVM),Logistic Regression(LR),and Multinomial Naive Bayes(MNB),as well as deep learning algorithms such as Long Short Term Memory(LSTM). Keywords: Sentiment 3 Analysis, Natural Language Processing, Supervised Model, Random Forest(RF), K-Nearest Neighbors(KNN), Decision Tree(DT), Support Vector Machine(SVM),Logistic Regression(LR),and Multinomial Naive Bayes(MNB),Long Short Term Memory(LSTM)

25 viii Table of Contents

xi List of Figures

1 CHAPTER 1 INTRODUCTION 1.1 Overview Bangla has the second-highest population of speakers in the Indian subcontinent and it ranks sixth among the most widely spoken languages globally.People are using 33 social networking sites like

Facebook,Twitter,Instagram and others to voice their thoughts on a variety of topics frequently in their own native language.Throughout the course of the past 10 years as the use of social media has grown.Since numerous simple-to-use Bangla keyboard apps were launched in the last few years,the use of Bangla on social media has increased as well.On social networking sites,people frequently talk about OTT platform movies,web series etc.Even there are dedicated groups,pages where people can discuss these topics.It is possible to determine whether or not people like such a particular movie or web series by analyzing the sentiment of their comments.Another practical application could be analyzing the audience's reaction to a web series trailer,which can indicate whether the movie is positively or negatively initially expectd by the general public. However, manually evaluating every single comment is a time-consuming and tedious task. As a result, this research investigates the effectiveness of some machine learning models in analyzing the sentiment of OTT movie, web-series related comments made in Bangla. On this dataset, various machine learning methods such as Random Forest (RF), K-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machines (SVM), Logistic Regression (LR), and Multinomial Naive Bayes (MNB) were used.As Deep Learning based approaches are being used in various sectors recently, Long Short Term Memory (LSTM) also applied for comparison. This research paves the way for further development of sentiment analysis methods in the Bangla language in other sectors by providing a method for automated sentiment analysis.   1.2 Problem Statement There are [38] more and more people using the internet every day, and there is a global competition to automate everything. Natural language processing is currently being used by several researchers to automate the translation of various languages has already invented ways for people to convey their emotions in many different languages. However, Bengali lags behind all other languages in the world, hence the purpose of this research is to find out how well some [29] machine learning and deep learning models work at detecting the sentiment of comments posted on OTT movies and web series in Bangla.

2 1.3 Motivation □ So far [32] a lot of work has been done on Bengali movies and text, but no work has been done on the content of Bengali OTT Platform. □ Sentiment analysis is currently occupying a leading position in the field of research. It is helpful for getting results without wasting time and brain. [37] Sentiment analysis is a process to automatically extricate sentiment or opinion from OTT contents review data. □ There are scopes for improving existing work through increasing accuracy and enhancing dataset. 1.4 Objective □ To build a new Bengali dataset from social media websites comments. □ To help the new viewers in OTT to acknowledge a true description of old viewers and also the sites can upgrade their service or contents through users review. □ The purpose of this thesis is to extract effectiveness of some machine learning and Deep learning models in analyzing the sentiment of OTT content related comments made in Bangla dataset and leveled it 3 classes(positive, negative, neutral. 1.5 Organization of the Thesis Document This thesis's clear message is organized as follows: Chapter 2 – Literature Review: This chapter summarized related research work and described research work comparison. Chapter 3 – Overview of Methodology and Machine Learning Algorithm: This chapter summarized the overview of Machine Learning Algorithm and methodology and its working method. Chapter 4 – Performance Evaluation: The results and discussion regarding the system's accuracy in various [29] machine learning and deep learning models were summarized in this chapter. The machine learning model and the deep learning model are compared in this section. Chapter 5 – Required Tools: All of those tools that we used in our work were outlined in this chapter. Chapter 6 – Conclusion and Future Work: This chapter summarized the result's conclusion. It also includes our limitations and future work for this outcome.

3 CHAPTER 2 LITERATURE REVIEW 2.1 [37] An overview of sentiment comparison The previous research on sentiment analysis and movie reviews on Bengali text are presented in this chapter. 2.2 Related Work □ In the research paper titled “Analyzing Sentiment of Movie Reviews in Bangla by Applying Machine Learning

Techniques”, by Rumman Rashid Chowdhury, Mohammad Shahadat Hossain, Sazzad Hossain and Karl Andersson in 2020. They collected from 4000 reviews data on social media websites and using model SVM, MNB, LSTM. By this model, they got best accuracy 88.90% using SVM.The limitation of this paper is the small amount of labelled data [1]. "Evaluation of Naive Bayes and Support Vector Machines on Bangla Textual Movie Reviews" was the title of a research paper written in 2018 by Nayan Banik and Md. Hasan Hafizur Rahman. They used model NB, SVM to compile data from 800 reviews they found on social media platforms and the Bangla Movie Database (BMDb). Their best precision was 86% using this model. The very small number of labelled data and the sparse use of models are the paper's main limitations.[2] In 2020, Atiqur Rahman and Md. Sharif Hossen will publish a study titled "Sentiment Analysis on Movie Review Data Using Machine Learning Approach." They utilized the models BNB, DE, SVM, ME, and MNB to collect data from 2000 reviews on social networking websites. They achieved the best accuracy with this model, 88.5%. The small amount of labeled data is the paper's main drawback.[3]

4 Saeed Mian Qaisar published a research paper titled "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory" in 2020. They used the LSTM model to collect data from the Internet Movie Database (IMDb) from 10000 datasets. They achieved the highest accuracy of 89.9% using this model. The limitation of this paper is the use of fewer models [4]. Abdul Hasib Uddin,DurjoyBapery,Abu Shamim Mohammad Arif published a paper titled " Depression Analysis from Social Media Data in Bangla Language using Long Short Term Memory (LSTM) Recurrent Neural Network Technique" in 2019. They collected 5000 data from Twitter comments and annotate it with models such as LSTM. According to this model, with lstm size 128, the model generated highest accuracy 86.3% where batch size 25 and epoch no 20 [5].

