# East West University

**Department of Electronics and Communications Engineering**

## Course Code: ICE498

## Course Title: Undergraduate Research Project

## Title

# Sentiment Extraction in Bangla Language Using Unsupervised Learning Approach

# Declaration

We hereby declare that this research project record is an authentic piece of work carried out with the aid of us, below the education and supervision of Dr. Mohammad Arifuzzaman. This file is the requirement for the successive completion of B.Sc. in Information and communications Engineering beneath the branch of Electronics and Communications Engineering.

We nation that the record along with its literature that has been established two in this report papers, is our own work with the masterly education and fruitful help of our supervisor for the finalization of our document efficaciously.

Signature of Student

__________________

Most. Afrin Al Jannat

ID:2016-2-50-017

Signature of Student

__________________

Shaon Sikder

ID:2016-2-50-026

Signature of Student

__________________

Ismail Hossen

ID:2016-2-50-027

Signature of   Supervisor

_______________________

Dr. Mohammad Arifuzzaman

Assistant Professor

Department of ECE

# Acknowledgement

We would like to specifics our gratitude and understanding to all these who gave us the opportunity to whole this research work. A extraordinary thanks to our supervisor Dr. Mohammad Arifuzzaman, whose help, suggestions and encouragements helped us to take our thesis specially on Data mining, we have been craving to work on it for so long. He supported us via showing one of a kind method of facts collection whilst doing this work. He constantly helped us when required and he gave required course toward completion of this work.

We additionally want to thank all faculty members and staffs of the Department of Electronics and Communications Engineering of East West University for their full cooperation and support to complete our graduation degree.

# Approval

This report on “Sentiment Extraction in Bangla Language using Unsupervised Learning Approach” ,submitted by Most. Afrin Al Jannat, ID:2016-2-50-017, Shaon Sikder, ID:2016-2-50-026, Ismail Hossen, ID:2016-2-50-027 to Dept. of ECE, East West University. It is submitted in some achievement of the requirement for the degree of B.Sc. in Information and Communications Engineering.

# Table of contents

# List of Figures

# Abstract

In this thesis, we mentioned a system that assigns scores indicating positive or negative to translated data. Much works have been done on sentiment analysis, document clustering for newspaper data in bangla language. News from one Bengali newspaper is used for the purpose of the project. we are using web crawler to get necessary news to make a dataset to use for this project. With the significant increase of person interactions through outstanding advances of the Web, sentiment analysis has acquired more focus from an educational and a industrial point of view. Recently, sentiment analysis in the Bangla language is step by step being regarded as an vital task, for which previous methods have tried to notice the universal polarity of a Bangla document. This can be described as being due to the lack of reachable datasets for bangle text analysis .These high massive unstructured web contents will be utilized to create smarter tools to help people by natural language Processing (NLP). Though Bengali NLP tools are still insufficient because of its natural complexities, research on Sentiment Analysis in Bengali is flourishing as a challenging area and is getting researcher's attention at a rapid pace. In this paper, we apply publicly reachable datasets to operate sentiment analysis in Bangla text. One of the datasets consists of human-annotated person remarks on dataset consists of person opinions of overall review. We additionally describe a baseline approach for the subtask of component class extraction to consider our datasets. Through this thesis, our destination is to establish a system, which can identify positive an negative news accurately.

# Chapter 1

# Introduction

## 1.1 Introduction

Bangla is spoken as the first language by using almost 200 million human beings worldwide, one hundred sixty million whom are Bangladeshi. Bangladeshi humans are determined to get more and more concerned in online activities such as - getting linked to friends and families via expressing their opinions and ideas on popular - running a blog and social networking sites, sharing opinions and thoughts by means of comments on reviewsof newspaper, doing purchasing through marketplaces and other such application. Sentiment analysis (or opinion mining) is a technique to determine the point of view of a man or woman on a certain topic (1). It classifies the polarity of a record (i.e. review, tweet, blog, or news), that is, whether the communicated opinion is positive, negative, or neutral. There are three stages at which sentiment is analyzed the report level, sentence level, and thing level. The record level considers that a file has an opinion on an entity, and the project is to classify whether an whole report expresses a positive or negative sentiment. The project at the sentence level regards sentences and identifying whether every sentence expresses a positive, negative, or neutral opinion (2). Neither the record level nor the sentence level analysis find out precisely what humans preferred and not like. The thing performs a finergrained analysis that identifies the components of a given report or sentence and the sentiment expressed towards every aspect. This level of analysis is the most elaborate version that is susceptible of discovering complicated opinions from newspaper reviews. Since sentiment analyses are section of the data mining that can take a look at public reduce about a variety of opinion.

Sentiment analysis, also called opinion mining, is the field of study that explorepeople’s opinions, sentiments, evaluations and emotions towards entities such as services, organizations, individuals, issues, events, topics, and theirattributes. For our work, we are focusing on newspaper data. News can be positive, negative, or neutral. Although full considerate of natural text language is well by capabilities of machines, statistical analysis of relatively simple statement can give surprisingly meaningful categorization of positive and negative sentiment. Analyzed data then can be used in much sectors. It can be used to predict financial state of a country and other predictions. The sentiment analysis is beneficial in several approaches and consists of many branches of computer science such as Natural Network, data mining, Gaming theory and coding, machine learning (3).

## 1.2 Sentiment Analysis

Sentiment analysis models track out polarity between a textual content, whether it’s a entire document, paragraph, sentence, or clause. By inspecting people feedback, from survey responses to people opinion.For example, one of our customers used sentiment analysis to automatically analyze opinions about their review, and found that clients have been completely satisfied about their quality but complained a lot about their client service (1) (2). There are some types of sentiment analysis: subjectivity/objectivity identification and feature sentiment analysis.

Sentiment analysis is conversant as data mining, data analysis or emotion in Artificial Intelligent . It is a present research topic in machine learning. It is the stem of report topic is applied to the newspaper data (4). It can be removed, evaluated and recognized from the opinion and survey responses, all social media, and health-care materials. Basically it deals with the user understanding about individual fact,sharing their concept and opinion with a summary message via different network , data analysis, Gaming theory, machine learning methods (5). It is an important social networking that brings extent for developing originative client service solutions. Sentiment analyses of text which is take out by different technologies and data mining techniques. The sentiment analyses the company have got idea about the review how positive or not are people, political events can be recognized how a lot human help their work. These sentiment analysis is evaluated by classifying the polarity of a text file. The classifier is evaluating a text by positive, or not polarity or showing emotional .

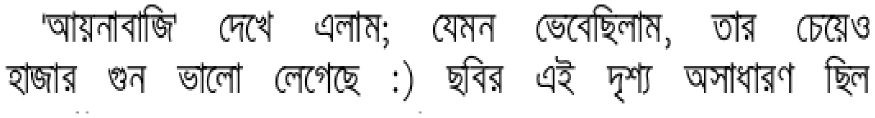
Emotion detection goals at detecting emotions, like happy, sad, anger, and so on. Many emotion detection structures use lexicons (i.e. lists of words and the feelings they convey) or complex machine learning algorithms.

## 1.3 Sentiment Analysis for Bangla Text

News Articles can be collected daily by manually copying the news that are mostly available in every news website like prothomalo, kaler kantho, jugantor, samakal News etc. By following as data sources, these websites facilitate the research on knowing the polarity of the news articles that can be helpful to readers as this can be beneficial to them (1). For our thesis, we have chosen to analyze the sentiment of Bangla newspaper data. Bangla is one of the highest spoken language, ranked seventh in the Page 10 of 37 world, but surprisingly very several works had been done with Bangla sentiment analysis using unsupervised learning approach. It is quite unfortunate that there is no general collection of data, such as datasets for Bangla texts. One effort for standardization came from an automatic comment of positive and negative opinion. However, no corpse used to be created from this work, thereby limiting its usage to dedication of sentiment, as an alternative than the more complicated two natural language processing methods (5).

A big amount of bangla text are used for this work and also used Bangla datasets for training. This data was annotated manually by using native speakers. However, in terms of usability the dataset's small measurement is a limiting aspect for modern deep learning techniques (4). Bangla Datasets were gathered automatically. However, their dataset is publicly available, and the dimension of the dataset is as an alternative small.

With most of the other works proceeded in the comparable way, the largest problems with the present day nation of affairs in Bangla Sentiment Analysis research are - first and foremost, the absence of a standard and large sufficient dataset to evaluate against, which makes evaluation of research work extremely difficult, and secondly, none of the Bangla SA research takes into account the very outstanding practical aspect (6). In this thesis, we are going to extract the sentiment conveyed by one bangla newspaper and will finding the polarity of the data  as positive or negative.Through sentiment analysis, we are going to classify given news data into one ofthree categories – positive, negative or neutral.



'আয়নাবাজি' দেখে এলাম; যেমন ভেবেছিলাম, তার চেয়েও
হাজার গুন ভালো লেগেছে :) ছবির এই দৃশ্য অসাধারণ ছিল

# Chapter 2

# Related Work

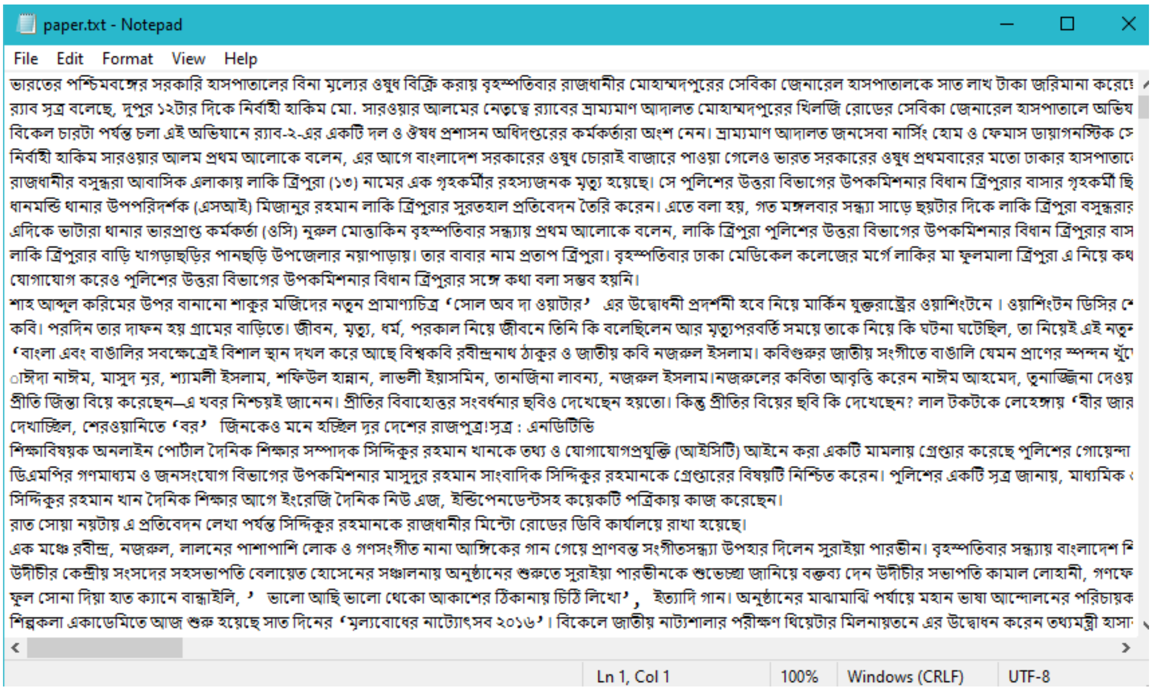
## 2.1 Related Work

There are some related works from which we based the concept of this topic. The person comment overview datasetwas used to enhance rating predictions. Their annotations blanketed six aspect categories and common sentence polarities. They had not organized a complete bangle dataset, as the aspect category used to be existing but the corresponding polarity of that aspect was absent. They published their dataset with four fields being contained for each review, that is, with the aspect time period happening in the sentences, the aspect term's polarity, the aspect category, and the aspect category's polarity. They additionally provided a news-review dataset and manually annotated with similar entities as for the bangla dataset.

