SENTIMENT ANALYSIS ON BANGLA CONVERSATION DATA USING MACHINE LEARNING APPROACH

BY

MD. MAHMUDUL HASSAN

ID: 171-15-8991

And

MD. SHAHRIAR SHAKIL

ID: 171-15-8558

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Nazmun Nessa Moon
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Ms. Nazmun Nessa Moon
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

31st MAY 2021

APPROVAL

This Project/internship titled **“Sentiment Analysis on Bangla Conversation Data Using Machine Learning Approach”**, submitted by **“Md. Mahmudul Hassan”** and **“Md. Shahriar Shakil”**, ID No: **171-15-8991** and **171-15-8558** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 31-05-2021.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Zahir

Gazi Zahirul Islam
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Raja Tariqul Hasan Tusher
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Farid

Dr. Dewan Md. Farid
Associate Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ms. Nazmun Nessa Moon and co-supervision of Ms. Nazmun Nessa Moon, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

____

Ms. Nazmun Nessa Moon
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

____

Ms. Nazmun Nessa Moon
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Mahmudul Hassan

____

Md. Mahmudul Hassan
ID: 171-15-8991
Department of CSE
Daffodil International University

Shahriar Shakil

____

Md. Shahriar Shakil
ID: 171-15-8558
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Supervisor Ms. Nazmun Nessa Moon and Co-Supervisor Ms. Nazmun Nessa Moon, Assistant Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning and Natural Language Processing" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Prof. Dr. Touhid Bhuiyan** and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

This research titled “Sentiment Analysis on Bangla Conversation Data Using Machine Learning Approach” is from conversations people's sentiment during the conversation period can be extracted as valuable information. In the field of NLP, text analysis and conclusion of any information as summarization can be done by Sentiment Analysis. The necessity of sentiment analysis of a conversation is increasing because of the use of conversation for customer support portal in many e-commerce platforms and crime investigations on digital evidence. Other languages, like English have enriched libraries and resources for natural language processing but there are very few works done over Bangla language. Because of the grammatical complexity in Bangla language, it is more difficult to extract sentiments from Bangla conversation data. That's why it opens the door of huge scopes of research. A machine learning approach was applied to extract sentiment from Bangla conversation. For that, Support Vector Machine, Multinomial Naïve Bayes, K-Nearest Neighbors, Logistic Regression, Decision Tree & Random Forest was used. From the dataset, extracted information was labeled as Positive and Negative.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

People have conversations in their daily life. People express their feelings and opinions in their conversations. These feelings and opinions can be categorized into sad, anger, happy, worried, disgusted, frightened, complement, motivation, suggestions, neutral etc. In our research work we merged them into two main categories of Positive and Negative. Our model can identify whether a part of any conversation is positive or negative. These two categories expose the sentiment of the people who said it. Analyzing sentiment from people's speech is a tough job because in a single sentence people can express various types of sentiment at the same time. Only the people who listen to it, can understand the sentiment properly. Our proposed model can extract sentiment from people's conversation with a closer accuracy of real life.

1.2 Motivation

Customer support and customer feedback plays an important role to analyze the performance, brand monitoring and reputation management of any organization. People's complaints and recommendations through online catboats are easier medium right now. That's why in every moment a lot of conversations are generated like phone call recording, email response, chat box etc. More than 228 million of people use Bangla as their first choice. People all around the world use Bangla language to communicate with nearest and dearest one in daily basis. Most of the popular social media have Bangla version of their web or android app platforms. From the platforms a large amount of Bangla conversation data generated in every seconds. So it is needed to extract people's sentiment from these conversations. Because of the availability of various social media, now it is more convenient to gather reviews or recommendations through conversations. In case of product analysis, people discuss among them about the uses of any products whether it is good or bad. Crime investigation can also be done as every digital crime has its own footprint that sometimes remains in conversation between suspect to victim or suspect to suspect.

1.3 Rationale of the Study

In this era Artificial Intelligence and Machine learning is an important field of CSE. Natural Language Processing is one of the popular field for text analysis and text summarizations. Like people Machine learns by itself and improves its performance by experience. Data and Machine learning algorithms are complementary to each other. The main goal of machine learning is to build a model that can perform well based on the dataset it feeds during the training procedure and this performance gives a promising accuracy of that model. Machine learning works on a large scale to solve critical and complex prediction and analysis tasks. In our work, we apply machine learning algorithms to extract sentiment from people’s conversation.

1.4 Research Questions

- Does it predict an actual output by given sample data with your system?
- Can it extract sentiment of Bangla conversation data using a machine learning algorithm?
- Does every algorithm work perfectly (yes/no)?

People's subconscious minds are frequently changing based on the environmental conditions and situations they faced. It is a critical task to trace the specific sentiment of specific moments from someone's conversations. We have trained our model with a huge amount of data and got good accuracy. We are so confident about our model that we can correctly predict accurate results whether the conversation is positive or negative.

We have used seven machine learning algorithms such as Support Vector Machine, Multinomial Naïve Bayes, K-Nearest Neighbors, Logistic Regression, Decision Tree, Stochastic Gradient Descent and Random Forest. And we have obtained a good result from SVM and Multinomial Naïve Bayes. However, other algorithms also give satisfactory performance but not as good performing algorithms.

1.5 Expected Outcome

In this research work we proposed a model that can extract sentiment from conversation as positive or negative sentiment. To pursue that we split our dataset into 80:20 ratio. For training purposes we used 80% data and for testing purposes we used 20% data. It helps to increase the accuracy of the model. Based on the training dataset the accuracy of the model fully depends on the training dataset. We have used some techniques such as changing the parameters of machine learning models to get more accurate results. We achieved about 86% accuracy on the Support Vector Machine. Rest of the algorithms perform closely to the highest accuracy.

1.6 Report Layout

Chapter 1 In this section we have talked about the inspiration of our task, targets, and furthermore the normal result of our undertaking.
Chapter 2 We have presented about the foundation of our work and examine about others related works, similar investigations, the extent of the issues and difficulties in this section.
Chapter 3 we are taking about our research subject and instrument used and also our data collection procedure and statistical analysis and also implementation.
Chapter 4 In this chapter we are taking out our research experimental result and descriptive analysis and find out summary.
Chapter 5 In this chapter we are talking about summary of prediction and conclusion and we have also added further study process.

CHAPTER 2

BACKGROUND

2.1 Introduction

Sentiment refers to a subjective response of a person. It means any physical response, mental pleasure or pain or repulsion. The needs of sentiment analysis are increasing day by day because it's widely used all over the world in many sectors. The uses of sentiment analysis lies on customer support, customer feedback, product analysis, crime investigation etc. Without this time worthy facility provided by sentiment analysis, it is quite impossible to analyze huge amounts of text data from different languages. As a human being people can easily understand that but for machines it is quite tough to extract exact emotion or sentiment from any conversation. Every conversation there are different types of sentiment. In our entire research work we tried to find out the people's sentiment that relies on their conversation.