5 44 CHAPTER 3 METHODOLOGY This chapter contains details on workflow

maintenance. It will present a clear understanding of the research. 3.1 Work Overview Due to the lack of dataset available on the internet about the Bangla ott platform content reviews, we first create one from social media. There are no websites that provide ott content review summaries in Bengali. As a result, we chose online streaming platforms such as Hoichoi, Chorki, Bongo BD, Binge, and others. Following that, we manually labeled our dataset. Then we used Python to put my work into action. We import our dataset first. Then, we used the NLTK and BNLP packages to preprocess our dataset. On our dataset, we also used the TF-IDF, Word Embedding, and SMOTE methods. Our dataset is now prepared for the split technique. The data was then divided into an 80/20 split. Accordingly, 80% of the data will be useful for the training set and 20% will be useful for the testing set. Then, we used deep learning and machine learning algorithms. After the algorithms have been implemented, we now choose the best model for my research project. The best model was then applied to the prediction. Finally, we can assess how well this research work performed.

6 3.2 Methodology Data Collection (7003) Sentiment Label Preprocessing □ Tokenization. □ Stopwords and Emoji Removal. □ Removing extra Symbols □ Removing Punctuations □ Stemming Feature Extraction using TF - IDF & Bag of Word for Machine Learning Word Embedding for Deep Learning Split into 80% training & 20% testing data Classification Algorithm LR, MNB, SVM, RF, LSTM) (Model Evaluation Result

Figure 3 . 1 Methodology

7 3.3 Dataset Collection The dataset is the central component of this project. 36 One of the most difficult aspects for us was creating the dataset. In our project, we created our own dataset we have analyzed the data well and also in case of data validation we have taken the help of 40 friends and supervisors of the university and then leveled it.. First and foremost, we faced multiple issues 38 as a result of this dataset. Datasets on OTT content

reviews in Bengali are not available in online resources. There are no sites that provide OTT review summaries in Bengali. So we did go through some steps to create our dataset. □ Step 1: To collect OTT content reviews and plot summaries, we chose an online streaming platform. For example Hoichoi, Chorki, Bongo Bd, Binge, and so on. □ Step 2: We are having a lot of trouble with this step because we can't find any reviews on the OTT video streaming platform. Therefore, we gather information from 33 social media sites like Facebook, Instagram, Twitter, etc. □ Step 3: This step involved making a spreadsheet with two columns that had the labels Positive, Negative, and Neutral. Positive comment 1, Negative comment 2, and Neutral comment 0 are used to simplify the dataset. □ Now, our dataset is ready to perform. Figure 3.2 Sample Data

8 3.4 About Dataset We summarized our dataset in this section. There are 7003 total datasets and 3 classes and we have analyzed the data well and also in case of data validation we have taken the help of 40 friends and supervisors of the university and then leveled it. Table 1 About Dataset Content Total Total Data 7003 Data Label 3 (Positive, Negative, Neutral) Positive (1) 3662 Negative (2) 2261 Neutral (0) 1080 3.5 Preprocessing After data collection, this dataset needs to be preprocessed. Emoji, punctuation, digits, additional symbols, stop words, urls, user tags, and mentions were all taken out of the dataset in this section. Every piece of data was tokenized and stemmed. We utilized the NLTK and BNLP tools for this section. 3.5.1 Tokenization 6 Tokenization is a simple process that converts raw data into a useful data string. Tokenization is well known for its applications in cybersecurity and the creation of NFTs, but it is also an important part of the NLP process. Tokenization is a technique used in natural language processing to divide paragraphs and sentences into smaller units that can be assigned meaning more easily. So, We did tokenize the data for a single corpus/word. Before Tokenization:আজকে দেখলাম পুরাই টুইস্ট After Tokenization: <দেখলাম>< পুরাই>< টুইস্ট>

9 3.5.2 Removing Punctuations Text data contains [19] a large number of punctuation marks, yet they do not impart any meaning to a phrase that is repeated repeatedly. As a result, all punctuation is removed from the entire dataset. Example: !#$%^&*(),.?-_+\`·:={}|<> Figure 3.3 Removing Punctuation 3.5.3 Emoji's Removing In sentiment analysis, emoji and emoticons are both used to convey emotions in text data. There are several instance where it is used to describe difficult-to-put-into-words expression. Emoji's can be used to extract sentiment from OTT content reviews data, which is particulary valuable for sentiment analysis. However, because we are only concerned with text data in our research, emoji's are excluded from all of data. For this reason,We removed all emoji from data set. Figure 3.4 Emoji Remove 3.5.4 Removing Stop words The terms that appear most frequently in phrases and provide only extremely basic information are known as stop words, and they are unimportant in the context of text mining. There isn't much [34] to be learned from the most basic words in any language (such as articles, preposition, pronouns, conjunction, and other similar words).Examples of a few stop words are "অতএব, অথচ ,অথবা, অনুযায়ি, অনননে".To get better results ,stop words are deleted from sentiment analysis to concentrate on more relevant information. There is no most common list of stop words

10 that can be applied to any given language or dialect. Stopwords mainly meaningless words. So, we removed all Stopwords. Figure 3.5 Removing Stop Words 3.5.5 Stemming Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words "খানন, খয়খনয়লি, খাননল "converted into "খায়নলা" Figure 3.6 Visualization Dataset after cleaning 3.6 Feature Extraction Feature extraction is a method of dimensionality reduction in which a preliminary set of raw data is reduced to a smaller number of possible businesses for processing. A function of those large statistics sets is a massive variety of variables that necessitate a massive amount of computing resources for the system. Feature extraction refers to the need for strategies

that select 30 and/or combine variables into capabilities, thereby reducing the number of records that must be processed while still accurately and completely describing the original data set. For this technique, we used separately the TF-IDF and Countvectorizer method for machine learning and word embedding used for deep learning.