In this chapter, we are introduced with several sentiment analysis of Bangla newspaper Data Set . Social networking sites have come to be a platform for online learning, sharing opinion, thinking, ideas where humans can share and express their views about affairs and making discussion with various communities across the world.

### 2.1.1 Dataset Details

According to this thesis paper,This is the simplest sentiment analysis techniques where the words of a posting newspaper data to a unlabeled word list. Though standard datasets are obtainable for few rich languages, most of the low resource languages conduct manual dataset either manufactured by the adept system or on all fours diverse online sources by custom crawlers.Raw data contain irrelevant noise like emoticons, stop words, digit, hashtags, URLs must be dispelled. Thisresearchpaper, We collect parsed newspaper data from specific website(prothomalo.com). Dataset contains a big amount of text datawhich is not labeled.



ভারতের পশ্চিমবঙ্গের সরকারি হাসপাতালের বিনা মূল্যের ওষুধ বিক্রি করায় বৃহস্পতিবার রাজধানীর মোহাম্মদপুরের সেবিকা জেনারেল হাসপাতালকে সাত লাখ টাকা জরিমানা করে

র‍্যাব সূত্র বলেছে, দুপুর ১২টার দিকে নির্বাহী হাকিম মো. সারওয়ার আলমের নেতৃত্বে র‍্যাবের ভ্রাম্যমাণ আদালত মোহাম্মদপুরের খিলজি রোডের সেবিকা জেনারেল হাসপাতালে অভিয

বিকেল চারটা পর্যন্ত চলা এই অভিযানে র‍্যাব-২-এর একটি দল ও ঔষধ প্রশাসন অধিদপ্তরের কর্মকর্তারা অংশ নেন। ভ্রাম্যমাণ আদালত জনসেবা নার্সিং হোম ও ফেমাস ডায়াগনস্টিক সে

নির্বাহী হাকিম সারওয়ার আলম প্রথম আলোকে বলেন, এর আগে বাংলাদেশ সরকারের ওষুধ চোরাই বাজারে পাওয়া গেলেও ভারত সরকারের ওষুধ প্রথমবারের মতো ঢাকার হাসপাতা

রাজধানীর বসুন্ধরা আবাসিক এলাকায় লাকি ত্রিপুরা (১৩) নামের এক গৃহকর্মীর রহস্যজনক মৃত্যু হয়েছে। সে পুলিশের উত্তরা বিভাগের উপকমিশনার বিধান ত্রিপুরার বাসার গৃহকর্মী ছি

ধানমন্ডি থানার উপপরিদর্শক (এসআই) মিজানুর রহমান লাকি ত্রিপুরার সুরতহাল প্রতিবেদন তৈরি করেন। এতে বলা হয়, গত মঙ্গলবার সন্ধ্যা সাড়ে ছয়টার দিকে লাকি ত্রিপুরা বসুন্ধরার

এদিকে ভাটারা থানার ভারপ্রাপ্ত কর্মকর্তা (ওসি) নুরুল মোত্তাকিন বৃহস্পতিবার সন্ধ্যায় প্রথম আলোকে বলেন, লাকি ত্রিপুরা পুলিশের উত্তরা বিভাগের উপকমিশনার বিধান ত্রিপুরার বাস

লাকি ত্রিপুরার বাড়ি খাগড়াছড়ির পানছড়ি উপজেলার নয়াপাড়ায়। তার বাবার নাম প্রতাপ ত্রিপুরা। বৃহস্পতিবার ঢাকা মেডিকেল কলেজের মর্গে লাকির মা ফুলমালা ত্রিপুরা এ নিয়ে কথ

যোগাযোগ করেও পুলিশের উত্তরা বিভাগের উপকমিশনার বিধান ত্রিপুরার সঙ্গে কথা বলা সম্ভব হয়নি।

শাহ আব্দুল করিমের উপর বানানো শাকুর মজিদের নতুন প্রামাণ্যচিত্র ‘সোল অব দা ওয়াটার’   এর উদ্বোধনী প্রদর্শনী হবে নিয়ে মার্কিন যুক্তরাষ্ট্রের ওয়াশিংটনে । ওয়াশিংটন ডিসির শে

কবি। পরদিন তার দাফন হয় গ্রামের বাড়িতে। জীবন, মৃত্যু, ধর্ম, পরকাল নিয়ে জীবনে তিনি কি বলেছিলেন আর মৃত্যুপরবর্তি সময়ে তাকে নিয়ে কি ঘটনা ঘটেছিল, তা নিয়েই এই নতু

‘বাংলা এবং বাঙালির সবক্ষেত্রেই বিশাল স্থান দখল করে আছে বিশ্বকবি রবীন্দ্রনাথ ঠাকুর ও জাতীয় কবি নজরুল ইসলাম। কবিগুরুর জাতীয় সংগীতে বাঙালি যেমন প্রাণের স্পন্দন খুঁ

◌াঈদা নাঈম, মাসুদ নূর, শ্যামলী ইসলাম, শফিউল হান্নান, লাভলী ইয়াসমিন, তানজিনা লাবন্য, নজরুল ইসলাম।নজরুলের কবিতা আবৃত্তি করেন নাঈম আহমেদ, তুনাজ্জিনা দেওয়

প্রীতি জিন্তা বিয়ে করেছেন—এ খবর নিশ্চয়ই জানেন। প্রীতির বিবাহোত্তর সংবর্ধনার ছবিও দেখেছেন হয়তো। কিন্তু প্রীতির বিয়ের ছবি কি দেখেছেন? লাল টকটকে লেহেঙ্গায় ‘বীর জার

দেখাচ্ছিল, শেরওয়ানিতে ‘বর’   জিনকেও মনে হচ্ছিল দূর দেশের রাজপুত্র!সূত্র : এনডিটিভি

শিক্ষাবিষয়ক অনলাইন পোর্টাল দৈনিক শিক্ষার সম্পাদক সিদ্দিকুর রহমান খানকে তথ্য ও যোগাযোগপ্রযুক্তি (আইসিটি) আইনে করা একটি মামলায় গ্রেপ্তার করেছে পুলিশের গোয়েন্দা

ডিএমপির গণমাধ্যম ও জনসংযোগ বিভাগের উপকমিশনার মাসুদুর রহমান সাংবাদিক সিদ্দিকুর রহমানকে গ্রেপ্তারের বিষয়টি নিশ্চিত করেন। পুলিশের একটি সূত্র জানায়, মাধ্যমিক

সিদ্দিকুর রহমান খান দৈনিক শিক্ষার আগে ইংরেজি দৈনিক নিউ এজ, ইন্ডিপেনডেন্টসহ কয়েকটি পত্রিকায় কাজ করেছেন।

রাত সোয়া নয়টায় এ প্রতিবেদন লেখা পর্যন্ত সিদ্দিকুর রহমানকে রাজধানীর মিন্টো রোডের ডিবি কার্যালয়ে রাখা হয়েছে।

এক মঞ্চে রবীন্দ্র, নজরুল, লালনের পাশাপাশি লোক ও গণসংগীত নানা আঙ্গিকের গান গেয়ে প্রাণবন্ত সংগীতসন্ধ্যা উপহার দিলেন সুরাইয়া পারভীন। বৃহস্পতিবার সন্ধ্যায় বাংলাদেশ শি

উদীচীর কেন্দ্রীয় সংসদের সহসভাপতি বেলায়েত হোসেনের সঞ্চালনায় অনুষ্ঠানের শুরুতে সুরাইয়া পারভীনকে শুভেচ্ছা জানিয়ে বক্তব্য দেন উদীচীর সভাপতি কামাল লোহানী, গণফে

ফুল সোনা দিয়া হাত ক্যানে বান্ধাইলি, ’   ভালো আছি ভালো থেকো আকাশের ঠিকানায় চিঠি লিখো’ ,   ইত্যাদি গান। অনুষ্ঠানের মাঝামাঝি পর্যায়ে মহান ভাষা আন্দোলনের পরিচায়ক

শিল্পকলা একাডেমিতে আজ শুরু হয়েছে সাত দিনের ‘মূল্যবোধের নাট্যোৎসব ২০১৬’। বিকেলে জাতীয় নাট্যশালার পরীক্ষণ থিয়েটার মিলনায়তনে এর উদ্বোধন করেন তথ্যমন্ত্রী হাসা

### 2.1.2 Lexicon-based and Learning-based Methods

In this paper, a new entity level sentiment analysis method is proposed for bangla data set where sentiment analysis on entities these are products, organizations and people. They claimed that their method should provide high precision and low recall. The proposed approach is combining lexicon-based and learning based method (7).

The lexicon-based method for opinion mining depends on opinion (or sentiment) words, which are words that categorical positive or negative sentiments. Words that encode a applicable state (e.g., "great" and "good") have a positive polarity, while words that encode an undesirable state have a negative polarity (e.g., "bad" and "awful"). Although opinion polarity usually applies to adjectives and adverbs, there are verb and noun opinion words as well.

By using lexicon-based method in any report or any sentence, the sentence polarity can be decided whether it is positive or negative through some feature of opinion word which is known as opinion lexicon. The method of using opinion words (the lexicon) to determine opinion orientations is called the lexicon-based approach to sentiment analysis. This method is efficient and can be employed to analyze textual content at the document, sentence or entity level. Opinion lexicon is a dictionary which consists of opinion word to determine and perceive the sentimental prediction such as positive , negative and neutral.

In their proposed technique, at first they are crawling bangle text data from social media. Then preprocessing text data through cleaning noise data such as ”RT”, external links, user names before sentiment analysis. Then realize the type of sentence these are Declarative, Imperative or Interrogative sentence.The semantic orientation in any sentence or report give an explanation for that the infer of opinion or sentence is either positive, negative or neutral.