In the upcoming part, we discuss our related works, project summary, challenges. In related works we will discuss the research works that have been done so far in this field. In the project summary and challenges section we will discuss the technical and analytical things along with the challenges to increase the accuracy of machine learning algorithms.

2.2 Related Works

Extracting sentiment from Bangla conversation data is an approach to classify any conversation whether it is positive or negative. In this paper [1] data was collected from Facebook groups through API. To simplify data, Tf-Idf, Bag of Words, word2vec methods used here. Tf-Idf counts the frequency of each word in the sample text. To find synonyms of each word to increase variation of train data word2vec was used. They labeled them into positive, negative and neutral. By applying a deep learning approach, in case of character level and word level mode they obtained about 80% and 77% accuracy to extract sentiment from their data. [2] They tried to establish a system that can detect fraud activity from real time messages in Facebook messenger. Content analysis, matrix pre-processing and natural language processing using semantic model and cosine similarity was also proposed to detect fraud activity. At the same time an android app was developed to analyze the chat content. In [3], they aimed to identify the sentiment from social media data. From the Facebook group post they collect data and two methods were applied to find the polarity of a post. Naive Bayes approaches used to lexical resources. They found lexicon based approach gives better performance in that specific domain. In paper [4], on Romanized Bangla and Bangla text that was collected from various social media a sentiment analysis was performed. With categorical cross entropy loss they obtained 78% accuracy by applying a RNN LSTM to train their model. This paper [5], tried to detect sentiment from twitter posts using a semi-supervised method. Firstly they used a rule based classifier to train data and split the post into positive and negative polarity. Support vector machine and Maximum Entropy algorithm used here and 93% accuracy achieved. This paper [6] proposed a model that can analyze paragraph text data. They used bag of words method and lexical analysis approach to extract sentiment from a paragraph. In this paper [7] they want to analyze twitter posts with the help of machine learning. Moreover, an element vector was presented to gather and infer people's feelings to these posts. Another paper [8] tried to analyze twitter post sentiment in a specific domain. They proposed a new feature vector that can classify positive and negative sentiment from twitter posts. This research work [9] used Naive Bayes and Decision tree machine learning algorithms to analyze twitter data for sentiment analysis. Their proposed model uses Apache spark, because it is scalable and fast. This paper [10] tried to analyze customer reviews of a restaurant applying some classifier based machine learning algorithms. From their dataset they obtained 94.56% accuracy on the SVM classifier.

2.3 Research Summary

In this research work we tried to develop a model that can extract sentiment from Bangla conversation data which is the most significant for this study. Support Vector Machine, Multinomial Naïve Bayes, K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest and Stochastic Gradient Descent implies to python to get outcome from our research work. In terms of communication, conversation is the significant medium to communicate. We differentiate the conversation into positive and negative sentiment. It was our approach to collaborate communication along with machine learning.

2.4 Challenges

To collect data we faced a lot of difficulties as Bangla language dataset and resources are not so available. That’s why we were challenged to collect our data. Data collection is not so easy in a manual manner. We have collected our data from Bangla movie and short film scripts as it is a huge source of Bangla conversation data. This procedure was not as easy as the movie and short film script has copyright issues. In this case our honorable supervisor ma’am refer us to script writers so that we can collect our data.

- Collecting proper conversation data
- Preprocessed them for machine learning model
- Collecting stop words for Bangla language
- Finding appropriate python library packages
- Improving the model accuracy

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The research methodology and procedures will be described in this chapter. In addition, tools for the research project, data collection, data pre-processing, model selection, statistical analysis, and its implementation will be discussed in this session. In Figure 3.1, we can see the full methodology at a glance.

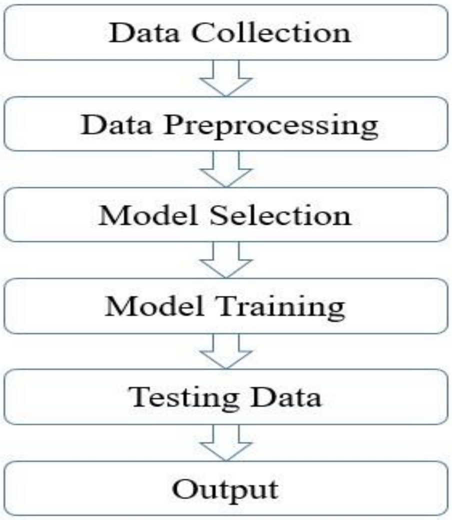


Figure 3.1: Methodology at a Glance

3.2 Research Subject and Instrumentation

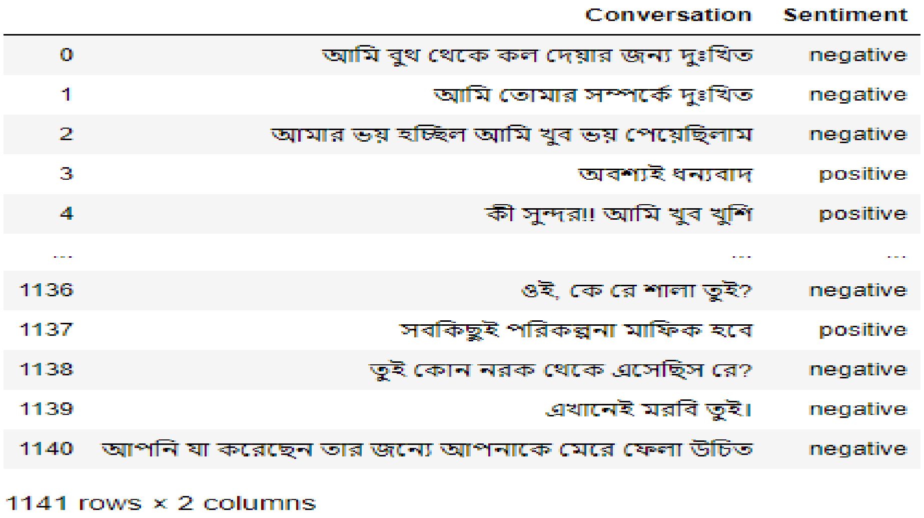
To find appropriate research methods, research instruments play a vital role. It helps the researcher to select which method keeps good pace with their work. Every research work follows a common manner to gain the expected outcome which denotes the research objective. To accomplish that several terms need to classify:

- Which data should be collected?
- How to ensure that the collected data are okay?
- How should each data be organized?

- How should each data be labeled?

3.3 Data Collection Procedure

From various Bangla movie and short film scripts we collected conversation data for our research work. These conversations covered a large scale of topic like food, family, motivation, fraud, business, friends etc.