11 3.6.1 TF-IDF The term Frequency Inverse Document Frequency is abbreviated as TF-IDF. A potent feature selection technique that extracts important words from textual data is called Term Frequency - Inverse Document Frequency. It is very common to use a set of rules to convert text into a meaningful representation of numbers that can then be used to fit machine prediction algorithms. The significance of each word in a document is quantified by the TF-IDF statistic. Term Frequency :- The number of times a word appears in a textual content document. Inverse report Frequency :- Measure the word is a unprecedented word or general word in a document. The approach for TF - IDF is founded on Figure 3 . 7 Feature Extraction for TF - IDF .

12 3.6.2 Bag of Word (Count) Countvectorizer tokenizes (tokenization means breaking down a sentence or paragraph or any text into words) the text along with performing very basic preprocessing like removing the punctuation marks, converting all the 10 words to lowercase, etc. The vocabulary of known words is formed which is also used for encoding unseen text later. An encoded vector is returned with a length of the entire vocabulary and an integer. 3.6.3 Word Embedding More semantic representation then Countvectorizer and TF-IDF. It is a method of representing texts and documents. A word is represented in a lower dimensional space by a numeric vector input called a word embedding or word vector. It enables the representation of words with similar meanings to be similar. Additionally, they can suggest meaning. 50 different features can be represented by 45 a word vector with 50 values. Create Vocabulary our vocabulary size=51500 One hot representation: representation of the sentences using word indexes From vocabulary Sample: য়ঘী এই এটেট গকে ভাল অয়ভনি কেরকে। After preprocessing: য়ঘী এই এটেট গকে

ভাল অয়ভনি কেরকে। One hot vector representation: য়ঘী এই এটেট গকে ভাল অয়ভনি কেরকে = [8527,9012,2571, 2571,523,2560,3676] Vector representation after pre padding: [0 0 0…2571 523 2560 3673] 39 count for the number of times each word appe ared in the document. Figure 3 . 8 Feature Extraction for Countvectorizer(BOW)

13 3.7 Labeling This is the dataset after preprocessing. Figure 3.9 Labeled Dataset 3.8 Classification Model This section will give a brief description 43 of Machine Learning and Deep Learning algorithms. 3.9 Machine Learning Machine learning (ML) 19 is a subset of artificial intelligence that studies computer algorithms that can improve themselves automatically through experience and data. Figure 3.10 Machine Learning Image

14 Types of ML (Machine Learning) □ Supervised Learning : Training samples are labeled here, and it is extremely powerful when used correctly. □ Unsupervised Learning : In this case, the training samples are unlabeled, and the system naturally groups the input samples into 34 a limited number of classes. □ Reinforcement Learning : Also referred to as behavioral machine learning. It is similar to supervised learning, except that it is not trained on sample data. It has an algorithm that uses trial and error to improve itself and learn from new situations. 3.10 Naïve Bayes (NB) Naive Bayes classifiers are a set of classification algorithms that are used in supervised learning and are based on the Bayes' Theorem. It is especially useful in the text category, where a high-dimensional education dataset is included. The Nave Bayes Classifier is one of the most fundamental class algorithms for developing fast machine learning models capable of making quick predictions. It is a probabilistic classifier, which means it predicts based on an object's probability. Working Method :- Find the frequency by using class given data set. Find the probability by using class given data set. Find the class following unseen data or attribute given class. Following this method , P(C|A)=P(A|C)P(C) / P(A) Expressed by , P(C|A)=P(A|C)P(C) Where, A= Attribute and C=Class

15 Gaussian Naïve Bayes 15 In Gaussian Naive Bayes, continuous values related to every characteristic are assumed to be dispensed consistent with a Gaussian distribution. A Gaussian distribution is also known as normal distribution. whilst plotted, it gives a bell fashioned curve that's symmetric approximately the suggest of the characteristic values as proven under: Figure 3.11 Gaussian Naive Bayes Classifier Multinomial Naïve Bayes The multinomial Naive Bayes classifier is appropriate for discrete function categories (e.g., word counts for text classification). Normally, the multinomial distribution requires integer characteristic counts. However, in practice, fractional counts consisting of tf-idf may also work. Bernoulli Naïve Bayes 40 Bernoulli Naïve Bayes is a type of Naïve Bayes which used for discrete data and its work on Bernoulli distribution.The method for Bernoulli Naïve Bayes is based on, 3.11 Decision Tree (DT) Decision Tree is a supervised learning technique 41 that can be used both classification and regression problem but it is mostly used in classification problems. Decision tree is a tree structured classifier wherein internal nodes constitute the features of a dataset, branches constitute the decision rules and every leaf node constitute the outcome. There are two nodes which are Decision Node (Which used to make any decision and have multiple branches) and Leaf Node (Which are the output of those decision and don't