Alternatively, we can follow a machine learning-based method to operate sentiment analysis . That is, we train a sentiment classifier to decide positive, negative and neutral sentiments. The method has been usually used for sentiment classification of documents or sentences. However, it is not handy to follow in our case because manual labeling of a massive set of tweet examples is labor intensive and time consuming. Moreover, manual labeling needs to be executed for every application domain, as it is accepted that a sentiment classifier can also operate very properly in the domain that it is trained, but performs poorly when it is utilized to a different domain . The learning based method is consequently not very scalable for sentiment analysis which covers nearly all domains as humans can specific opinions about something on text data.

### 2.1.3 Machine Learning Approach

The machine learning-based approach typically trains sentiment classifiers the use of facets such as unigrams or bigrams. Most techniques use some structure of unsupervised learning by applying different learning methods such as Hierarchical clustering, K-means clustering, K-NN (k nearest neighbors), Maximum Entropy and. These methods want manual labeling of training examples for each application domain (8).

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of enter facts except labeled responses. The most common unsupervised learning approach is cluster analysis, which is used for exploratory facts analysis to find hidden patterns or grouping in data.

Unsupervised learning is a type of machine learning algorithm used to draw conjecture from datasets component of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for experimental data analysis to search hidden patterns or grouping in data. Unsupervised machine learning search all kind of unknown patterns in data.Unsupervised methods help you to search features which can be useful for categorization (9).It is held taken place in real time, so all the input data to be analyzed and labeled in the appearance of learners. It is simply to get unlabeled data from a computer than labeled data, which needs manual intervention.Clustering is a significant concept when it comes to unsupervised learning (10). It mainly find out a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters if they remain in the data. You can also modify how much clusters your algorithms should identify. It allows you to adjust the granularity of these clusters. In machine learning technique, two types of datasets are needed:

1.Trainingset

2.Testing set

Machine learning (ML) is the learn about of computer algorithms that improve automatically through experience. It is considered as a subset of artificial intelligence. Machine learning algorithms construct a mathematical model based on sample data, recognized as "training data", in order to make predictions or decisions except being explicitly programmed to do so. Machine learning algorithms are used in a large variety of applications, such as email filtering and computer vision, where it is tough or infeasible to improve conventional algorithms to perform the needed tasks (11).

Machine learning is intently related to computational statistics, which focuses on making predictions the use of computers. The learn about of mathematical optimization can provide methods, theory and application domains to the field of machine learning (12). Data mining is a related field of study, focusing on investigative data analysis through unsupervised learning. In its application throughout commercial enterprise problems, machine learning is additionally referred to as predictive analytics.

In this case, they can use an unsupervised clustering algorithm such as k-means or hierarchical clustering to search those strong and weak customer bases.

### 2.1.4 K-Nearest Neighbours (KNN)

KNN is a statistical model reorganization algorithm which has been studied considerably for text categorization functions . It is a method for classifying objects build on closest training examples in the feature space (12). The summery of the algorithm is as follows: given a check paper x, search the K nearest neighbours of x amongst all the training documents, and rating the category candidates based totally the category of K neighbours. The similarity of x and every neighbour paper is the rating of the class of the neighbour document. If a number of of the K nearest neighbour archives belong to the identical category, then the sum of the rating of that category is the similarity score of the category in regard to the check paper x. By sorting the ratings of the candidate categories, method assigns the candidate class with the best possible score to the check paper x. The output build on whether k-NN is used for classification or regression. In k-NN classification, the output is a category member.The thing is classified by means of a plurality opinion of its neighbors, with the thing being assigned to the category most frequent amongst its k nearest neighbors (k is a positive integer, usually small). If k = 1, then the thing is actually assigned to the category of that single nearest neighbor. In k-NN regression, the output is the property render for the thing . This value is the average of the values of ok nearest neighbors.

# Chapter 3

# Proposed Work

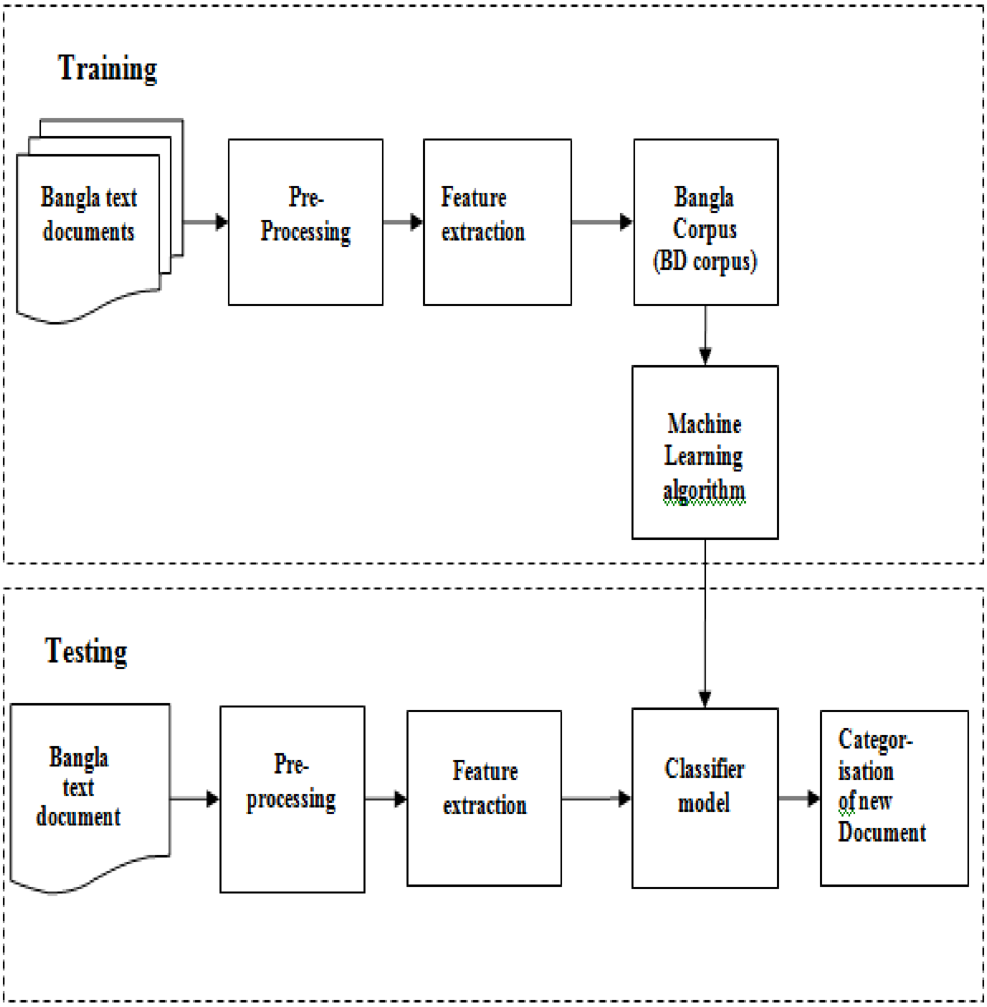


Fig-1: Bangla text classification process

## 3.1 Natural Language Processing

Natural Language Processing, generally shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language (13). The ultimate objective of NLP is to read, decipher, understand, and make experience of the human languages in a manner that is valuable. Most NLP methods depend on machine learning to derive which means from human languages. Natural language processing (NLP) is a area of artificial intelligence in which computers analyze, understand, and derive that means from human language in a smart and useful way (14). By using NLP, developers can organize and structure understand operating tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation (15). Bangla Text analysis can be implemented by using data mining technique with Natural language processing. To create solution with NLP entails applying algorithms to identify and extract the natural language rules such as that the unstructured language data is converted into a form that computers can understand. These tools are containing necessary tools to text process, and sentiment analysis.In this part, we introduce several definitions of general terminologies towards with different processing activities of SA research those are construct in our investigation.

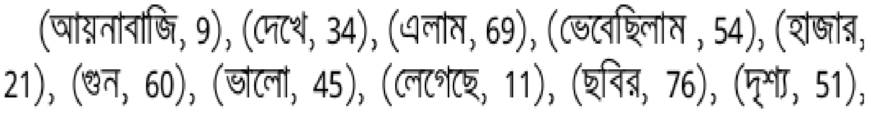
## 3.2 Googletrans

Google uses a model called Neural machine translation, where a sort of recurrent neural network based model is trained to get a sequence of tokens from another sequence of tokens. It provides a website interface, a mobile app for Android and iOS, and an application programming interface that helps developers construct browser extensions and software applications. Google Translate can translate multiple types of text and media, which consists of text, speech, and text within still or transferring images (7). Specifically, its functions include: Written Words Translation, Website Translation ,Document Translation, Speech Translation, Mobile App Translation Image Translation Handwritten Translation.You can actually do lots of things with the assistance of the Google Translate API starting from detecting languages to simple text translation, setting source and destination languages, setting source and destination languages, and translating entire lists of text data.In this article, you'll see some way to work with the Google Translate API within the Python programming language.

```
from googletrans import Translator
```

## 3.3 Vectorization

Vectorization may be a process to reconstruct a set of texts during a document into numerical feature vectors. A document is considered as a set of words with or without considering the data of their relative position (6). High frequent words having little meaning must be excluded to prioritize rarer yet more interesting terms within the corpus. so as to re- weight these count features into floating point values, Term- Frequency and Inverse Document-Frequency (TF IDF ) transformation or word embeddings are used example illustrates the vectors of words  after removing smaller words and performing tokenization operation.



(আয়নাবাজি, 9), (দেখে, 34), (এলাম, 69), (ভেবেছিলাম , 54), (হাজার,
21), (গুন, 60), (ভালো, 45), (লেগেছে, 11), (ছবির, 76), (দৃশ্য, 51),

## 3.4 Numpy

NumPy is a linear algebra library or fundamental package for scientific computing in python.Almost all of the libraries at pydata ecosystem in python rely on NumPy (15).If you have Anaconda,you can installation NumPy by using going to your command prompt by using ”pip install numpy” It contains a powerful N-dimensional array object, many broadcasting func- tions,using tools for integrating C/C++ and Fortran code.It can be used for linear algebra, Fourier transform, and random number generating.NumPy array has two component i.e Vector & Matrix where vector is 1-D and matrix is 2-D.

## 3.5 Pandas

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the Python programming language. Pandas is the most popular python library that is used for data analysis (16). It provides highly optimized performance with back-end source code is purely written in C or Python. Pandas help us to tabular data and manipulate the data and arbitrary matrix data and observational or statistical data sets.Pandas do well to manage missing data, automatic and explict records alignment, column insertion and deletion, easy to convert python and NumPy data structure into data frame object.