	Conversation	Sentiment
0	আমি বুথ থেকে কল দেয়ার জন্য দুঃখিত	negative
1	আমি তোমার সম্পর্কে দুঃখিত	negative
2	আমার ভয় হচ্ছিল আমি খুব ভয় পেয়েছিলাম	negative
3	অবশ্যই ধন্যবাদ	positive
4	কী সুন্দর!! আমি খুব খুশি	positive
...	...	...
1136	ওই, কে রে শালা তুই?	negative
1137	সবকিছুই পরিকল্পনা মাফিক হবে	positive
1138	তুই কোন নরক থেকে এসেছিস রে?	negative
1139	এখানেই মরবি তুই।	negative
1140	আপনি যা করেছেন তার জন্যে আপনাকে মেরে ফেলা উচিত	negative

1141 rows × 2 columns

Figure 3.2: Sample Data

We have collected about 1141 data. These conversations include emotions like happy, sad, anger, worried, afraid etc. These categories help us to differentiate the whole dataset into two main categories of Positive and Negative. Among 1141 data there was 570 data for positive sentiment and for negative it was 571 data. Figure 3.2 shows the sample dataset.

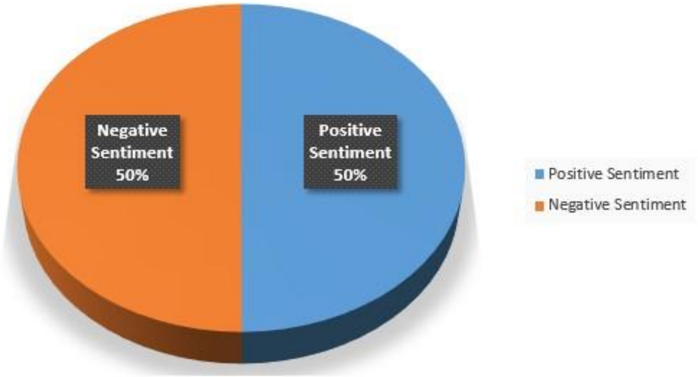
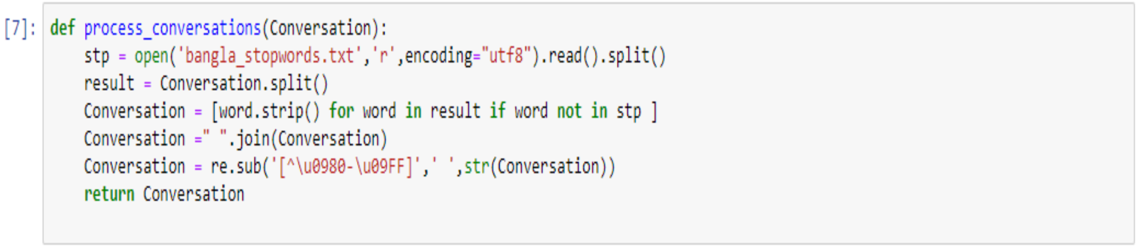


Figure 3.3: Class Label Distribution

3.3.1 Data Preprocessing and Organizing

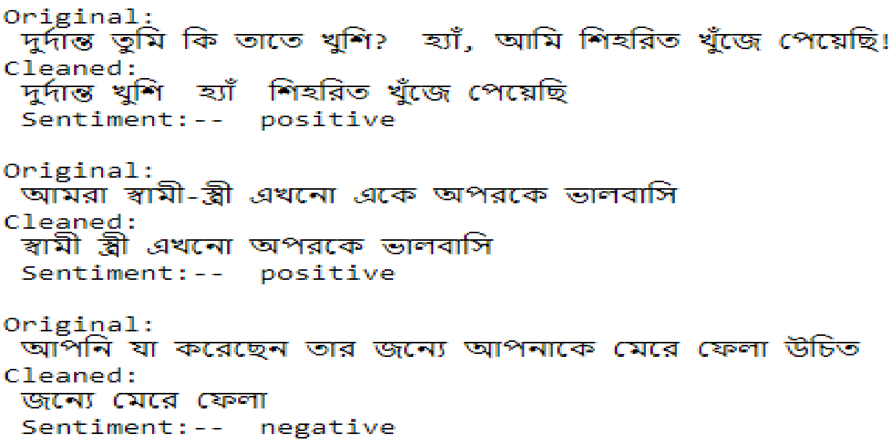
Firstly, we collect data from scripts and store them into an xlsx file. The dataset we have collected has two attributes. These are Positive and Negative. As we already discussed, we collect data from movie and short film scripts as conversation. Every conversation starts with a single word or single sentence. People can express their feelings, emotions and thoughts through a single word or sentence. To classify these expressions into two main attributes we merged happy, joy, motivation and thankfulness into Positive conversations and for Negative conversation we merged sad, anger, backbiting and worries.

During pre-processing, we remove punctuation in the first step. In Natural Language Processing, for every language it is essential to identify and remove stop words. For our research work we have collected Bangla stop words and removed them to clean our data. There were about 410 stop words in Bangla language. For example: ‘অতএব’, ‘অথচ’, ‘এই’, ‘একই’, ‘একটি’, ‘হয়’, ‘হয়তো’, ‘কিন্তু’, ‘কী’, ‘কে’ etc.



```
[7]: def process_conversations(Conversation):
         stp = open('bangla_stopwords.txt','r',encoding="utf8").read().split()
         result = Conversation.split()
         Conversation = [word.strip() for word in result if word not in stp ]
         Conversation =" ".join(Conversation)
         Conversation = re.sub('[^\u0980-\u09FF]',' ',str(Conversation))
         return Conversation
```

Fig 3.4: Removing Stop Words and Punctuations



```
Original:
 দুর্দান্ত তুমি কি তাতে খুশি?  হ্যাঁ, আমি শিহরিত খুঁজে পেয়েছি!
Cleaned:
 দুর্দান্ত খুশি  হ্যাঁ  শিহরিত খুঁজে পেয়েছি
 Sentiment:--  positive

Original:
 আমরা স্বামী-স্ত্রী এখনো একে অপরকে ভালবাসি
Cleaned:
 স্বামী স্ত্রী এখনো অপরকে ভালবাসি
 Sentiment:--  positive

Original:
 আপনি যা করেছেন তার জন্যে আপনাকে মেরে ফেলা উচিত
Cleaned:
 জন্যে মেরে ফেলা
 Sentiment:--  negative
```

Fig 3.5: Cleaned Data

Fig 3.4 shows the python code for removing Bangla stop words and punctuations. Fig 3.5 shows the cleaned data what we pre-processed.