16 contain any further branches). The decision or test are performed based on feature of 5 given dataset. It is a graphical constitute for getting all the possible solutions to a problem or decision based on given conditions. Figure 3.12 Decision Tree Classifier Above diagram explain the general structure of a decision tree classifier. It's called decision tree because similar to a tree .It starts with a root node which expands on further branches and constructs a tree-like structure. 3.12 Random Forest (RF) A well-known machine learning model from the supervised learning approach is random forest. It is applicable to classification and regression issues in ML. It is wholly based on the idea of ensemble learning, which is a technique for combining several classifiers to address a

challenging issue and enhance system performance. According to the call, "Random Forest area is a classifier that incorporates several decision trees on subsets of the given dataset and takes the common to enhance the predictive accuracy of that dataset." The random forest uses the predictions from each tree to predict the very last output based on the precedence votes of predictions rather than relying solely on one decision tree. The more wide variety [8] of trees in the forest leads to higher accuracy and prevents the hassle of over fitting.The beneath diagram explains the working of the Random forest set of rules:

17 Figure 3.13 Random Forest Classification 3.13 Logistic Regression (LR) Logistic regression is the process of estimating the probability of a discrete outcome from an input variable. The majority of logistic regression models have a binary outcome that can be true or false, yes or no, or another value. Modeling [42] situations with more than two discrete outcomes can be done using multinomial logistic regression. A helpful analysis technique for classification issues is logistic regression, which [36] can be used to determine whether a new sample belongs in a particular category. Logistic regression is a helpful analytical method because cyber security issues like attack detection are classification issues. 3.14 K-Nearest Neighbors (KNN) The K-nearest Neighbors (KNN) algorithm is a supervised machine learning method that is mostly applied to classification and predictive issues. The KNN algorithm uses feature similarity to predict the values of new data points, and assigns a value to a new data point based on how much it resembles the points in the training set. KNN uses some mathematics to represent the idea of proximity, such as calculating the distance between points on a graph. There are many ways to calculate distance, and one method may be preferable depending on the situation. The straight-line distance, also known as the Euclidean distance, is a popular and well-known way to calculate distance.

18 3.15 Support Vector Machine (SVM) Support vector machines are a class of supervised learning methods for classification, regression, and detecting outliers. All of

these are common machine learning tasks. It can detect cancerous cells using [1] millions of images or predict future driving routes using a well-fitted regression model. SVMs are used for specific machine learning problems, such as support vector regression (SVR), which is an extension of support vector classification (SVC). The important thing to remember here is that these are simply math equations tuned to provide you with the most accurate answer possible as quickly as possible. SVMs differ from other classification algorithms in that they select the decision boundary that maximizes the distance from the nearest data points for all classes. The maximum margin classifier or maximum margin hyper plane is the decision boundary generated by SVMs. Figure 3.14 Support Vector Machine

19 3.16 Long Short Term Memory (LSTM) A recurrent neural network structure includes Long Short-Term Memory (LSTM) units or blocks. [19] Recurrent neural networks are designed to use specific types of artificial memory processes that can assist these artificial intelligence programs in better imitating human thought. [4] Long short-term memory blocks are used by the recurrent neural network to provide context for how the program receives inputs and generates outputs. [17] The long short-term memory block is a complex unit that includes weighted inputs, activation functions, inputs from previous blocks, and eventual outputs. Because the program uses a structure based on short-term memory processes to create longerterm memory, the unit is known as a long short-term memory block. These systems are frequently used in natural language processing, for example. Long shortterm [4] memory blocks are used by the recurrent neural network to evaluate a specific word or phoneme in the context of others in a string, where memory can be useful in sorting and categorizing these types of inputs. In general, LSTM is a wellknown and widely used concept in pioneering recurrent neural networks. Figure 3.15 [4] Long Short-Term Memory (LSTM)

20 CHAPTER 4 Performance Evaluation In this section, we discuss

the accuracy, classification report, confusion matrix, sensitivity, and specificity of this system using the various types of models that we utilized in our research. Additionally, show how the OTT platform content reviews are organized based on opinion and plot summary level. It can infer from a plot summary whether OTT platform content reviews are positive, negative, or neutral. It also offers a visual representation of the system's AUC, ROC, and confusion matrix. 4.1 Performance Evaluation In order to accomplish its goals, performance evaluation expresses values in a quantifiable manner. Several performance measures 19 have been used to evaluate the efficacy of our proposed model. Confusion matrix, precision score, recall score, f1 score, accuracy score, sensitivity score, recall score, area under the curve, and ROC analysis have all been carried out. 4.1.1 Confusion Matrix The Confusion Matrix is a fantastic resource for analyzing the behavior and comprehending the efficacy of a binary or categorical classifier. The confusion matrix is a two-dimensional array that contrasts the true label with the predicted category labels. Binary classification is indicated by the terms True Positive, True Negative, False Positive, and False Negative. 4.1.2 Precision Score Precision measures what proportion of predicted positive label is actually positive. 4.1.3 Recall Score Recall measures what proportion of actual positive label is correctly predicted as positive.

21 4.1.4 F1 Score F1-score is yet another effective performance matrix that makes use of the recall and precision matrices. The "Harmonic Mean" 9 of precision and recall can be used to determine an F1-score. Contrary to recall, which primarily makes a speciality of falsenegative, and precision, which generally specializes in false-positive, the F1-score focuses on false positive and false negative. 4.1.5 AUC-ROC Curve AUC is increased as region 14 Under Curve and ROC is increased as “Receiver Operating Characteristics”. It is also called AUROC and is increased as Area Under Receiver Operating Characteristics. AUC-ROC is one of the most important performance matrix used to test model performance. AUC-ROC is used for binary and multi-class classification but generally used for binary classification problems. 4.2 Performance of Bangla OTT

Platform Content Reviews Classification This section of the article concentrated on summaries of reviews of Bangla OTT Platform Content that we had collected from social ,media. Figure 4.1 Dataset distribution on Bangla OTT Platform Content Reviews . Now, we present and discuss the accuracy, classification report, confusion matrix, sensitivity, specificity, and ROC Curve of this system in various types of models that we used for our work.