## 3.6 Tokenization and Normalization

In tokenization essentially refers to splitting up a large body of text into smaller lines, words or even developing words for a non-English language (6). The quite a number tokenization functions in built into the nltk module itself and can be used in programs as shown below. Then it will be simpler to a programmer to write a program that can be apprehended a single sentence then apprehend the entire sentence. The sentence damage into separate words which is recognized as ”token” & the method is referred to as Tokenization. Normalization is that the process of converting each token into unified scale for further processing (3).The method of tokenization of texts after removing stop-words punctuations, hashtags, URLs of following example to shut look the method of recognized individual words in a very system.

```
Input: রাজপণ্ডিতহবমনেআশাকরে|সপ্তশ্লোকভেটিলামরাজাগৌড়েশ্বরে
Output:'রাজপণ্ডিত','হব','মনে','আশা','করে','|','সপ্তশ্লোক','ভেটিলাম','
রাজা','গৌড়েশ্বরে'
```

## 3.7 Stemming

Stemming is Text Normalization (or sometimes called Word Normalization) techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing (16). Stemming is the manner of decreasing inflection in words to their root forms such as mapping a group of words to the identical stem even if the stem itself" is not a valid word in the Language. One of the biggest challenges in doing word lookups is to derive the suitable base word for any given word in Bengali (6). The basic concept to the solution of the problem is to remove inflections from a given word to derive its stem word. Stemmers try to reduce a word to its root form the usage of stemming process, which reduces an inflected or derived word to its stem or root form. Existing works in the literature use lookup tables either for stem words or suffixes, growing the overheads in terms of memory and time. This paper develops a rule-based algorithm that eliminates inflections stepwise besides always searching for the preferred root in the dictionary. To the satisfactory of our knowledge, this paper first investigates that, in Bengali morphology, for a massive set of inflections, the stems can be computed algorithmically slicing down the inflections step by using step.For example,

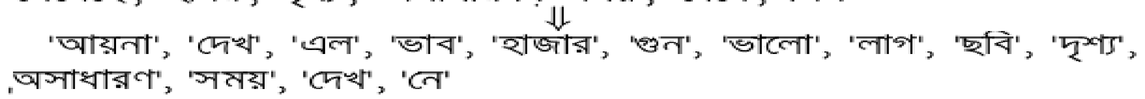
```
from bangla_stemmer.stemmer import stemmer

wordlist = [' করছে', 'কাজটি ', 'নামাবার']

stmr = stemmer.BanglaStemmer()

stm = stmr.stem(word)

print(stm)

output: ['কর', 'কাজ', 'নামা']
```

## 3.8 Lemmatization

In previous section which is talking about stemming we can see a base or root form is ”Intelligen” that doesn’t make any meaning. So,we can figure out that intermediate representation of the word may not have any meaning.When working in a computer with text,it helps to know the root form of the word which explain that the different sentence are explaining the same concept.

Lemmatization is the method of converting a word to its base form. The difference between stemming and lemmatization is, lemmatization considers the context and converts the word to its significant base form, whereas stemming just removes the last few characters, often leading to incorrect meanings and spelling errors (17). Lemmatization is a bit more complex in that the computer can group together words that do not have the same stem, but still have the equal inflected meaning. Grouping the word “good” with words like “better” and “best” is an example of lemmatization. Lemmatization is moreover a harbinger of extended artificial intelligence sophistication – as natural language processing advances in accommodating lemmatization, it is greater capable to parse inputs and provide intelligent outputs . This will be an essential factor of NLP as technology gets closer to passing the Turing test with computers that hear, understand, think and talk like humans.

For example,



'আয়নাবাজি', 'দেখে', 'এলাম', 'ভেবেছিলাম', 'হাজার', 'গুন', 'ভালো', 'লেগেছে', 'ছবির', 'দৃশ্য', 'অসাধারণ', 'সময়', 'দেখে', 'নিন'

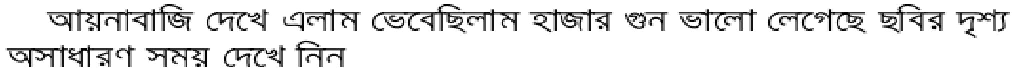
⇓

'আয়না', 'দেখ', 'এল', 'ভাব', 'হাজার', 'গুন', 'ভালো', 'লাগ', 'ছবি', 'দৃশ্য', ,অসাধারণ', 'সময়', 'দেখ', 'নে'

## 3.9 Stop Word, Hash Tags, Digit Removal

stop words are words which are filtered out  processing of natural language data  . Stop words are typically the most common words in a language; there is no single common list of stop words used by means of all natural language processing tools, and indeed not all tools even use such a listing. Stop words are usually used words that are excluded from searches to help index and parse web pages faster. Some case of stop words are: "a," "and," "but," "how," "or," and "what.". While the majority of all Internet search engines make use of stop words, they do not stop a user from the usage of them, however they are ignored.In general stop words are language specific (13).For english words like a, an, the, they're stop words. we are also add some extra words as per your requirement. Generally, they're used as segment of pre-processing eliminate extra/common words in every documents.. As they don't put any extra influence on result.

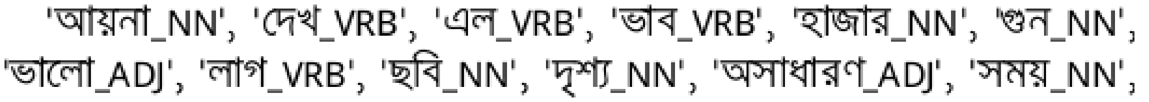
so when we are doing sentiment analysis these words have no impact on sentiment whether or not the sentiment is positive or negative. This is the purpose in most of the instances we actually want to eliminate these different stop words to get better performance.Hashtags in social networks also are marked with a number sign (#) ahead of a word or unspaced phrase during a document The shows the output of example after removing stop words, punctuations, hash tags, decibel number and URLs to know the importance of this process in SA (17). A general Bengali document may contain Bengali similarly as English digits. But as meaningful Bengali words don't contain digits, we remove these digits by using their Unicode representation.



আয়নাবাজি দেখে এলাম ভেবেছিলাম হাজার গুন ভালো লেগেছে ছবির দৃশ্য অসাধারণ সময় দেখে নিন

## 3.10 Parts Of Speech Tagger

In this pipeline,program is looking each tokenize word and gives its part of speech.Knowing the part of speech of each word in the document will help to understand what the sentence is talking about.The model of parts of speech was trained by millions of English sentence with its each word's POS so that it can learn the behaviou of that word.After Tokenization each tokenize word can show their POS by using



```
'আয়না_NN', 'দেখ_VRB', 'এল_VRB', 'ভাব_VRB', 'হাজার_NN', 'গুন_NN',
'ভালো_ADJ', 'লাগ_VRB', 'ছবি_NN', 'দৃশ্য_NN', 'অসাধারণ_ADJ', 'সময়_NN',
```

The meaning of Part of Speech in NLTK:

- CC coordinating conjunction
- CD cardinal digit
- DT determiner
- EX existential there (like: "there is" … think about it like "there exists")
- FW foreign word
- IN preposition/subordinating conjunction
- JJ adjective 'big'
- JJR adjective, comparative 'bigger'
- JJS adjective, superlative 'biggest'

- LS list marker 1)
- MD modal could, will
- NN noun, singular ‘desk’
- NNS noun plural ‘desks’
- NNP proper noun, singular ‘Harrison’
- NNPS proper noun, plural ‘Americans’
- PDT pre-determiner ‘all the kids’
- POS possessive ending parent’s
- PRP personal pronoun I, he, she
- PRP$ possessive pronoun my, his, hers
- RB adverb very, silently,
- RBR adverb, comparative better
- RBS adverb, superlative best
- RP particle give up
- TO, to go ‘to’ the store.
- UH interjection, errrrrrrrm
- VB verb, base form take
- VBD verb, past tense took
- VBG verb, gerund/present participle taking

- VBN verb, past participle taken
- VBP verb, sing. present, non-3d take
- VBZ verb, 3rd person sing. present takes
- WDT wh-determiner which
- WP wh-pronoun who, what
- WP$ possessive wh-pronoun whose
- WRB wh-abverb where, when

## 3.11 Named Entity Recognization

Named-entity Recognization (also called entity identification, entity chunking and entity extraction) could be a subtask of data extraction that seeks to detect and classify named entity noted in unstructured textual content material into pre-defined classes such as individual names, organizations, locations, scientific codes, time expressions, quantities, economic values, percentages, etc. For example, an NER machine learning (ML) model may observe the word “Canotic” in a text and classify it as a “Company”. NER model is a two step process: 1. Detect a named entity 2. Categorize the entity Step one includes detecting a word or string of words that structure an entity. Individually a word represents a token: “The Great Lakes” is a string of three tokens that represents one entity. Inside-outside-beginning tagging is a frequent way of indicating where entities start and end. We’ll discover this similarly in a future blog post (4).The 2nd step requires the introduction of entity categories. Using named entity recognition facts program, you supply us your raw text and preferred entities and categories. We’ll label the text you send and return a excessive high-quality training dataset that you can desire train and tailor your NER model.

## 3.12 Bangla Text Modeling

Natural Language Processing is the subarea of computer science and artificial intelligence which refers the normal language that we use to communicate to converse language which will understand computer such as computer code.In natural language processing,Text modeling which is also called statistical model to use for discovering the abstract from the document.Basically Text modeling is using as a text mining tool to find out semantic structure from a text body (14).We are using some text model to evaluating sentiment analysis by building a classifier. We are testing by two model to train & testing the classifier which provides us more accuracy.

- Bag of Words
- TF-IDF Model

### 3.12.1 Bag Of Words

We need some way to represent text information for machine learning algorithm and also the bag-of-words model helps us to accumulate that task. The bag-of-words model is easy to know and implement. it's some way of extracting aspects from the text to be used in machine learning algorithms. The bag-of-words model could be a simplifying illustration utilized in language processing and data retrieval (IR). during this model, a text (such as a sentence or a document) is represented because the bag (multiset) of its words, dismissing grammar and even word order however maintaining multiplicity (14). The bag-of-words model has additionally been used for computer vision. The bag-of-words model is sometimes utilized in methods of document classification where the happening of every word is used as a function for training a classifier.

In the face of being a relatively basic model, BOW is sometimes used for language Processing tasks like Text Classification. Its strengths be its simplicity: it’s inexpensive to compute, and sometimes simpler is best when positioning or contextual info aren’t relevant.

For Example:

**1. অনুগ্রহ পূর্বক ধীরে বলুন**.

**2.অনুগ্রহ পূর্বক পুনরায় বলুন**.

Now Create a Bag of words from the document of different sentence:

| Bag of Words | | | | | |
|---|---|---|---|---|---|
| Words/Documents | অনুগ্রহ | পূর্বক | ধীরে | বলুন | পুনরায় |
| sentence1 | 1 | 1 | 1 | 1 | 0 |
| sentence2 | 1 | 1 | 0 | 1 | 1 |

So, within the columns there are all the everyday words and within the rows there are all thedifferent documents and also the simplest scoring method is to mark the presence of wordsas a binary value, 0 for absent, 1 for present.Using the arbitrary ordering of words listed above in our vocabulary, we will step through the first document (“**অনুগ্রহ পূর্বক ধীরে বলুন**“) and convert it into a binary vector. All ordering of the words is nominally discarded and that we have the same way of extracting features from any document in our corpus, ready to be used in modeling.