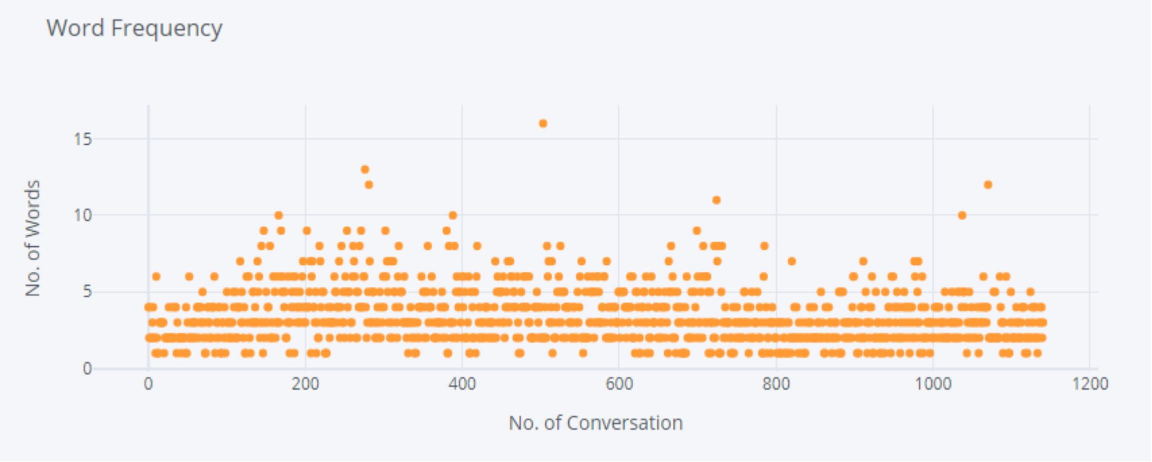


Fig 3.6: Word Frequency

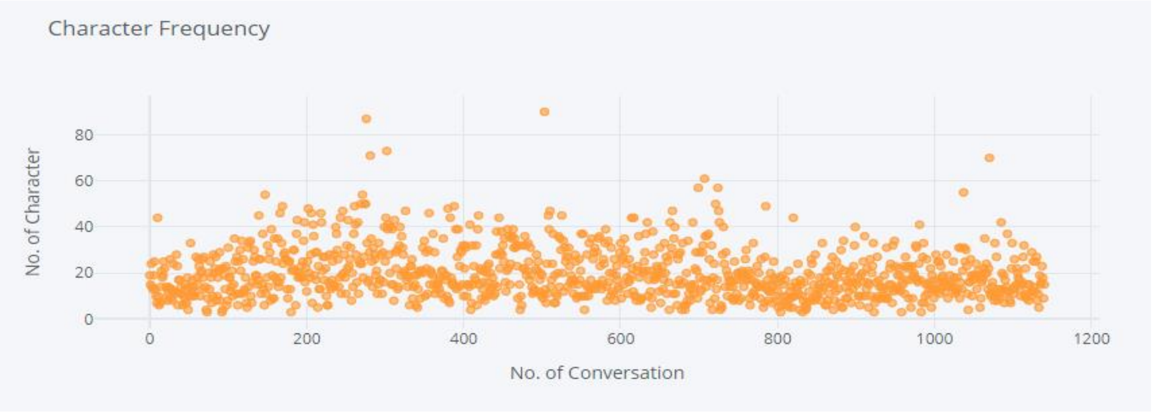
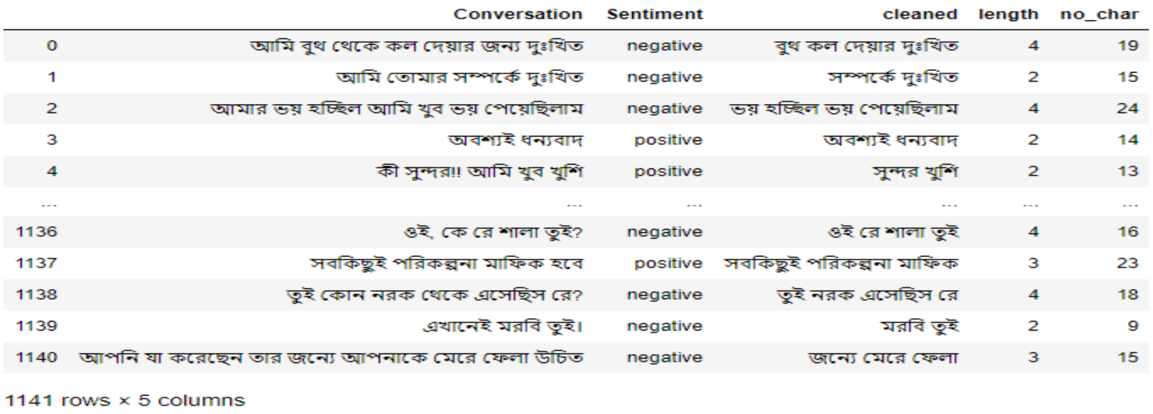


Fig 3.7: Character Frequency

To extract feature from each of the conversation, number of words and number of characters is needed. Fig 3.6 and Fig 3.7 shows the result respectively.

After preprocessing procedure label encoding method applied to the sentiment column. And then a pickle file generated. Pickle file contains temporary data for reuse and also saves time during runtime execution. In this work, our cleaned data is stored as a pickle file for upcoming procedures. We need to demonstrate our dataset data where highlights are age, Occupation, house type, want switch job and we are giving low highlighting to other attributes. In Fig 3.8, cleaned data along with counts of each conversation length and character is shown.



	Conversation	Sentiment	cleaned	length	no_char
0	আমি বুথ থেকে কল দেয়ার জন্য দুঃখিত	negative	বুথ কল দেয়ার দুঃখিত	4	19
1	আমি তোমার সম্পর্কে দুঃখিত	negative	সম্পর্কে দুঃখিত	2	15
2	আমার ভয় হচ্ছিল আমি খুব ভয় পেয়েছিলাম	negative	ভয় হচ্ছিল ভয় পেয়েছিলাম	4	24
3	অবশ্যই ধন্যবাদ	positive	অবশ্যই ধন্যবাদ	2	14
4	কী সুন্দর!! আমি খুব খুশি	positive	সুন্দর খুশি	2	13
...	...	...	...	...	...
1136	ওই, কে রে শালা তুই?	negative	ওই রে শালা তুই	4	16
1137	সবকিছুই পরিকল্পনা মাফিক হবে	positive	সবকিছুই পরিকল্পনা মাফিক	3	23
1138	তুই কোন নরক থেকে এসেছিস রে?	negative	তুই নরক এসেছিস রে	4	18
1139	এখানেই মরবি তুই।	negative	মরবি তুই	2	9
1140	আপনি যা করেছেন তার জন্যে আপনাকে মেরে ফেলা উচিত	negative	জন্যে মেরে ফেলা	3	15

1141 rows × 5 columns

Fig 3.8: Sample of Cleaned Dataset

In case of data summary, we extract data statistics along with respective class names. Fig 3.8 shows number of total words, documents and total number of unique words in our dataset.

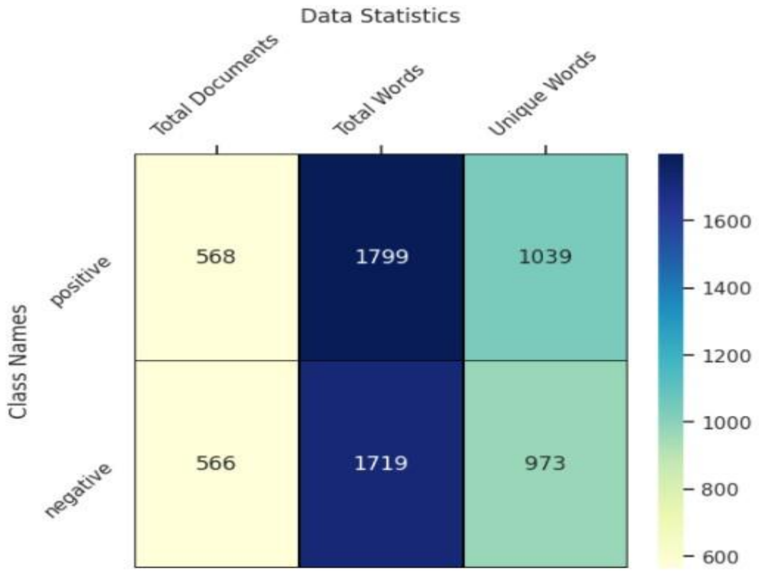


Fig 3.9: Data Statistics of Cleaned Data

3.4 Machine Learning Algorithms

To know the accuracy on our dataset we applied some classifier based algorithms. These are Support Vector Machine, Multinomial Naïve Bayes, K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest and Stochastic Gradient Descent.