22 4.3 Logistic Regression We got almost 87.89% accuracy by using Logistic Regression but using Countvectorizer feature we got 73.22% accuracy. Accuracy Score: Figure 4.2 Accuracy of LR using TF-IDF. Figure 4.3 Accuracy of LR using BOW.. Classification Report: Figure 4.4 Classification report of LR using TF-IDF.

23 Figure 4.5 Classification report of LRusing BOW. Confusion Matrix: Figure 4.6 Confusion Matrix of LR using TF-IDF. Figure 4.7 Confusion Matrix of LR using BOW.

24 4.4 Multinomial Naïve Bayes Using Multinomial Naive Bayes, we achieved 8 an accuracy of nearly 88.81% but using Countvectorizer feature we got 62.32% accuracy. Accuracy Score : Figure 4.8 Accuracy of MNB using TF-IDF. Figure 4.9 Accuracy of MNB using BOW. Classification Report: Figure 4.10 Classification report of MNB using TF-IDF.

25 Figure 4.11 Classification report of MNB using BOW. Confusion Matrix: Figure 4.12 Confusion matrix of MNB using TFIDF. Figure 4.13 Confusion matrix of MNB using BOW.

26 4.5 K-Nearest Neighbors Using K-Nearest Neighbors, we obtained 8 an accuracy of

nearly 60.95% but using Countvectorizer feature we got 54.06% accuracy. □ Accuracy Score: Figure 4.14 Accuracy of KNN using TF-IDF. Figure 4.15 Accuracy of KNN using BOW. □ Classification Report: Figure 4.16 Classification report of KNN using TF-IDF.

27 Figure 4.17 Classification report of KNN using BOW. □ Confusion Matrix: Figure 4.18 Confusion Matrix of KNN using TF-IDF. Figure 4.19 Confusion Matrix of KNN using BOW.

28 4.6 Linear Support Vector Machine By using Linear SVM, we achieved 8 an accuracy of nearly 88.61% but using Countvectorizer feature we got 68.82% accuracy. □ Accuracy: Figure 4.20 Accuracy of Linear SVM using TF-IDF. Figure 4.21 Accuracy of Linear SVM using BOW. □ Classification Report: Figure 4.22 Classification report of Linear SVM using TFIDF.

29 Figure 4.23 Classification report of Linear SVM using BOW. □ Confusion Matrix: Figure 4.24 Confusion matrix of Linear SVM using TF-IDF Figure 4.25 Confusion matrix of Linear SVM using BOW.

30 4.7 Decision Tree We got almost 75.90% accuracy by using Decision Tree model but using Countvectorizer feature we got 65.49% accuracy. □ Accuracy Score: Figure 4.26 Accuracy of DT using TF-IDF. Figure 4.27 Accuracy of DT using BOW. □ Classification Report: Figure 4.28 Classification report of DT using TF-IDF.

31 Figure 4.29 Classification report of DT using BOW. □ Confusion Matrix: Figure 4.30 Confusion matrix of DT using TF-IDF. Figure 4.31 Confusion matrix of DT using BOW.

32 4.8 Random Forest [5] Using the Random Forest model, we achieved an accuracy of almost 84.26% but using Countvectorizer feature we got 70.43% accuracy. Accuracy Score: Figure 4.32 Accuracy of RF using TF-IDF. Figure 4.33 Accuracy of RF using BOW. Classification Report: Figure 4.34 Classification report of RF using TF-IDF.

33 Figure 4.35 Classification report of RF using BOW. Confusion Matrix: Figure 4.36 Confusion Matrix of RF using TF-IDF. Figure 4.37 Confusion Matrix of RF using BOW.

34 4.9 [28] Long Short Term Memory Using the Long Short Term Memory, we achieved an accuracy of almost 81.00%. Accuracy Score: Figure 4.38 Accuracy of LSTM. Classification Report: Figure 4.39 Classification report of LSTM. Confusion Matrix: Figure 4.40 Confusion Matrix of LSTM.

35 4.10 Overall Model Performance Shows [9] the performance of the six classifiers in terms of accuracy, precision, recall, and f1 score measures also one deep learning model. The algorithms performed were machine learning classifiers (LR, MNB, KNN, SVM, DT, RF) with TF-IDF based feature extraction techniques and Deep Learning classifiers (LSTM) with Word Embedding based feature extraction techniques. The best result was given by Multinomial Nave Bayes classifier. The Multinomial Nave Bayes classifier achieved 88.81% accuracy, Linear Support Vector Machine classifier we achieved 88.61% accuracy, Random Forest classifier achieved 84.26% accuracy also Logistic Regression 87.89% accuracy gain. On the other hand, Long Short Term Memory achieved 81% accuracy However, KNN and DT had the lowest accuracy (60.95%) and (75.90%). Multinomial Nave Bayes outperformed the other models. In terms of f1score, [9] precision, and recall score, the Multinomial Nave Bayes classifier achieves the highest score of 89%

among all models. In the table below, we also show the overall performance of all models, including Accuracy, Precision Score, Recall Score, F1 Score, and AUC score. Table 2 Overall Comparison of all model performances. Features Algorithm Accuracy(%) Precision(%) Recall(%) F1score(%) TF-IDF LR 87.89 87.89 87.89 87.89 MNB 88.81 88.81 88.81 88.81 KNN 60.95 60.95 60.95 60.95 Linear SVM 88.61 88.61 88.61 88.61 DT 75.90 75.90 75.90 75.90 RF 84.26 84.26 84.26 84.26 Count LR 73.22 73.22 73.22 73.22 MNB 62.32 62.32 62.32 62.32 KNN 54.06 54.06 54.06 54.06 Linear SVM 68.82 68.82 68.82 68.82 DT 65.49 65.49 65.49 65.49 RF 70.43 70.43 70.43 70.43 Word Embedding LSTM 81.00 82.00 79.00 80.00