### 3.12.2 TF-IDF Model

TF-IDF means Term Frequency-Inverse Document Frequency. There are some problems have to be find out before discussing about TF –DF in previous section BOW model was explained. For Bag of words models we'll build a model which is more usefully. Bag of Words just creates a set of vectors containing the count of word occurrences in the document , while the TF-IDF model contains information on the more important words and the less important ones as well.

Bag of Words vectors are easy to interpret. However, TF-IDF usually performs better in machine learning models.

1.It doesn't preserve any semantic information.

2.It gives all words have the same importance.

So, Term frequency-inverse document frequency weight could be a weight often utilized in information retrieval and text mining. TF-IDF will be successfully used for stop-words filtering in various subject fields including text summarization and classification. This is a statistical quantity accustomed measure the importance of a word with regard to a document corpus.

TF-IDF is determined by multiplying a local component like term frequency (TF) with a global component, that is, inverse document frequency (IDF) and the result to unit length optionally normalized.

Term frequency (TF) is used in connection with information retrieval and shows how frequently an term, word) occurs in a document. Term frequency shows the importance of a single term within the overall document. Besides, term frequency is frequently divided by the total number of terms in the document as a way of normalization. TF is individual to each document and word, hence we can formulate TF as follows.

$$\text{Term frequency,TF} = \frac{\text{Number of times term appears in a document}}{\text{Total number of terms in the document}}$$

IDF is that the inverse of the document frequency which measures the informativeness of term t When we calculate IDF, it will be very low for the most occurring words such as stop words. The inverse document frequency (IDF) could be a statistical weight used for measuring the importance of a term during a text document collection. The document frequency DF of a term is defined by the quantity of documents during which a term appears.

$$\text{Inverse Document Frequency, IDF=log } \frac{\text{Total number of documents}}{\text{Number of document with term in it}}$$

So,

$$\text{TF-IDF=TF* log } \frac{\text{Total number of documents}}{\text{Number of document with term in it}}$$

Example:Taking twosentence

**1.অনুগ্রহ পূর্বক ধীরে বলুন**.

**2.অনুগ্রহ পূর্বক পুনরায় বলুন**.

The frequency list of most frequent word:

| Frequency List | |
|---|---|
| Word | Frequency |
| অনুগ্রহ | 2 |
| পূর্বক | 2 |
| ধীরে | 1 |
| বলুন | 2 |
| পুনরায় | 1 |

Now the Term Frequency for every of the various words of frequency:

| Term Frequency | | |
|---|---|---|
| Word | Sentence1 | Sentence2 |
| অনুগ্রহ | 0.25 | 0.25 |
| পূর্বক | 0.25 | 0.25 |
| ধীরে | 0.25 | 0 |
| বলুন | 0.25 | 0.25 |
| পুনরায় | 0 | 0.25 |

According to formula of Inverse Document Frequency the IDF values are:

| Inverse Document Frequency | |
|---|---|
| Word | IDF Values |
| অনুগ্রহ | 0 |
| পূর্বক | 0 |
| ধীরে | 0.693 |
| বলুন | 0 |
| পুনরায় | 0.693 |

Now we've the values of TF & IDF model. Then calculating the final value of TF-IDF model which is given below.

| TF-IDF Values | | | | | |
|---|---|---|---|---|---|
| Words/Documents | অনুগ্রহ | পূর্বক | ধীরে | বলুন | পুনরায় |
| sentence1 | 0 | 0 | 0.173 | 0 | 0 |
| sentence2 | 0 | 0 | 0 | 0 | 0.173 |

The bag of words provide only zeros & ones. But this model carries zeroes because of the idea of values but it includes these fractional values additionally such as 0.173.So when we are doing it on a massive corpus of data we have this bunch of decimal values and if we look carefully at these specific values then we will be able to find out that most of the essential words in the entire document have a higher value and the higher fractional value of a document. So the word " **পুনরায়**" is a very essential word in our two document corpus So it has a very excessive value which is 0.173 &amp; so on. So you can see in the word " **বলুন**" is a very frequent word right because it regarded in all the documents. So it obtained the lowest TF-IDF value which is zero. So similarly in this way in the TF-IDF model can provide more importance to some precise word and that is the purpose why this model is significantly used in case of text classification are in our opinion mining and many different applications.

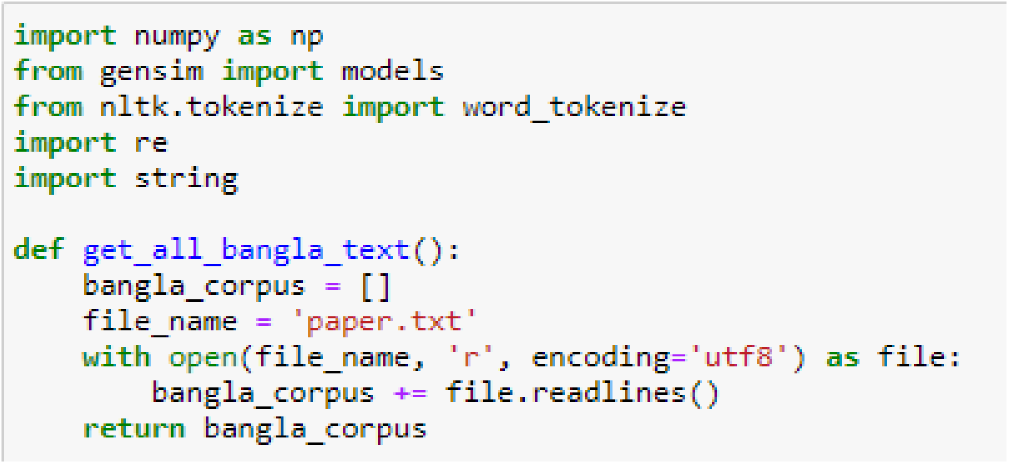
# Chapter 4

# Model  Validation

As we mentioned before, we used raw text data from newspapers. In order to process further work, we have to import data.

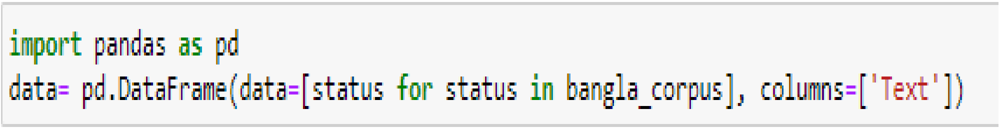
## 4.1  Creating Data Frame and Cleansing

First of all we imported all the necessary libraries. After that we import the dataset using a function as given in the next screenshot. Here, “paper.txt” is our dataset file which must be in the same folder of python code. At first we are importing several  library to access one module of python code from another module.It searches for the named module that it binds the results of search to a name into a local variable.The quite a number tokenization functions in built into the nltk module itself and can be used in programs as shown below.Then it will be simpler to a programmer to write a program that can be apprehended a single sentence then apprehend the entire sentence. The sentence damage into separate words which is recognized as ”token” & the method is referred to as Tokenization.



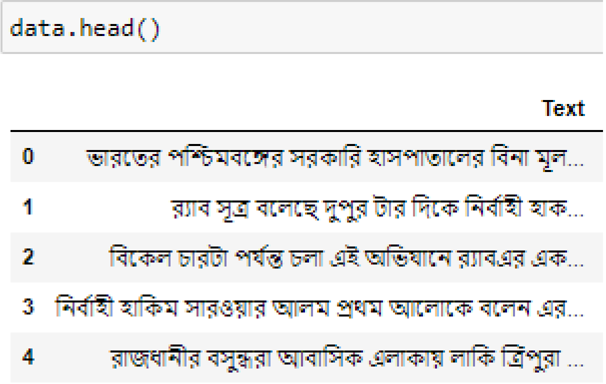
```
import numpy as np
from gensim import models
from nltk.tokenize import word_tokenize
import re
import string

def get_all_bangla_text():
    bangla_corpus = []
    file_name = 'paper.txt'
    with open(file_name, 'r', encoding='utf8') as file:
        bangla_corpus += file.readlines()
    return bangla_corpus
```

To show the storing data in dataframe, we created a pandas dataframe.“import re” for Regular expression which specifies a purchase of string matches with the given regular expression. “import pandas” uses to manage data and ”import NumPy” uses for computing number. This module contains variety of functions to process standard Python strings. In Python, many string operations are made available as string methods further, and many functions within the string module are simply wrapper functions that call the corresponding string method.



```
import pandas as pd
data= pd.DataFrame(data=[status for status in bangla_corpus], columns=['Text'])
```

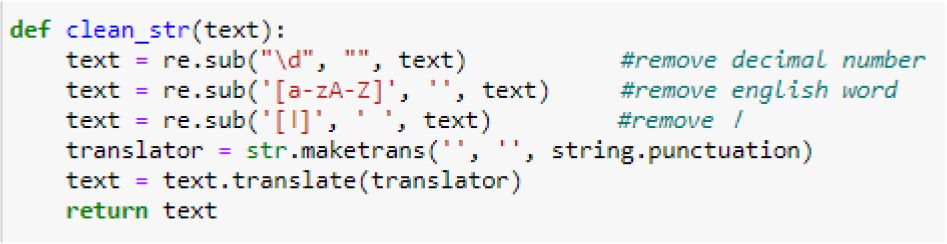
We checked the dataset if it is working properly or not. As given in picture its perfect.



```
data.head()
```