Decision Tree classifier works like a flowchart as a tree structure, where each internal node indicates test on an attribute, every single branch represents an outcome of the test, and every leaf node holds a class label.

Among all of the machine learning algorithms **K-NN** is the simplest one. The motivation behind the algorithm is to find a predefined number of training samples closest in distance to the new point and also predict the label from these.

Support Vector Machine algorithm is a supervised machine learning algorithm. It can solve both classification and regression problems. It is also suitable for both linear and nonlinear separable data. We can make the result more efficient by using kernel tricks.

Random Forest is the well-known learning method for classification and regression. During training time, it constructs a number of decision trees. It sends the new case to each of the trees to classify the new case.

Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It is a technique based on Bayes' Theorem. This model is easy to build and particularly useful for very large datasets. Naïve Bayes is known to outperform even highly sophisticated classification methods.

Logistic Regression model can build a model of probability from a class or event. For example, to determine an image category that contains photos of various animal that can be explored to a model of several classes.

Stochastic Gradient Descent is known for optimizing any algorithm that mainly propagated in machine learning algorithms in order to find the related parameter of the model to fit predicted and actual outputs.

Needed all algorithms and libraries are imported. Used libraries are Pandas, Numpy, Tensor flow, Keras, Seaborn, Pyplot, Scikit-learn, Tf-idf, Matplotlib etc.

3.4.1 Statistical Analysis

About 571 records for positive and 570 records are for negative conversations in our dataset. For the dataset splitting purpose we used train-test split function. We followed supervised machine learning techniques. To train our model we used 80% of our data and for test 20% of data used. In number 912 data used for trains and 229 data used for test purposes. In figure 3.6, a dataset flowchart along with the uses of the dataset for our proposed model. Support Vector Machine, Multinomial Naïve Bayes, K-Nearest Neighbors, Logistic Regression, Decision Tree, Random Forest and Stochastic Gradient Descent algorithms are applied to build our model.

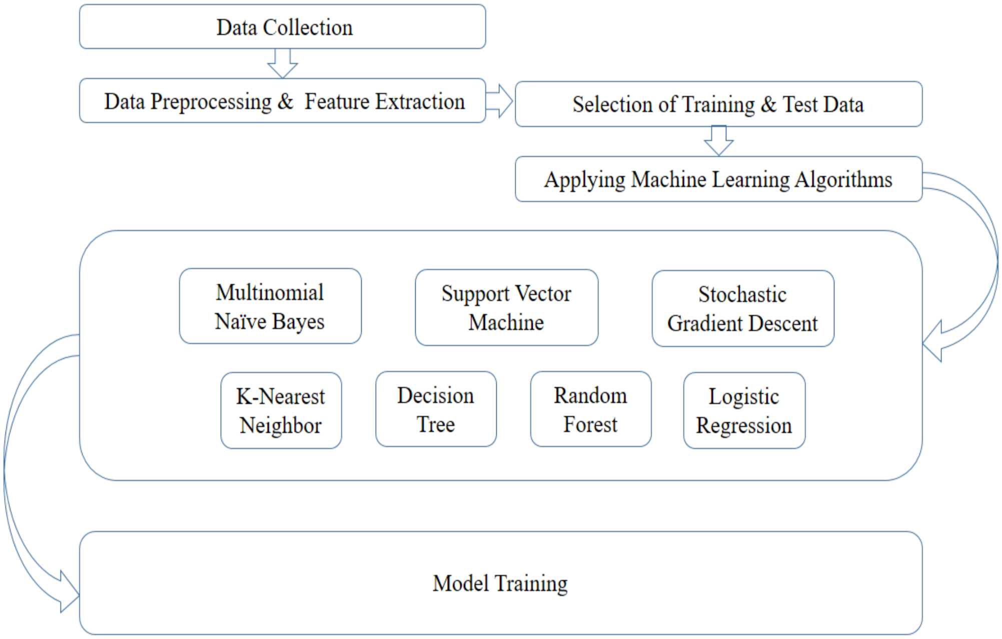


Figure 3.10: Proposed Model Structure

In this figure, we have shown that how we have done our research shortly details. In this figure (3.6), we can know how to go ahead for our target step by step.

3.5 Implementation Requirements

- **Python 3.9:** Python 3.9 is the latest version till now. It is a high-level programming language. Most of the researchers use it to do their research.
- **Google Colab:** Google Colab is a free to use open-source distributor of Python programming language. For online teamwork purposes we use it.
- **Anaconda:** For scientific computations Anaconda is an offline distribution for Python and R programming language. We used Jupyter Notebook for faster execution.
- **Operating System:** Windows (7, 8.1 or 10).

- **Browser:** Google Chrome.
- **RAM & ROM:** Hard Disk (Minimum 4 GB), ROM (Minimum 4 GB)

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Setup

To create our model and to execute codes some information is needed. The following setup process are maintained:

- As we want to extract sentiment from conversations the first and foremost task is to collect conversation data.
- It was a challenge for us to collect our desired data. We tried to make the dataset more suitable and subjective to our aims. For that we collect conversation data from Bangla movie and short film scripts.
- All possible types of conversation exist in communication systems founded from the data resource.
- We collect them as sentences and store them into an xlsx file by naming the conversation type whether it is positive or negative conversation.
- After preprocessing process data is usable for utilizations.
- This was the finished line point and we started the further preparation

4.2 Model Summary

Here in this work we use machine learning algorithms to pursue Natural Language Processing objectives. From two main attributes our model gets trained by extracting all features of each sentence. For this procedure a method named tokenizer is introduced here. Tokenizer breaks sentences into segments of words. These unique or more frequent words make attributes more identical.

Moreover, Tf-idf is also such a numerical statistic that explores the necessity of a word in a document. Some of the relevant works follow this method for different languages. Their success influenced us and we found out the best accuracy from our machine learning algorithms.

4.3 Experimental Result and Analysis

According to our requirement we update our model and dataset. From this modification, we can accomplish that our used classifier is exactly usable for a wide range of use according to our dataset. As our expectations we achieved 86% accuracy from our proposed mode which is a fruitful outcome. This performance of the model creates a path to think about the improvement in results.