36 4.11 AUC-ROC Curve of All Models Figure 4.41 AUC-ROC Curve of overall model performances based on TF-IDF feature. Figure 4.42 AUC-ROC Curve of overall model performances based on Countvectorizer

37 The experiment is run again for graph analysis for all classifiers. Figure 5.43 to 5.45 show the AUC-ROC curves of the seven selected classifiers models with TF-IDF and Countvectorizer features respectively. Within the case of TF-IDF feature, the AUCROC curve of all models performances true positive rates are adequate, except KNearest Neighbors classifier, 9 the true positive rate of remaining models is much better. The best result was given by Multinomial Nave Bayes classifier. The Multinomial Nave Bayes classifier achieved 88.81% accuracy, Linear Support Vector Machine classifier we achieved 88.61% accuracy, Random Forest classifier achieved 84.26% accuracy also Logistic Regression 87.89% accuracy gain. On the opposite hand, due to the use of Countvectorizer feature, the performances of the classifiers models is seen to be slightly reduced, especially in the Decision Tree and K-Nearest Neighbors models. Overall, reviewing all classifiers models and using features Multinomial Nave Bayes provide best

result. Thus, the Multinomial Nave Bayes(MNB) classifier was selected of the final model. 4.12 Experimental Input and Output To get the necessary results, seven distinct classification algorithms are used. Multinomial Nave Bayes classifier achieves around 89 percent accuracy on the Bangla OTT Platform Content Review dataset. For regression model, the Logistic Regression (LR) has an accuracy of approximately 89 percent also deep learning model 28 Long Short Term Memory achieves around 81 percent accuracy.

38 4.13 Bangla OTT Platform Content Review Results The figure shows some of 8 the predictions made by the ML and DL Classifier for Bangla OTT Platform Content Review dataset Figure 4 . 43 Experimental Output.

39 4.14 Decision We look at the issue of classifying Bengali plot summaries by overall plot summary rather than topic, such as identifying whether a plot summary is positive, negative, or neutral. We discover that common machine learning techniques unquestionably outperform baselines created by humans using data from OTT platform content reviews plot. The six machine learning techniques and one deep learning models we used, such as Logistic Regression, Multinomial Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, and LSTM, perform just as well on the classification of content reviews on the Bangla OTT platform as they do on conventional topic-based categorization.

40 CHAPTER 5 REQUIRED TOOLS 5.1 Python Python : Python is an interpreted high-level programming language. 35 Its design philosophy emphasizes code clarity with its use of good sized indentation For machine learning python language is the best. For that reason we use python for implement our model. Pandas : Pandas is a library function of Python Language. We used pandas for data manipulation and analysis. Numpy : Numpy is a library function of Python

Language. We used numpy for provides a high performance multidimensional array and basic tools to compute with and manipulated these array. Sklearn : [7] Scikit-learn (Sklearn) is the maximum useful and strong library for machine learning in Python. It presents a spread of efficient tools for machine studying and statistical modeling together with type, regression, clustering and dimensionality discount via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib. Matplotlib : Matplotlib is a python library used to create 2d graphs and plots by way of the use of python scripts. [13] It has a module named pyplot which makes matters smooth for plotting by way of providing feature to manipulate line styles, font homes, formatting axes and so on. It supports a totally wide form of graphs and plots namely - histogram, bar charts, energy spectra, errors charts and so forth. It's miles used together with NumPy to offer an surroundings that is an effective open supply opportunity for MatLab. It may also be used with portraits toolkits like PyQt and wxPython. Keras : Keras is a python-based deep learning API that runs on top of TensorFlow, a machine learning platform. It was created to allow for quick experimentation. It's crucial [41] to be able to go from idea to result as quickly as possible when conducting research. Seabron : Seaborn is a library by and large used for statistical plotting in Python. It's far built on pinnacle of Matplotlib and presents stunning default styles and color palettes to make statistical plots extra attractive.

41 5.2 Google Colab For this research purpose we use Google Colab. [22] Colab is a free Jupyter notebook surroundings that runs completely in the cloud. Most significantly, it does now not require a setup and the notebooks. Google Colab is an 0nline platform which works on python programming language. All python packages and resource are available in Google Colab. It is an user friendly platform. 5.3 NLTK NLTK is a trendy python library with prebuilt functions and utilities [24] for the benefit of use and implementation. It's miles one of the most used libraries for natural language processing and computational linguistics. NLTK is a leading platform for constructing Python programs to paintings with

human language records. It presents smooth-to-use [18] interfaces to over 50 corpora and lexical assets together with WordNet, along side a set of text processing libraries for type, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-energy NLP libraries, and an energetic dialogue discussion board. 5.4 BNLPTK BNLP is an open source language processing toolkit for Bengali language consisting with tokenization, word embedding, POS tagging, NER tagging facilities. [2] BNLP provides pre-trained model with high accuracy to do model based tokenization, embedding, POS tagging, NER tagging task for Bengali language. BNLP pre-trained model achieves significant results in Bengali text tokenization, word embedding, POS tagging and NER tagging task. BNLP is using widely in the Bengali research communities with 16K downloads, 119 stars and 31 forks.