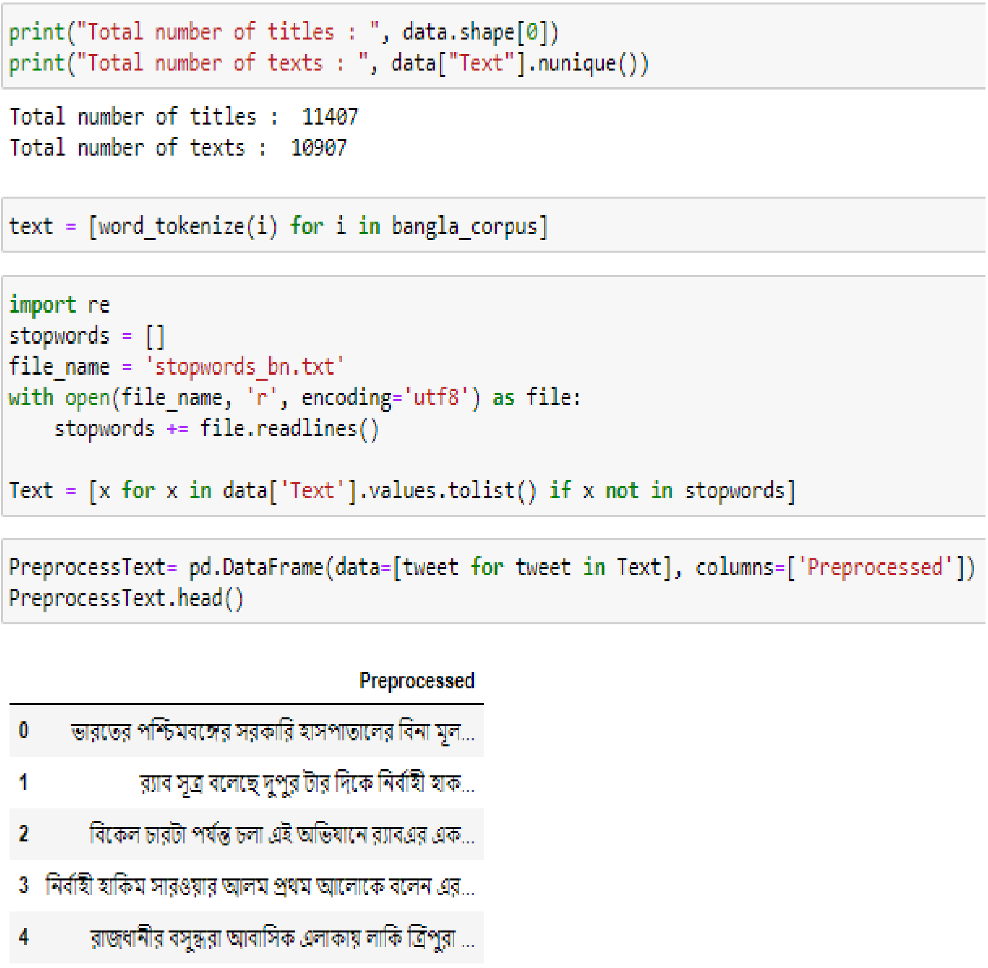
```
                                              Text
0       ভারতের পশ্চিমবঙ্গের সরকারি হাসপাতালের বিনা মূল...
1                 র‍্যাব সূত্র বলেছে দুপুর টার দিকে নির্বাহী হাক...
2            বিকেল চারটা পর্যন্ত চলা এই অভিযানে র‍্যাবএর এক...
3  নির্বাহী হাকিম সারওয়ার আলম প্রথম আলোকে বলেন এর...
4            রাজধানীর বসুন্ধরা আবাসিক এলাকায় লাকি ত্রিপুরা ...
```

As we know, raw dataset contain a huge amount of garbage data. So we have to clean it. So, we tried to remove all unnecessary content from the text. Data cleaning or cleansing is that the process of detecting and correcting (or removing) garbage or inaccurate records from a record set, table, or database and prescribe to identifying incomplete, incorrect, inaccurate or irrelevant parts of the info and so replacing, modifying, or deleting the nasty or unclean data.



```
def clean_str(text):
    text = re.sub("\d", "", text)            #remove decimal number
    text = re.sub('[a-zA-Z]', '', text)      #remove english word
    text = re.sub('[।]', ' ', text)          #remove ।
    translator = str.maketrans('', '', string.punctuation)
    text = text.translate(translator)
    return text
```

Now we have to tokenize from the bangla_corpus. After tokenization, we removed the stopwords. We need to remove stopwords because it does not make any sense in terms of sentiment. Stop words are typically the most common words in a language; there is no single common list of stop words used by means of all natural language processing tools, and indeed not all tools even use such a listing. Stop words are usually used words that are excluded from searches to help index and parse web pages faster. So, our data is now preprocessed. Code and output is given below-



```
print("Total number of titles : ", data.shape[0])
print("Total number of texts : ", data["Text"].nunique())
```

```
Total number of titles :  11407
Total number of texts :  10907
```

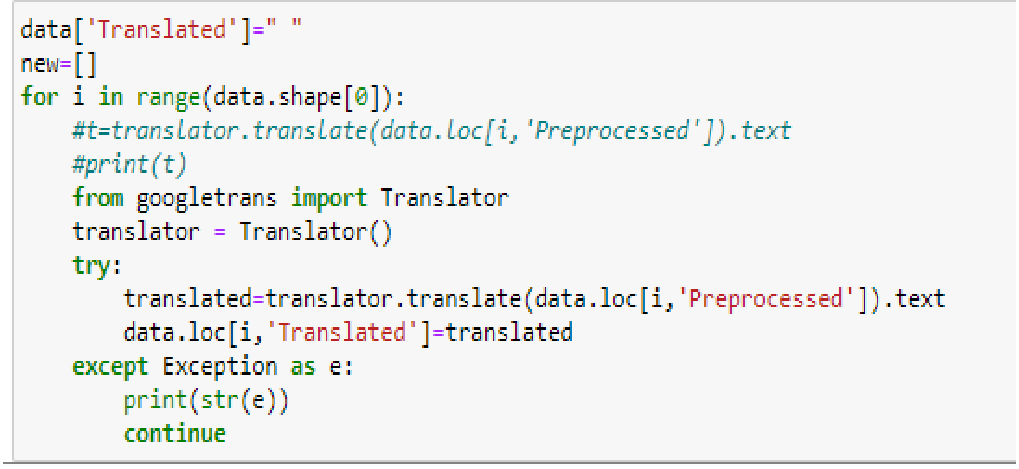
```
text = [word_tokenize(i) for i in bangla_corpus]
```

```
import re
stopwords = []
file_name = 'stopwords_bn.txt'
with open(file_name, 'r', encoding='utf8') as file:
    stopwords += file.readlines()

Text = [x for x in data['Text'].values.tolist() if x not in stopwords]
```

```
PreprocessText= pd.DataFrame(data=[tweet for tweet in Text], columns=['Preprocessed'])
PreprocessText.head()
```

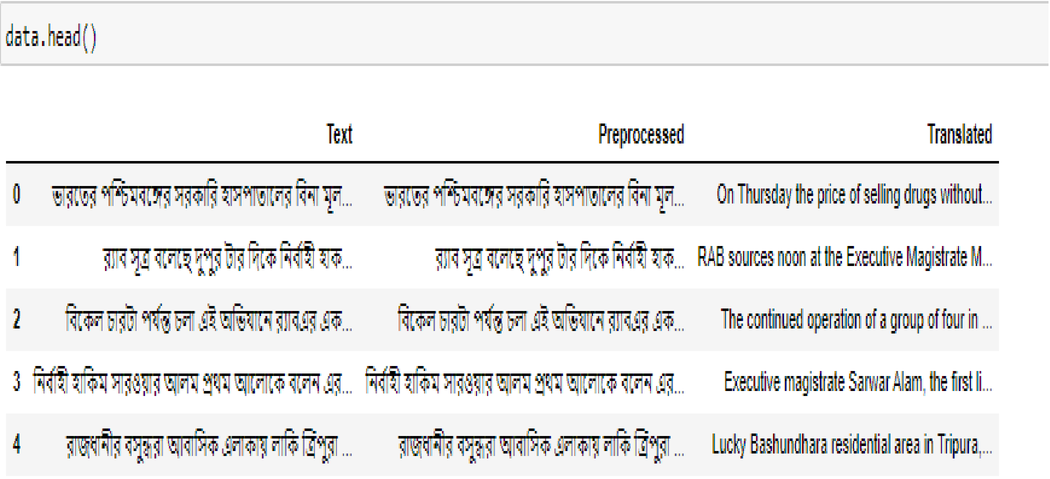
| | Preprocessed |
|---|---|
| 0 | ভারতের পশ্চিমবঙ্গের সরকারি হাসপাতালের বিনা মূল... |
| 1 | র‍্যাব সূত্র বলেছে দুপুর টার দিকে নির্বাহী হাক... |
| 2 | বিকেল চারটা পর্যন্ত চলা এই অভিযানে র‍্যাবএর এক... |
| 3 | নির্বাহী হাকিম সারওয়ার আলম প্রথম আলোকে বলেন এর... |
| 4 | রাজধানীর বসুন্ধরা আবাসিক এলাকায় লাকি ত্রিপুরা ... |

We want to analyze sentiment from bangla text. But it is not possible in Textblob libray. At the beginning of sentiment analysis, we worked with a construct in Python library which is ”Textblob”.But when we able to analysis the same data by using Natural Language Processing then we have found several problem and also found some wrong prediction from ”Textblob”. We translated the preprocessed text in order to analyze. Google Translate can translate multiple types of text and media, which consists of text, speech, and text within still or transferring images. Specifically, its functions include: Written Words Translation, Website Translation ,Document Translation, Speech Translation, Mobile App Translation Image ,Translation Handwritten Translation.



```
data['Translated']=" "
new=[]
for i in range(data.shape[0]):
    #t=translator.translate(data.loc[i,'Preprocessed']).text
    #print(t)
    from googletrans import Translator
    translator = Translator()
    try:
        translated=translator.translate(data.loc[i,'Preprocessed']).text
        data.loc[i,'Translated']=translated
    except Exception as e:
        print(str(e))
        continue
```

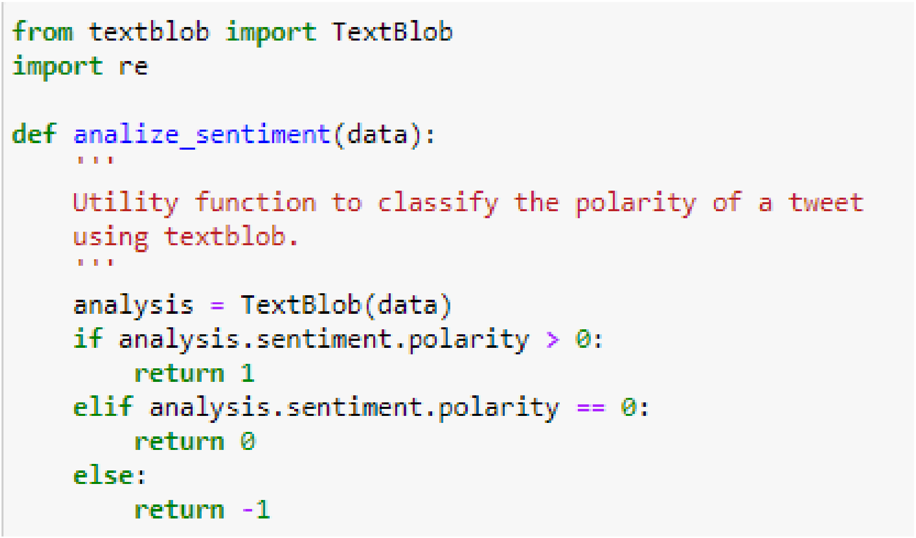
After preprocessing and all essential formation our dataset is ready for work. Our final dataframe looks like below picture. We have actually do lots of things with the assistance of the Google Translate API starting from detecting languages to simple text translation, setting source and destination languages, setting source and destination languages, and translating entire lists of text data. In this article, you'll see some way to work with the Google Translate API within the Python programming language.



```
data.head()
```

|   | Text | Preprocessed | Translated |
|---|---|---|---|
| 0 | ভারতের পশ্চিমবঙ্গের সরকারি হাসপাতালের বিনা মূল... | ভারতের পশ্চিমবঙ্গের সরকারি হাসপাতালের বিনা মূল... | On Thursday the price of selling drugs without... |
| 1 | র‍্যাব সূত্র বলেছে দুপুর টার দিকে নির্বাহী হাক... | র‍্যাব সূত্র বলেছে দুপুর টার দিকে নির্বাহী হাক... | RAB sources noon at the Executive Magistrate M... |
| 2 | বিকেল চারটা পর্যন্ত চলা এই অভিযানে র‍্যাবএর এক... | বিকেল চারটা পর্যন্ত চলা এই অভিযানে র‍্যাবএর এক... | The continued operation of a group of four in ... |
| 3 | নির্বাহী হাকিম সারওয়ার আলম প্রথম আলোকে বলেন এর... | নির্বাহী হাকিম সারওয়ার আলম প্রথম আলোকে বলেন এর... | Executive magistrate Sarwar Alam, the first li... |
| 4 | রাজধানীর বসুন্ধরা আবাসিক এলাকায় লাকি ত্রিপুরা ... | রাজধানীর বসুন্ধরা আবাসিক এলাকায় লাকি ত্রিপুরা ... | Lucky Bashundhara residential area in Tripura,... |