The research result was focused to identify whether a conversation is positive or negative. We have applied classifiers based on different machine learning models to extract the conversation type. The result has two criteria of "positive" and "negative". There were 1141 data for training each of the models. We get various accuracy on different models. Among 7 models the Support Vector Machine (SVM) and Multinomial Naive Bayes perform well with highest accuracy. As we already discussed, we collect data from scripts as conversation. All conversations have people's emotions like, happy, sad, worries, annoyed, motivation etc. We merged and categorized them into two main types, positive and negative. The decision making capability of the classifiers was measured by their performance. Accuracy, precision, recall and F-score were used to determine the performance of classifiers. For a classifier the overall accuracy was considered as adequate standard. In the test set it is necessary to have a notion of the correctly classified samples.

Table 4.1: Accuracy of Classifiers

Classifier	Accuracy
Random Forest	74.24%
Decision Tree	76.42%
Logistic Regression	82.53%
KNN	82.97%
SGD	83.41%
Multinomial Naïve Bayes	85.15%
SVM	85.59%

In Table 4.1 the accuracy scores obtained for the classifiers built are given. Here it is clear that the Support Vector Machine gives the highest accuracy score of 0.85589 and

Multinomial Naive Bayes gives almost similar accuracy of 0.8513. That's why it was needed to calculate the other performance measures to decide a suitable classifier for our dataset.

Table 4.2: Precision of Classifiers

Classifier	Precision
Random Forest	67.01%
Decision Tree	69.62%
Logistic Regression	79.23%
KNN	79.39%
SGD	79.55%
Multinomial Naïve Bayes	85.96%
SVM	81.68%

To measure the class agreement of the data labels with the positive labels given by the classifier the precision is used. We have to calculate the precision scores for each of the two class labels because it is directly relevant to class labels. In table 4.2 the values for each of the classifiers is given along with the 2 labels we used in this research work. We can see that the classifier Random Forest gives a score of 0.93 and Multinomial Naive Bayes gives 0.85 for positive conversation.

Table 4.3: Recall of Classifiers

Classifier	Recall
Random Forest	96.55%
Decision Tree	94.83%
Logistic Regression	88.79%
KNN	89.66%
SGD	90.52%
Multinomial Naïve Bayes	84.48%
SVM	92.24%

To identify class labels Recall is known as sensitivity of the measurement that represents the effectiveness of the classifier. We also focused here to get a score close to 1 for the positive class label. In table 4.3 the recall scores for two class labels and the classifiers are shown. For positive conversation the recall score 0.92 was obtained for Decision tree and Support Vector Machine.

The relationship between positive labels and those given by the classifier can be obtained by F1-score. We can calculate it by taking the harmonic mean of precision and recall for all the 2 labels across all the classifiers. For deciding the best model of classifier the score close to 1 for the positive class label was considered. In table 4.4 the F1-scores for the class labels are shown. The classifiers Support Vector Machine and Multinomial Naive Bayes and Stochastic Gradient Descent have the best to decide the optimal classifier for our dataset.

Table 4.4: F1-Score of Classifiers

Classifier	F1-Score
Random Forest	79.15%
Decision Tree	80.29%
Logistic Regression	83.74%
KNN	84.21%
SGD	84.68%
Multinomial Naïve Bayes	85.22%
SVM	86.64%

Our objective is to extract sentiment from Bangla conversation data with higher precision which was achieved by Random Forest, Multinomial Naïve Bayes and Support Vector Machine (SVM). With remarkable accuracy SVM, Multinomial Naïve Bayes and SGD perform well among the classifiers. It is clearly seen from the tables SVM, Multinomial Naive Bayes and Random Forest give the best performance as individual classifiers. As our dataset is much more concise and because the labels are weakly known Support vector machines work well for the problem. And as there are fewer dimensions or features K-Nearest Neighbor works well. The assumption of class conditional independence will work only for a huge dataset; that's why the Decision tree does not give a good performance here.

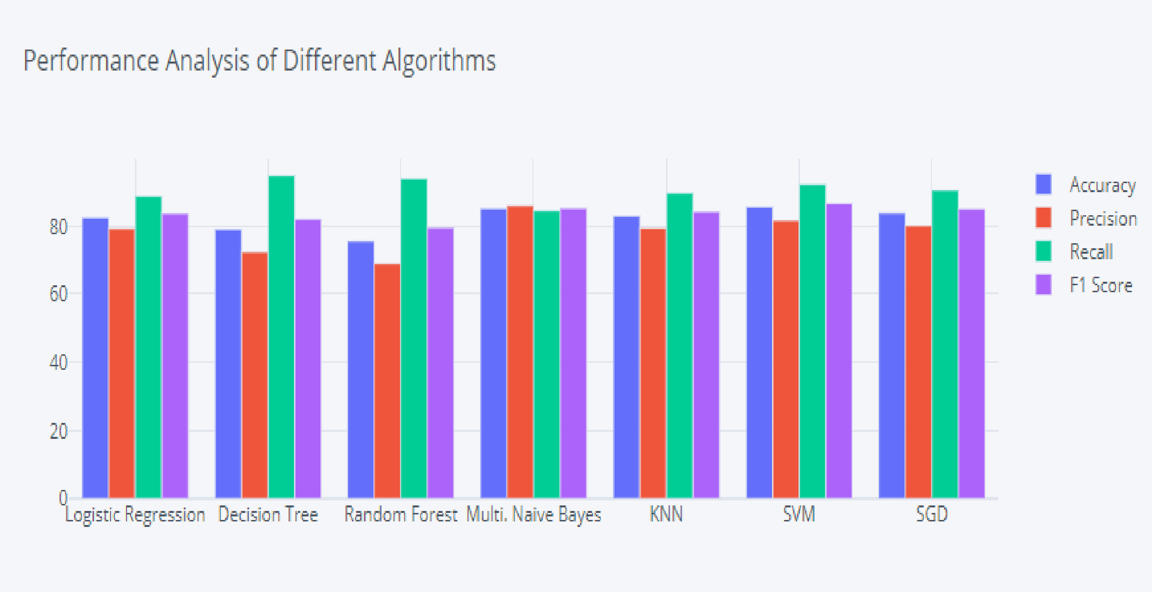
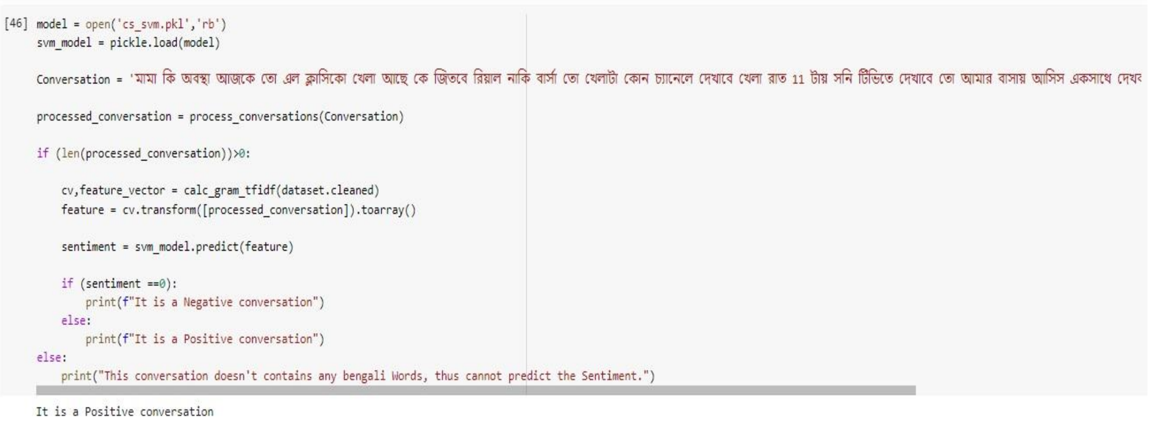