42 CHAPTER 6 CONCLUSION and FUTURE WORK 6.1 Conclusion In this experiment, various techniques were used to identify the polarity of the Bangla OTT platform content reviews using ML classifiers (MNB, SVM, RF, LR) with TFIDF and Countvectorizer based feature extraction techniques and Deep Learning classifiers (LSTM) with Word Embedding based feature extraction technique applied. Here I have created a dataset of 7003 Bangla sentences by myself and leveled the data in 3 classes (positive, negative, neutral). Through this research work, 89% accuracy is achieved using the MNB classifier and 81% for LSTM model. The dataset will need to be updated in the future by including more data samples. It'll be will fascinating to see how the models performs with a huge a dataset. To boost accuracy, applying more deep learning approaches and complex feature extraction methods such as word2vec or the BERT model can be used or create hybrid methods so that accuracy of the results can be increased. [11] Finding the polarity of the reviews can help in various domain. Intelligent systems can be developed which can provide the users with comprehensive reviews of Bangla ott platform contents, services etc. without requiring the user to go through individual reviews, he can directly take decisions based on the results provided by the

intelligent systems. 6.2 Future Work In the future, additional robust algorithms and features will be added to find the semantic relationships between the words in a comment, helping us to reliably detect sentiment. Other techniques for material comparison must be used. Increasing the dataset size should also be possible to achieve the goal. In future, we will work in large dataset on Bangla Text and Romanized Bangla Text. The value of accuracy can be increased by using a variety of review data. BERT is now helpful for text classification tasks as a result. But it's abundant resources. It will be viable choice for sentiment classification in the future.

43 REFERENCES [1]Chowdhury, R. R., Hossain, M. S., Hossain, S., & Andersson, K. (2019, September). 23 Analyzing sentiment of movie reviews in bangla by applying machine learning techniques. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-6). IEEE. [2]Banik, N., & Rahman, M. H. H. (2018, September). Evaluation of naïve bayes and support vector machines on bangla textual movie reviews. In 2018 international conference on Bangla speech and language processing (ICBSLP) (pp. 1-6). IEEE. [3]Rahman, A., & Hossen, M. S. (2019, September). Sentiment analysis on movie review data using machine learning approach. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 14). IEEE. [4]Qaisar, S. M. (2020, October). 12 Sentiment analysis of IMDb movie reviews using long short-term memory. In 2020 2nd International Conference on Computer and Information Sciences (ICCIS) (pp. 1-4). IEEE. [5]Tripto, N. I., & Ali, M. E. (2018, September). Detecting multilabel sentiment and emotions from bangla youtube comments. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-6) IEEE. [6]Zulfiker, M. S., Kabir, N., Biswas, A. A., Zulfiker, S., & Uddin, M. S. (2022). 31 Analyzing the public sentiment on COVID-19 vaccination in social media: Bangladesh context. Array, 15, 100204. [7]Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. 12 International Journal of Computer Applications, 179(7), 45-49. [8]Daeli, N. O. F., & Adiwijaya, A.

(2020). Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. Journal of Data Science and Its Applications, 3(1), 1-7. [9]Hossain, N., Ahamad, M. M., Aktar, S., & Moni, M. A. (2021, February). Movie genre classification with deep neural network using poster images. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp.195-199). IEEE.

44 [10]Haydar, M. S., Al Helal, M., & Hossain, S. A. (2018, February). Sentiment extraction from bangla text: A character level supervised recurrent neural network approach. 26 In 2018 international conference on computer, communication, chemical, material and electronic engineering (IC4ME2) (pp. 1-4). IEEE. [11]Sarkar, K., & Bhowmick, M. (2017, December). Sentiment polarity detection in bengali tweets using multinomial Naïve Bayes and support vector machines. In 2017 IEEE Calcutta Conference (CALCON) (pp. 3136). IEEE. [12]Al-Amin, M., Islam, M. S., & Uzzal, S. D. (2017, February). A comprehensive study on sentiment of bengali text. In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 267-272). IEEE. [13]Taher, S. A., Akhter, K. A., & Hasan, K. A. (2018, September). Ngram based sentiment mining for bangla text 21 using support vector machine. In 2018 international conference on Bangla speech and language processing (ICBSLP) (pp. 1-5). IEEE. [14]Azmin, S., & Dhar, K. (2019, December). 16 Emotion detection from bangla text corpus using naive bayes classifier. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-5). IEEE. [15] Uddin, A. H., Bapery, D., & Arif, A. S. M. (2019, July). 20 Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique. 27 In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2) (pp. 1-4). IEEE.