The Textblob package for python is a convenient way to do Natural Language Processing tasks. The lexicon it refers to is in en-sentiment.xml. In the time of calculating a single word, it uses a sophisticated technique known to mathematicians as “averaging”. We used Textblob as given bleow-



```
from textblob import TextBlob
import re

def analize_sentiment(data):
    '''
    Utility function to classify the polarity of a tweet
    using textblob.
    '''
    analysis = TextBlob(data)
    if analysis.sentiment.polarity > 0:
        return 1
    elif analysis.sentiment.polarity == 0:
        return 0
    else:
        return -1
```

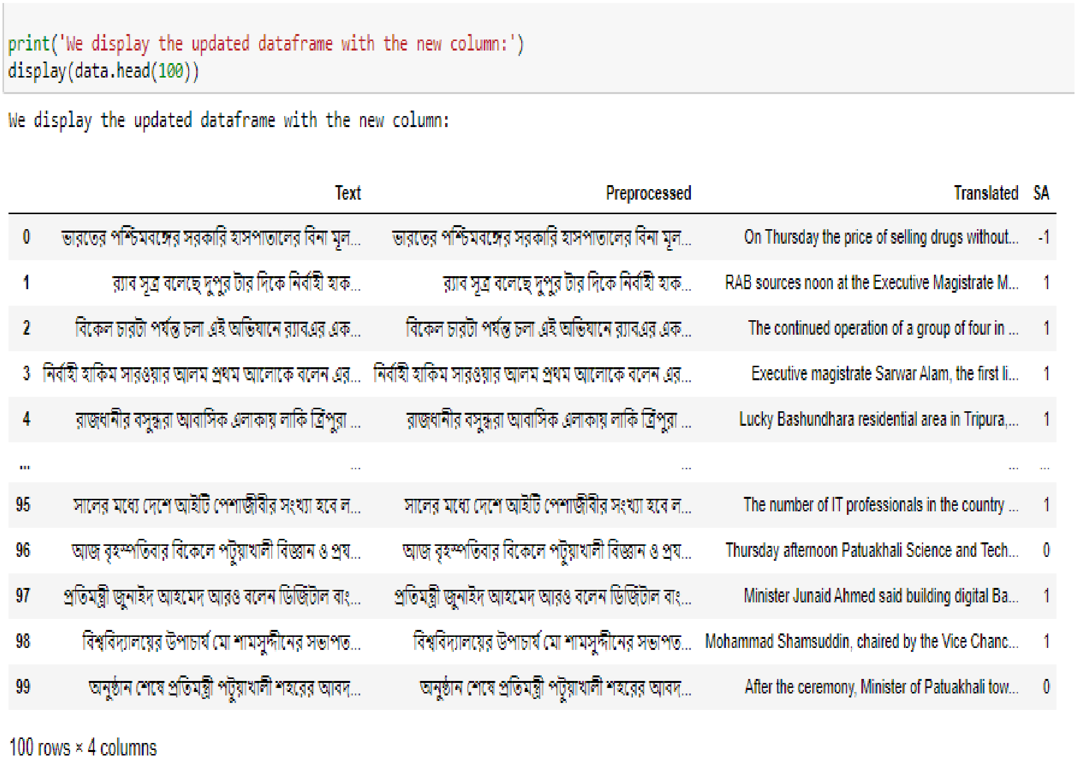
The ”Textblob” library predicts as negative sentiment, positive sentiment or neutral against those data but . we can use analize_sentiment function and this function will provide numeric value 1 (positive sentiment) or  0 (negative sentiment)or -1(neutral sentiment).

# Chapter 5

# Performance Evaluation  and Result

## 5.1Result

After cleansing dataset, we used Textblob library package for our sentiment analysis. The result of our work so far is given below which denoted sentiment at very most right column-



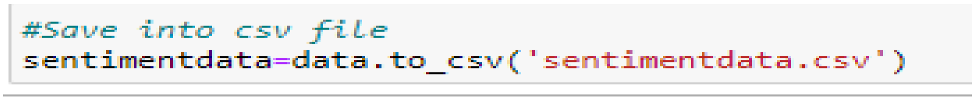
```
print('We display the updated dataframe with the new column:')
display(data.head(100))
```

```
We display the updated dataframe with the new column:
```

|  | Text | Preprocessed | Translated | SA |
|---|---|---|---|---|
| 0 | ভারতের পশ্চিমবঙ্গের সরকারি হাসপাতালের বিনা মূল... | ভারতের পশ্চিমবঙ্গের সরকারি হাসপাতালের বিনা মূল... | On Thursday the price of selling drugs without... | -1 |
| 1 | র‍্যাব সূত্র বলেছে দুপুর টার দিকে নির্বাহী হাক... | র‍্যাব সূত্র বলেছে দুপুর টার দিকে নির্বাহী হাক... | RAB sources noon at the Executive Magistrate M... | 1 |
| 2 | বিকেল চারটা পর্যন্ত চলা এই অভিযানে র‍্যাবএর এক... | বিকেল চারটা পর্যন্ত চলা এই অভিযানে র‍্যাবএর এক... | The continued operation of a group of four in ... | 1 |
| 3 | নির্বাহী হাকিম সারওয়ার আলম প্রথম আলোকে বলেন এর... | নির্বাহী হাকিম সারওয়ার আলম প্রথম আলোকে বলেন এর... | Executive magistrate Sarwar Alam, the first li... | 1 |
| 4 | রাজধানীর বসুন্ধরা আবাসিক এলাকায় লাকি ত্রিপুরা ... | রাজধানীর বসুন্ধরা আবাসিক এলাকায় লাকি ত্রিপুরা ... | Lucky Bashundhara residential area in Tripura,... | 1 |
| ... | ... | ... | ... | ... |
| 95 | সালের মধ্যে দেশে আইটি পেশাজীবীর সংখ্যা হবে ল... | সালের মধ্যে দেশে আইটি পেশাজীবীর সংখ্যা হবে ল... | The number of IT professionals in the country ... | 1 |
| 96 | আজ বৃহস্পতিবার বিকেলে পটুয়াখালী বিজ্ঞান ও প্রয... | আজ বৃহস্পতিবার বিকেলে পটুয়াখালী বিজ্ঞান ও প্রয... | Thursday afternoon Patuakhali Science and Tech... | 0 |
| 97 | প্রতিমন্ত্রী জুনাইদ আহমেদ আরও বলেন ডিজিটাল বাং... | প্রতিমন্ত্রী জুনাইদ আহমেদ আরও বলেন ডিজিটাল বাং... | Minister Junaid Ahmed said building digital Ba... | 1 |
| 98 | বিশ্ববিদ্যালয়ের উপাচার্য মো শামসুদ্দীনের সভাপত... | বিশ্ববিদ্যালয়ের উপাচার্য মো শামসুদ্দীনের সভাপত... | Mohammad Shamsuddin, chaired by the Vice Chanc... | 1 |
| 99 | অনুষ্ঠান শেষে প্রতিমন্ত্রী পটুয়াখালী শহরের আবদ... | অনুষ্ঠান শেষে প্রতিমন্ত্রী পটুয়াখালী শহরের আবদ... | After the ceremony, Minister of Patuakhali tow... | 0 |

100 rows × 4 columns

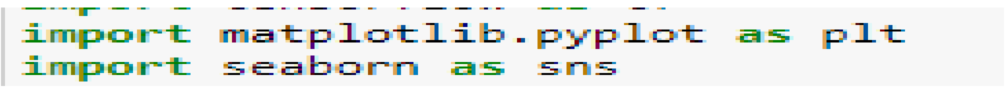
Here, 1 is for positive, -1 for negative and 0 for neutral sentiment. The Chart illustrates overall results or outcome in three different plotted chart for bangla newspaper. Initially bar, it representing the opinion of every text data in negative & positive& neutral manner. Then secondly, this chart exhibit positive and negative and neutral  opinion in out of 100 scales. It displaying aggregate a part of many people opinion for csv sentiment data file.



```
#Save into csv file
sentimentdata=data.to_csv('sentimentdata.csv')
```