Fig 4.1: Performance Analysis

To avoid over fitting and robustness, it is needed to have a strong correlation over fitting nuts; though it is not exceptional. As it is not robust to noise and does not generalize well, future observed data Decision trees do not work too well. In Fig 4.1 the overall performance comparison is shown.

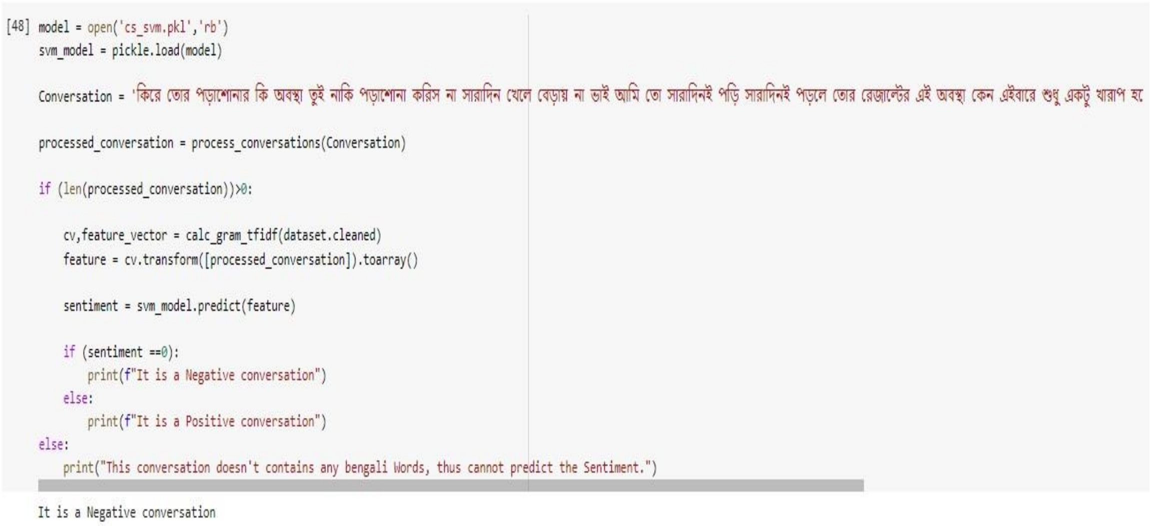
4.3.1 Prediction

We have tried to test our model by using a random conversation data and we got a result. In Figure 4.2 and 4.3, we can see that our proposed model can extract sentiment from Bangla conversation data.



```
[46] model = open('cs_svm.pkl','rb')
     svm_model = pickle.load(model)

     Conversation = 'মামা কি অবস্থা আজকে তো এল ক্লাসিকো খেলা আছে কে জিতবে রিয়াল নাকি বার্সা তো খেলাটা কোন চ্যানেলে দেখাবে খেলা রাত 11 টায় সনি টিভিতে দেখাবে তো আমার বাসায় আসিস একসাথে দেখব

     processed_conversation = process_conversations(Conversation)

     if (len(processed_conversation))>0:

         cv,feature_vector = calc_gram_tfidf(dataset.cleaned)
         feature = cv.transform([processed_conversation]).toarray()

         sentiment = svm_model.predict(feature)

         if (sentiment ==0):
             print(f"It is a Negative conversation")
         else:
             print(f"It is a Positive conversation")
     else:
         print("This conversation doesn't contains any bengali Words, thus cannot predict the Sentiment.")

     It is a Positive conversation
```

Figure 4.2: Predicting Positive Conversation



```
[48] model = open('cs_svm.pkl','rb')
     svm_model = pickle.load(model)

     Conversation = 'কিরে তোর পড়াশোনার কি অবস্থা তুই নাকি পড়াশোনা করিস না সারাদিন খেলে বেড়ায় না ভাই আমি তো সারাদিনই পড়ি সারাদিনই পড়লে তোর রেজাল্টের এই অবস্থা কেন এইবারে শুধু একটু খারাপ হ

     processed_conversation = process_conversations(Conversation)

     if (len(processed_conversation))>0:

         cv,feature_vector = calc_gram_tfidf(dataset.cleaned)
         feature = cv.transform([processed_conversation]).toarray()

         sentiment = svm_model.predict(feature)

         if (sentiment ==0):
             print(f"It is a Negative conversation")
         else:
             print(f"It is a Positive conversation")
     else:
         print("This conversation doesn't contains any bengali Words, thus cannot predict the Sentiment.")

     It is a Negative conversation
```

Figure 4.3: Predicting Negative Conversation

4.4 Discussion

According to our requirement we update our model and dataset. From this modification, we can accomplish that our used classifier is exactly usable for a wide range of use according to our dataset. As our expectations we achieved 86% accuracy from our proposed mode which is a fruitful outcome. This performance of the model creates a path to think about the improvement in results.

CHAPTER 5

SUMMARY, CONCLUSION, IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the Study

Conversation is a common medium used in our daily life and it is an essential part of communication. Each conversation contains people’s feelings, emotions and thinking. Sentiment analysis is becoming a crucial term in the digital era. In every moment all over the world a large amount of conversations are generated. To ensure best customer support, informative crime investigation it is needed to extract sentiment from conversation data. In our research, we tried to extract sentiment from a conversation with the help of various machine learning algorithms. Our proposed model can extract people’s sentiment from their conversation. We believed this outcome of our research work opens a door to the scope of conversation data analysis.

5.2 Conclusion

This research work concludes with an expected outcome of extracting sentiment from Bangla conversation data. Text mining and text analysis are very new terms in Bangla language. Though it's a tough task to work with some limitations, lacking the resources we tried to overcome these difficulties. Technology makes the communication sector easier with advancement. But embracing the advancement by ensuring the control of enormous data is necessary for us. We should be concerned about these terminologies to make the world of data more accessible and convenient.

5.3 Future Work

This research work proposes a methodology that finds the scopes to work with Bangla conversation data. To accomplish that, machine learning models were trained from Bangla conversation data and able to extract sentiment from those conversations. There is a scope to apply a deep learning approach in our dataset to improve the efficiency. Here in this work we extract sentiment as a positive and negative category. But on a large scale, people’s emotions and sentiments as an individual like sadness, anger, neutral, happiness, fear can also be extracted. For real time conversation data, by converting real time conversations into text and analysis sentiment from these conversations can also be done. However, scope lies in every possible opportunity. And opportunity revealed innovation and evolutions.