45 Appendix A List of Acronyms ML Machine

Learning DL Deep

Learning NLP Natural Language Processing TF-

IDF Term Frequency-Inverse

Document Frequency BOW

Bag of Word WE Word

Embedding SVM Support Vector

Machine MNB Multinomial Naïve

Bayes KNN K-Nearest Neighbors 46

RF Random Forest LR Logistic

Regression DT Decision

Tree CNN Convolutional Neural

Network ANN Artificial Neural

Network RNN Recurrent Neural

Network LSTM Long Short -Term Memory

Sources

1	https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/ INTERNET 1%
2	https://arxiv.org/abs/2102.00405 INTERNET 1%
3	https://portcity.edu.bd/HomePage/SubPageDetailsInfo/208/Teacher/farzina-akther INTERNET 1%
4	https://deepchecks.com/glossary/long-short-term-memory-lstm/ INTERNET 1%
5	https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm INTERNET 1%
6	https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/ INTERNET 1%
7	https://www.tutorialspoint.com/scikit_learn/index.htm#:~:text=Scikit-learn (Sklearn) is the most useful and robust,dimensionality reduction via a consistence interface in Python. INTERNET <1%
8	https://medium.com/analytics-vidhya/random-forest-4a2981aab4f7 INTERNET <1%
9	https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/ INTERNET <1%
10	https://www.studytonight.com/post/scikitlearn-countvectorizer-in-nlp INTERNET <1%
11	https://www.researchgate.net/publication/321843804_Sentiment_Analysis_of_Movie_Reviews_using_Machine_Learning_Techniques INTERNET <1%
12	https://computers.journalspub.info/index.php?journal=JCSPL&page=article&op=view&path[]=787 INTERNET <1%
13	https://www.tutorialspoint.com/python_data_science/python_matplotlib.htm INTERNET <1%
14	https://medium.com/analytics-vidhya/performance-metrics-precision-recall-f1-score-efb51ac111bd INTERNET <1%

15 https://www.geeksforgeeks.org/naive-bayes-classifiers/
INTERNET
<1%

16 https://link.springer.com/article/10.1007/s42979-022-01077-1
INTERNET
<1%

17 https://www.techopedia.com/definition/33215/long-short-term-memory-lstm#:~:text=The long short-term memory block is a complex,on short-term memory processes to create longer-term memory.
INTERNET
<1%

18 https://digitalhumanities.fas.harvard.edu/resources/choosing-digital-methods-and-tools/
INTERNET
<1%

19 https://learn.microsoft.com/en-us/azure/machine-learning/concept-deep-learning-vs-machine-learning
INTERNET
<1%

20 https://www.researchgate.net/publication/336854556_Depression_Analysis_from_Social_Media_Data_in_Bangla_Language_using_Long_Short_Term_Memory_LSTM_Recurrent_Neural_Network_Technique
INTERNET
<1%

21 https://www.sust.edu/d/cse/faculty-profile-detail/55
INTERNET
<1%

22 https://cppsecrets.com/users/1102811497104117108109111104116975048484864103109971051084699111109/How-to-run-C-and-C00-on-Google-Colab.php
INTERNET
<1%

23 https://ltu.diva-portal.org/smash/get/diva2:1349255/FULLTEXT01
INTERNET
<1%

24 https://ieeexplore.ieee.org/document/9083703/
INTERNET
<1%

25 http://eprints.utar.edu.my/2037/1/PE-2016-1105263-1.pdf
INTERNET
<1%

26 https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8433310
INTERNET
<1%

27 https://ieeexplore.ieee.org/xpl/conhome/9030693/proceeding
INTERNET
<1%

28 https://www.aiplusinfo.com/blog/introduction-to-long-short-term-memory-lstm/
INTERNET
<1%

29 https://www.ibm.com/topics/natural-language-processing
INTERNET
<1%

30 https://www.includehelp.com/data-science/feature-extraction.aspx#:~:text=Feature Extraction is the name for methods that,Feature extraction is a method of attribute reduction.
INTERNET
<1%

31 https://www.researchgate.net/publication/361258634_Analyzing_the_public_sentiment_on_COVID-19_vaccination_in_social_media_Bangladesh_context
INTERNET
<1%

32 https://ludwig.guru/s/a lot of work has been done
INTERNET
<1%

33 https://www.ielts-practice.org/many-people-believe-that-social-networking-sites-have-a-huge-negative-impact-on-both-individuals-and-society-band-8-ielts-essay-sample/
INTERNET
<1%

34 https://staff.fnwi.uva.nl/r.vandenboomgaard/MachineLearning/LectureNotes/Overview/index.html
INTERNET
<1%

35 https://www.knowprogram.com/python-mcq/
INTERNET
<1%

36 https://positivepsychology.com/how-to-express-gratitude/
INTERNET
<1%

37 https://research.aimultiple.com/sentiment-analysis-methods/
INTERNET
<1%

38 https://www.testbig.com/ielts-writing-task-ii-ielts-general-training-essays/more-and-more-people-are-using-internet-do
INTERNET
<1%

39 https://pianalytix.com/countvectorizer-in-nlp/
INTERNET
<1%

40 https://www.tutorialspoint.com/scikit_learn/scikit_learn_bernoulli_naive_bayes.htm
INTERNET
<1%

41 https://www.studocu.com/row/document/meru-university-of-science-and-technology/computer-science/methodology-chapter-3-lecture-notes-3/11380880
INTERNET
<1%

42 https://www.golinuxcloud.com/supervised-learning-algorithms/#:~:text=Multinomial logistic regression can be used to model,visitor choosing an offer made on your website.
INTERNET
<1%

43 https://www.javatpoint.com/machine-learning-vs-deep-learning
INTERNET
<1%

44	https://www.studocu.com/ph/document/technological-university-of-the-philippines/bs-in-information-system/chapter-3-methodology/36590081 INTERNET <1%	
45	https://www.geeksforgeeks.org/word-embeddings-in-nlp/ INTERNET <1%	
46	https://www.researchgate.net/figure/Hyperparameters-DT-decision-tree-kNN-k-nearest-neighbor-NB-naive-Bayes-RF-random_tbl2_355273097 INTERNET <1%	

EXCLUDE CUSTOM MATCHES OFF

EXCLUDE QUOTES ON

EXCLUDE BIBLIOGRAPHY ON