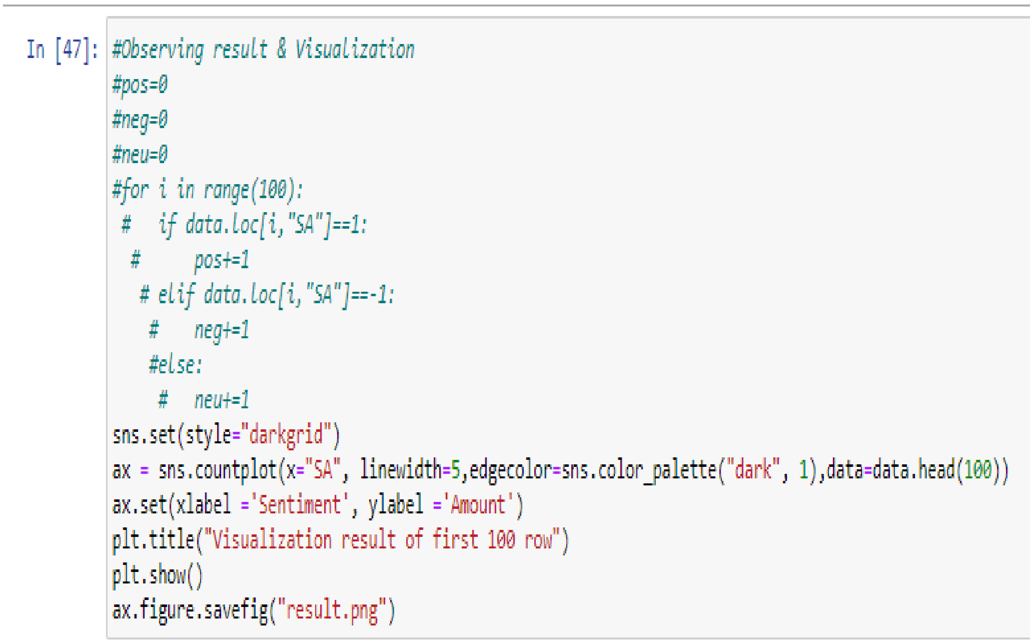
## 5.2  Performance Evaluation

The performance of SA system is evaluated by individual metrics including expert analysis, precision, recall, accuracy to justify the acceptance  of the system under considerations. Then, remarks on limitations and future directions are given by discussing the performance of a system. After the scenes, seaborn uses matplotlib to draw plots. Many tasks may be accomplished with only seaborn functions, but further customization might require using matplotlib directly. This can be explained in additional detail below. For interactive work, it's recommended to use a Python interface in matplotlib mode, alternatively you'll must call matplotlib.pyplot.show() after you want to determine the plot. We import seaborn, which is the only library necessary for this following example-



```
import matplotlib.pyplot as plt
import seaborn as sns
```

We can see the follwing observing result and visualization. The Seaborn python library is well known for its grey background and its general styling. However, note that a few other built in style are available darkgrid.



```
#Observing result & Visualization
#pos=0
#neg=0
#neu=0
#for i in range(100):
 #   if data.loc[i,"SA"]==1:
  #      pos+=1
   # elif data.loc[i,"SA"]==-1:
    #    neg+=1
    #else:
     #   neu+=1
sns.set(style="darkgrid")
ax = sns.countplot(x="SA", linewidth=5,edgecolor=sns.color_palette("dark", 1),data=data.head(100))
ax.set(xlabel ='Sentiment', ylabel ='Amount')
plt.title("Visualization result of first 100 row")
plt.show()
ax.figure.savefig("result.png")
```

Here we can see bangle newspaper with its opinion goes almost 63%with positive sentiment. ButFew people was neutral sentiment analysis with its opinionalmost 16%.Some people was negative sentiment analysis with its opinion almost 21% Following figure is showing the percentage bar chart of the mining opinion  It helps us to understand people reaction more clearly.

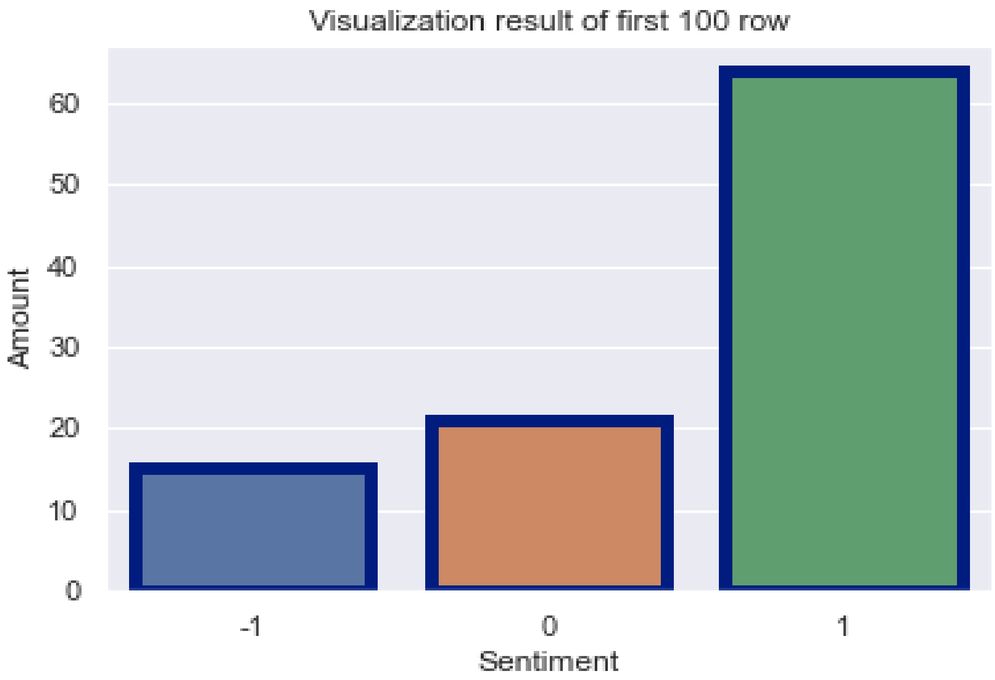


Fig-2: Bar chart of mining opinion for bangla newspaper

# Chapter 6

# Conclusion

## 6.1 Conclusion

Many of the research work on sentiment analysis has been dispensed within the English, but work in Bangla is restricted to only news corpus,websites and blogs that means social networking sites. Websites are becoming a valuable source for declaring huge size of user-generated information, as users reveal their views, opinions, and sentiments over various object.By this paper, we motive to automatically quiddity the emotions or opinions conveyed by users from Bangla blog posts then identify the general polarity of texts as either negative or positive (16).We built the model with several months news from a Bengali daily newspaper Prothom-alo.An opinion could also be positive, negative or neutral depends on individuals judgment or evaluation towards a subject or an event.

Usually sentiments may be varied in cultures and languages.This reconsideration substance demonstrates existing procedures, achievements, and limitations of obtainable works of Sentiment Analysis (SA) on Bengali texts. In our determination, we represent pre-processing steps with respect to the domains, context of research, parallelism of various experimental methods also as performance evaluation of a system which is able to help the new researchers to try and do their research during this field (13). In our judgment, a little number of evaluation metrics are satisfactory regarding problem domains on Bengali texts; however, there are a sizable amount of domains available that don't seem to be explored till now and hence it are often an exciting field to contribute. Moreover, Bangla Natural Language Processing (BNLP ) tools, standardized datasets for benchmarking don't seem to be up up to now for Bengali language to develop SA system effectively which must be kept in mind during the longer term experiments (14).

## 6.2 Future Work

The main future approach of our research is to test the influence of other text representation schemes on text classification, such as several other term weighting schemes , latent semantic Indexing, and multi-word for text depiction. For the sake of ability, it's also worth inquiring further popular dimensional lightening techniques including cosmopolitan term stemming and pruning. We may work with labeled dataset which is supervised learning method. A great update will be, if we use real time data for analysis and we are going to develop this kind of work very soon. Using different algorithm, clean dataset, emphasizing on update learning model will accumulate weight to the project.

# Chapter 7

# Reference

1. *Sentiment analysis on bangla and romanized bangla text using deep recurrent models.* **Hassan, Asif, et al.** Dhaka, Bangladesh : IEEE, 2016. International Workshop on Computational Intelligence (IWCI).

2. **Md. Atikur Rahman, Emon Kumar Dey.** Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation. Dhaka 1000, Bangladesh : s.n., 2018.

3. *On Predicting and Analyzing Breast Cancer using Data Mining Approach.* **M. R Basunia, I. A. Pervin, M. A. Mahmud, S. Saha and M. Arifuzzaman.** s.l. : IEEE, 2020. Region 10 Symposium (TENSYMP).

4. **Mohammad Samman Hossain, Israt Jahan Jui,Afia Zahin Suzana.** *Sentiment Analysis for Bengali Newspaper Headlines.* s.l. : BRAC University.

5. *Performing sentiment analysis in Bangla microblog posts.* **Chowdhury, Shaika and Chowdhury, Wasifa.** Dhaka, Bangladesh : IEEE, 2014.

6. *Survey on Text-Based Sentiment Analysis of Bengali Language.* **Nayan Banik, Md. Hasan Hafizur Rahman,Shima Chakraborty,Hanif Seddiqui, Muhammad Anwarul Azim.** Comilla - 3506, Bangladesh : ICASERT, 2019.

7. *A Comparative Sentiment Analysis On Bengali Facebook post.* **Salehin, S.M. Samiul.** Dhaka, Bangladesh : ICCA, 2020.

8. *Analyzing Sentiment of Movie Reviews in Bangla by Applying Machine Learning Techniques.* **Rumman Rashid Chowdhury, Mohammad Shahadat Hossain,Sazzad Hossain,and Karl Andersson.** Chittagong, Bangladesh : International Conference on Bangla Speech and Language Processing(ICBSLP), 2019.

9. *Finding emotion holder from Bengali blog texts -An unsupervised syntactic approach.* **S., Das D. and Bandyopadhyay.** s.l. : PACLIC 24 - Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, 2010.

10. *Unsupervised sentiment analysis with emotional signals.* **X. Hu, J. Tang, H. Gao, and**

**H. Liu.** s.l. : the 22nd international conference on World Wide Web, 2013. 607-618.

11. *SUPERVISED LEARNING METHODS FOR BANGLA WEB DOCUMENT CATEGORIZATION.* **Sen, Ashis Kumar Mandal and Rikta.** 5, Pahang, Malaysia : International Journal of Artificial Intelligence & Applications, 2014, Vol. 5.

12. *KNN based machine learning approach for text and document mining.* **Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, Jordan Pascual.** 1, s.l. : International Journal of Database Theory and Application, 2014, Vol. 7. 61-70.

13. *Automatic detection of opinion bearing words and sentences.* **Soo-Min Kim, Eduard Hovy.** s.l. : the International Joint Conference on Natural Language Processing (IJCNLP), 2005.

14. *Exploring Word Embedding for Bangla Sentiment Analysis.* **Sumit, Sakhawat Hosain, et al.** Sylhet, Bangladesh : IEEE, 2018.

15. *Sentiment Analysis with NLP on Twitter Data.* **Hasan, Md Rakibul, Maisha Maliha, and M. Arifuzzaman.** s.l. : IEEE, 2019. In 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). pp. 1-4.

16. *SMOTE Approach for predicting the Success of Bank Telemarketing.* **Islam, Md Shifatul,Mohammad Arifuzzaman and Md Saiful Islam.** s.l. : IEEE, 2019. In 2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON). pp. 1-5.

17. *Detecting Sentiment from Bangla Text using Machine.* **Muhammad Mahmudun, Md. Tanzir Altaf,Sabir Ismail.** 11, Sylhet, Bangladesh. : International Journal of Computer Applications, 2016, Vol. 153.

18. *A real-time Twitter sentiment analysis using an unsupervised method.* **Noureddine Azzouza, Karima Akli-Astouati, Amira Oussalah, Samy Ait Bachir.** s.l. : the 7th International Conference on Web Intelligence, Mining and Semantics, 2017. 1–10.