REFERENCES

[1] M. S. Haydar, M. Al Helal and S. A. Hossain, "Sentiment Extraction From Bangla Text : A Character Level Supervised Recurrent Neural Network Approach," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465606.

[2] Kuo-Hui Yeh, Nai-Wei Lo, Lin-Chih Chen and Ping-Hsien Lin, “A fraud detection system for real-time messaging communication on Android Facebook messenger,” 2015 IEEE 4th Global Conference on Consumer Electronics (GCCE), DOI: 10.1109/GCCE.2015.7398737, Osaka, Japan, 2015.

[3] Akter, Sanjida, and Muhammad Tareq Aziz. ”Sentiment analysis on facebook group using lexicon based approach.” Electrical Engineering and Information Communication Technology (ICEEICT), 2016 3rd International Conference on. IEEE, 2016.

[4] Hassan, Asif, et al. “Sentiment analysis on bangla and romanized bangla text using deep recurrent models.” Computational Intelligence (IWCI), International Workshop on. IEEE, 2016.

[5] Chowdhury, Shaika, and Wasifa Chowdhury. ”Performing sentiment analysis in Bangla microblog posts.” Informatics, Electronics & Vision (ICIEV), 2014 International Conference on. IEEE, 2014.

[6] Chowdhury, S. M. Mazharul Hoque & Abujar, Sheikh & Saifuzzaman, Mohd & Ghosh, Priyanka & Hossain, Syed. (2018). Sentiment Prediction Based on Lexical Analysis Using Deep Learning.

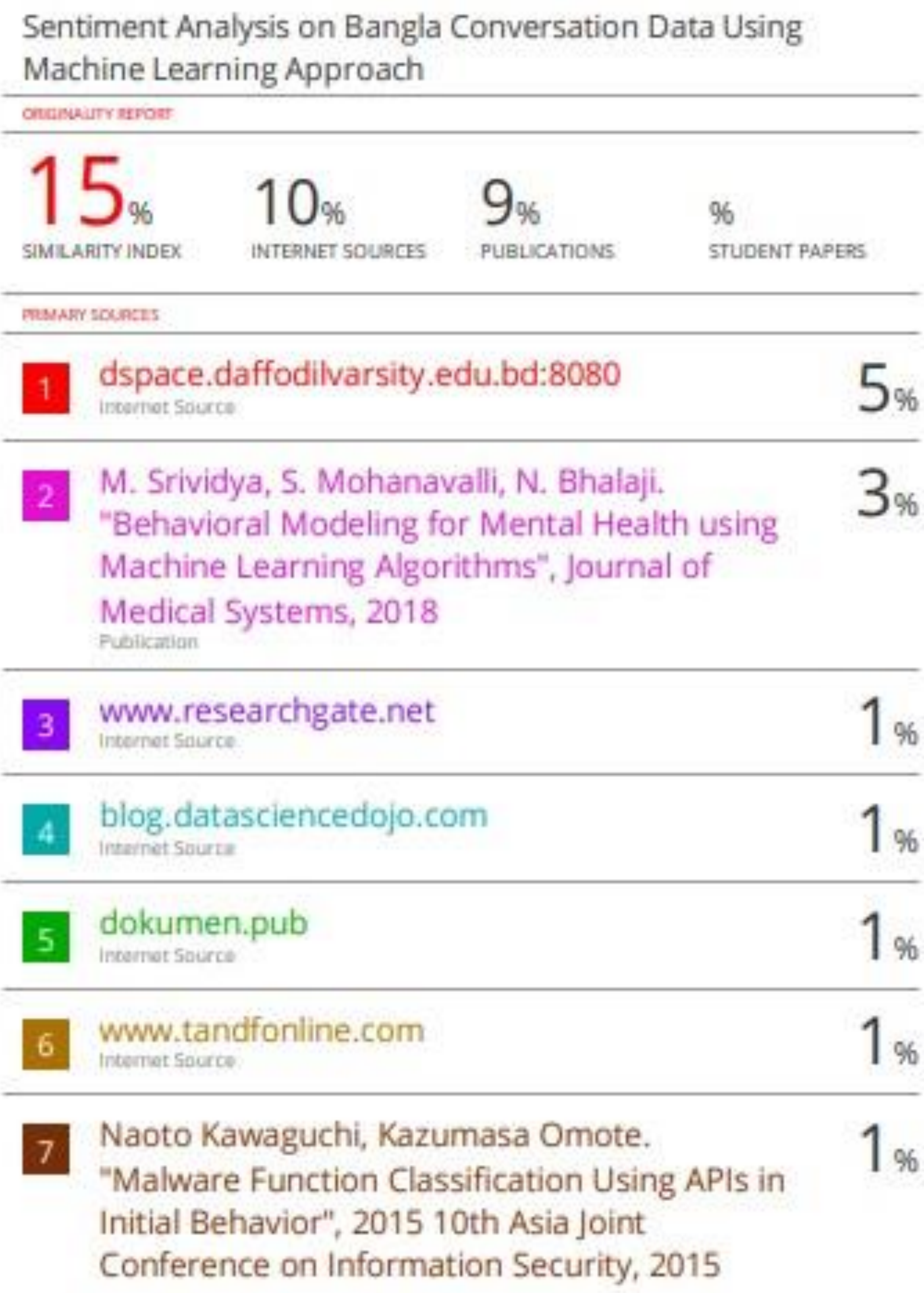
[7] Chandhok, Surbhi & Anand, Romil & Gupta, Soumay & Jamshed, Aatif. (2017). An Analysis of Sentimental Data using Machine Learning Techniques. International Journal of Computer Applications. 166. 35-39. 10.5120/ijca2017913955.

[8] Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). doi:10.1109/icccnt.2013.6726818.

[9] Jain, A. P., & Dandannavar, P. (2016). Application of machine learning techniques to sentiment analysis. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). doi:10.1109/icatcct.2016.7912076.

[10] Krishna, A., Akhilesh, V., Aich, A., & Hegde, C. (2019). Sentiment Analysis of Restaurant Reviews Using Machine Learning Techniques. Die Anästhesiologie, 687–696. doi:10.1007/978-981-13-5802-9_60.

Plagiarism Report



Sentiment Analysis on Bangla Conversation Data Using Machine Learning Approach

ORIGINALITY REPORT

15%	10%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	M. Srividya, S. Mohanavalli, N. Bhalaji. "Behavioral Modeling for Mental Health using Machine Learning Algorithms", Journal of Medical Systems, 2018 Publication	3%
3	www.researchgate.net Internet Source	1%
4	blog.datasciencedojo.com Internet Source	1%
5	dokumen.pub Internet Source	1%
6	www.tandfonline.com Internet Source	1%
7	Naoto Kawaguchi, Kazumasa Omote. "Malware Function Classification Using APIs in Initial Behavior", 2015 10th Asia Joint Conference on Information Security, 2015	1%